NTNU – Trondheim
Norwegian University of
Science and Technology

Defining Tags by Linking to Knowledge Bases

Geir Ivar Kihle Hanssen

Master of Science in Informatics
Submission date: March 2014
Supervisor: Herindrasana Ramampiaro, IDI

Norwegian University of Science and Technology
Department of Computer and Information Science

Defining Tags by Linking to Knowledge Bases

Geir Kihle Hanssen

March 1, 2014

1 Preface

This was written to partially fulfil the requirements for the degree of Master of Science in Computer Science, specialisation in Information Management, at the Department of Computer and Information Science, at the Norwegian University of Science and Technology (NTNU).

1.1 Acknowledgements

First and foremost, I would like to thank and express my gratitude for all the guidance and help given to me by my advisor Heri Ramampiaro. I would also like to thank my colleagues at Orakeltjenesten NTNU for their support, both with technical issues as well as their contributions on discussions regarding solutions and ideas. I would also like to thank the people who helped me proofread this thesis: Therese Krystad Myrvold, Jens Austerheim, Per Eigil Aalborg, Kristin Blindheim, Andreas K. L. Belboe and Gaute Lier Guldahl.

Lastly I would like to thank all those who participated in my surveys.

2 Abstract

This thesis looks into the process of automatically expanding image searches based on tags and the definitions of terms from public knowledge bases. To this end, we will try to extract terms related to a query. The process of finding these terms is known as feature extraction. The text collection on which we perform this feature extraction is, in this thesis, based on text retrieved from public knowledge bases using the original query. The program will in other words, first retrieve related documents. It will then pick out related terms using either Chi-Squared, or an approach I've coined "Neighbouring Terms", or NT. The latter is an approach that is much quicker to process, and may prove to give good precision to term extraction, despite not having to perform such a demanding process beforehand.

This thesis will also look into different variables in these kind of processes to find the best approach to both Chi-Squared and NT. Because this automatic term extraction is set to work on a limited size of articles, there is a question of how many articles would be needed to get the best results. There are also several similarity models to consider when building something like this. For that reason, this thesis also looks into the different results obtained when working with models like the Vector Space model, Okapi BM25 and the Language Model.

Other variables that this thesis looks into is whether or not term pre-processing, like stop word removal and stemming, are beneficial or not. Also, what gives the best results between searching for abstracts based on their title or their contents, and with how many terms can a query be expanded without losing too much relatedness.

To evaluate the terms suggested by these methods, this thesis looks into the $P@n$ values for 20 queries, as well as using metrics such as MAP (Mean Average Precision) to evaluate the sum of the results for each approach. To avoid biased evaluation, we also perform a user survey. We present the results of a survey where 32 people have given their opinion on the different terms suggested by the system, and how related to a given query they are.

The main conclusion in this thesis is that NT does run faster than Chi-Squared, but while results did vary, the precision values on an average fell in favour of Chi-Squared. That said, it did not perform better by much, and with future improvements it could prove a viable solution in automatically generating semantically related terms without having to perform heavy processing.

Contents

3 Introduction

3.1 Motivation

We have crossed the threshold where we now have so many pictures available to us when searching that we often struggle to find that perfect image that fits our needs. Somewhere out there in that sea of pictures is the picture you want, but the task of finding it can be challenging. What terms should you use when searching to find the correct one? It is very likely that someone posted a picture you want, but tagged it with other ideas in mind than what you have. It may even get dropped by the search engine because it is not deemed "relevant enough" based on your choice of tags to look for.

When the typical user searches for a given type of image, he/she will usually try different kinds of terms as queries until a fitting hit has been found. But what if a system could suggest those extra tags? What if, when a user searches for "computer", the system automatically adds related terms to the search that expands the base of tags being searched for, while still maintaining a high precision in the resulting list of images?

3.2 Problem Specification

This thesis will look into the use of two methods for automatically expanding a user query by processing text from a public knowledge base, such as DBpedia.org [22]. It will do this with the use of indexes built using the Apache Lucene project, which will be more thoroughly introduced later in this thesis. This task rises several problems.

First of all, we need to look into what kind of variable settings gives us the best results. Variables like similarity models, sizes of datasets and different ways to search for related articles from DBpedia. There is also the problem of what kind of expanded terms could we get for a given single term query given a certain extraction method? We can expect two different approaches to give different results, so which one gives the best results?

We also need to establish exactly how related the automatically gathered terms will be, and in a non-biased manner at that. Additionally, we need to establish how well an image retrieval search perform based on tags using the query expansion from the different approaches.

3.3 Organisation

This thesis is organised as follows: In chapter 4 we discuss different projects and studies that are similar or related to this thesis. Chapter 5 describes the different concepts used in this thesis, as well as giving insight into how the program written to prove our theories is built and how it works.

Chapter 6 describes the experimental setup, i.e. where the data set collections are from, how the approaches suggested work and how they are implemented, as well as some insight into the evaluation metrics used. We also describe the user survey done in this thesis. Lastly, this chapter presents the results of the evaluations done in this thesis.

Finally, in Chapter 7 we sum up our findings, both for the evaluations of the different approaches, as well as the results from using the optimal approaches in an image retrieval process. Furthermore, we present ideas for future work and suggestions on what could be done to possibly improve the results.

4 Related Work

4.1 Semantic-Link

Semantic-Link is a project created as a demonstration available through a web site. It utilises the mutual information concept on contents from Wikipedia. Every query returns 100 terms that are deemed relevant, where each term presented is a term that is used at least 1.000 times in Wikipedia [17].

This is more a project than a research paper, but works generally speaking as we intend in this thesis, although with some key differences. Semantic-Link works by pre-calculating Mutual Information values for a given term with other terms that appear in the same document.

Semantic-Link works quite fast and seems to also give a rather good term-wise precision, but is also dependent on a very large pre-built matrix. Another issue is that it makes its suggestions based on how the query relates to all terms in the entire Wikipedia collection. This means that terms that may be related to the *subject* of the query, but not necessarily used much together with the query term, may be viewed as non-related. My approach in this thesis attempts to remedy this, and find related terms to a given subject instead of a given term.

In our approach we hope to fix several of these shortcomings by firstly basing the calculations on a smaller set of documents where most of the documents already should be related to the query. We then find terms commonly used in these articles, instead of looking at terms that have some lexical relationship with the query term. In doing this we should be able to find commonly used terms regardless of whether or not the query is a commonly used term.

4.2 Finding Semantically Related Words in Large Corpora

Finding Semantically Related Words in Large Corpora is a very similar approach to that of the Semantic-Link. The difference is that in this paper, they are using a much larger text collection consisting of 100 million terms. They used Mutual Information equation in their calculations and created term clusters based on their calculations to create a hierarchy of semantically related terms.

Their approach with using a clustering hierarchy fixed to some degree the problem Semantic-Link had in how the resulted terms would be directly related to the query term, as opposed to the subject the query term could be within. It is still strongly and directly related to the query term, but by using the clustering approach they were able to group terms together in a way that allowed for a more precise and subject oriented term suggestion [23].

A problem with their approach was however the sheer size of the processing task. The process is so demanding that a super computer was needed for their processing [23]. The average person do not have access to a super computer, and this solution cannot be implemented in any practical means by a normal home computer. Their findings were largely successful, but highly impractical.

My approach aims to perform these calculations on the run on a much smaller set of data. Mutual information techniques are more efficient the more data they have to work with. However, the theory is that if the dataset used to calculate relatedness is mainly about a single subject (or a group of subjects related to each other), the resulting terms based on mutual information calculation will also be related to that subject.

4.3 Query Expansion Using Wikipedia and DBPedia

In this paper, the researchers try to find related terms through texts from Wikipedia and DBpedia. Their approaches work by searching Wikipedia for articles related to a user supplied query. These articles are then sorted according to TF-IDF (Term Frequency - Inverted Document Frequency) scores compared to the query.

A subset of these are then selected as "concept candidates". They then calculate an ESA (Explicit Semantic Analysis) relatedness score between the query and this list of candidates. The new ESA score is then used to select new concepts deemed relevant to the original query [1].

The approach is in other words an attempt to improve the results of searching through Wikipedia articles to build a concept, but is not focused on using that data to generate a list of single, stand-alone terms that should or should not be related to a given subject.

The approach described in our thesis will try to remedy the aforementioned problem by using a set of articles deemed related to a search, and extrapolate independent terms that are related to that of the collected article's subject.

4.4 RiTa.WordNet Hyponym Extraction

RiTa is a Java library that enables semantic evaluation of terms. Its functions allows the user to identify terms as nouns, verbs, adjectives or adverbs as well as retrieving synonyms and hyponyms to query terms. This is not work done through any form of co-occurrence or mutual information calculations, but is relevant to the work of expanding one term into more related terms.

Hyponyms are terms directly related to a given term. The RiTa API explains a hyponym as follows:

"X is a hyponym of Y if there exists an is-a relationship between X and Y. That is, if X is a subtype of Y. Or, for example, if X is a species of the genus Y. If X is a hypernym of Y, then Y is a hyponym of X. Examples: Artefact is a hyponym of object, object is a hypernym of artefact, carrot is a hyponym of herb, and herb is a hypernym of carrot" [11].

It is in other words *very* simple to get a selection of terms that are more or less guaranteed to be relevant to a query by using RiTa.WordNet. The precision of the relationship between query and return set is very high, and it is therefore also very likely that the precision of the relationship between query and images would also be very high.

A potential weakness with using the WordNet-approach, is that while a hyponym is very likely relevant to a query, it may not be diverse enough to give the result set the spread of vocabulary that other methods might give.

For example: Finding the hyponyms of computer would likely give other terms and names for a computer (like node, client, server, host etc.), but would not give any terms about other things *related* to a computer, (like network, process, programmer etc). The latter is exactly what we hope to achieve in this thesis.

5 Approach

There are different variables to consider when implementing this thesis' approach. Either one of which can potentially make or break the outcome, so it is important that they are carefully selected. In this section we introduce the different preprocessing steps, the different similarity models utilised, the feature extraction methods CSMI and NT, and how the program is built.

5.1 Data Processing

5.1.1 Stemming

In information retrieval, it is common to perform stemming on terms. Stemming is the process of attempting to return a term to its roots. I.e. automatically remove any suffixes that may be the result of some form of grammatical conjugation. The primary goal is to turn terms like "biologically" and "running", into "biology" and "run" [15].

A commonly used stemmer is the Porter Stemmer which works by following a linear set of rules. Each rule analyses a term and removes parts of the end of a term where the rules applies. The end result of "biologically" and "running" would here be "biolog" and "run" [18]. For some terms it works fine, and other terms becomes "crippled" according our human interpretation.

Another form of stemming is called lemmatisation which does a much more thorough job of analysing a term to return it to its roots without crippling the term like the Porter stemmer does. This process is a lot more demanding on the system however, and will usually not give a particularly superior result over standard stemming.

While stemming is a common preprocessing task, our experiments with this system only confirms what is already a known fact: Stemming may harm precision while being beneficial for recall [15].

5.1.2 Stop Words

A stop word is defined as a term with little or no semantic meaning in itself. Terms like "is", "are", "this", "be", "to", "from" and so on [19]. Apache Lucene also contains by default a basic stop word list used to remove certain stop words. Lucene's own stop word list is very limited (only 33 terms) and we have therefore chosen to implement a list manually. The list of stop words we have fetched from a comprehensive list of terms commonly ignored by most search engines [27]. This list contains 635 terms. It is worth noting that this list likely seems larger than it really is, as it covers many non-stemmed terms as well. Terms like "he", "he'd", "he'll" and "he's", instead of just "he".

Current research shows that removing stop words is helpful for information retrieval, and does indeed improve precision of searches like the ones this thesis is focused on [19]. My own experiments with removing vs. not removing stop words only confirm this. Without a doubt, every query performs much better when removing stop words, as is expected.

In this thesis we also often come across terms containing numbers and special characters. We will in this thesis also consider these terms very likely to be without semantic meaning and, by that extension, stop words.

5.2 Similarity Models

In this thesis we will compare results using three different similarity models for Lucene to evaluate whether or not the different methods may (or may not) yield better results than the other. The methods chosen for evaluation here are the Vector Space model, Language Model with Jelinek-Mercer smoothing, and Okapi BM25.

5.2.1 Vector Space Model

Vector Space model is the default similarity measure used by Lucene and is a compound similarity model based on TF-IDF, which is found through combining several measurements. First off is TF, or *term frequency*. This model calculates a weight by counting occurrences of terms within the different documents.

$$TF_{t,d} = 0.5 + \frac{0.5 \times f_{t,d}}{max\{f(_w, d) : w \in d\}}$$

Here the 0.5 values are inserted as a smoothing constant and $f_{t,d}$ is the frequency of term t in document d. It is divided by the frequency of the most frequent term in document d, regardless of which term it is [2].

This is a very simple measure that gives insight into how many times every term is used in every document. The measure is, however, not very good at identifying how meaningful a term actually is. Some terms are much more common than others, even if we remove stop words that are semantically insignificant on their own.

The second measurement needed is IDF, or *inverted document frequency*, which gives a measure that tells us something about how common a term is in a collection of documents. It starts off by counting a document frequency (DF) for a term. This means it looks at individual documents in a collection and looks for a specific term. If it finds it, it increases a counter and moves on. In the end, the DF tells us how many documents contain the term at least once. This value is then processed using the following equation:

$$IDF_t = log \frac{N}{DF_t}$$

Where N is the total number of documents, and DF_t is the document frequency for term t.

Finally we use both the TF and IDF value and multiply them together to find the TF-IDF value for the term in a collection:

$$TF - IDF_{t,d} = TF_{t,d} \times IDF_t$$

5.2.2 Language Model

A language model is a statistical measure that assigns a probability to a sequence of terms by means of a probability distribution. In information retrieval, using this model the approach in a search boils down to the computer attempting to estimate the *probability* that a document would generate terms matching the query.

$$P(q|d) = \prod_{i=1}^{n} P(q_i|d),$$

where q is a query (a set of terms), d is a document and n is the number of terms in q. q_i is one term within the query. This method, however, has a problem. It is likely to underestimate the probability of any unseen term [29]. In an attempt to fix this, the concept of smoothing was introduced. Several smoothing methods have been suggested. One of which is the Jelinek-Mercer method:

$$p_\lambda(w|d) = (1 - \lambda)Pml(w|d) + \lambda p(w|C)$$

This method uses a linear interpolation of the maximum likelihood model with the collection model, using a coefficient λ to control the influence of each model. Experimentation performed by Zhai and Lafferty indicate that using a λ coefficient of 0.1 is best for very short, title-like queries, which is what fits our implementation best [29].

There are other smoothing models, but Jelinek-Mercer tends to perform better on smaller training sets [14].

5.2.3 Okapi BM25

This method is a ranking method based on probabilistic retrieval framework. It works by ranking a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document. This method also uses IDF as a part of all the calculations it does.

$$\text{BM25 Score(d,Q)} = \sum_{i=1}^{n} IDF(q_i) \times \frac{f(q_i, d) \times (k_1 + 1)}{f(q_i, d) \times (1 - b + b \times \frac{|d|}{avgdlen})},$$

where d is the document, Q is the given query, q_i is one of the terms within Q, n is the number of terms within the query, $|d|$ is the length of the document in terms of terms, and *avgdlen* is the average document length in the text collection from which documents are fetched. k_1 and b are free parameters. Commonly used values for these are $k_1 \in [1.2, 2.0]$ and $b = 0.75$ [20].

5.3 Term Retrieval Algorithms

5.3.1 Term Co-Occurrence Using Chi-Squared

This concept is only *based* on the concept of mutual information statistics, and uses the Chi-Squared equation for calculation. The classic method compares the probability of observing term x and term y independently. If there is a clear relation between term x and term y, then the joint probability of both term x and y will be bigger then the probability of finding just term x multiplied by the probability of finding just term y. In other words then, the stronger the relationship between term x and y, the closer the resulting calculation (equation shown below) will be to 0 [5].

$$MI(x, y) = log \frac{P(x, y)}{P(x)P(y)}$$

A suggested alternative to this equation is using an equation called Chi-Squared:

$$Chi - Square_{(x,y)} = \frac{\left(\frac{P(x,y)-1}{|N| \times P(x) \times P(y)}\right)^2}{P(x) \times P(y)},$$

where $|N|$ is the total number of articles, and $P(x)$, $P(y)$ and $P(x, y)$ is the statistical probability that a term x and/or y occurs in a given document, given by

$$P(x) = \frac{|DF_x|}{|\mathrm{N}|}$$
$$P(y) = \frac{|DF_y|}{|\mathrm{N}|}$$
$$P(x, y) = \frac{|DF_{xy}|}{|N|}$$

where $|DF_n|$ is the number of documents containing at least one case of term n (document frequency), and again, $|N|$ is the total number of articles [6].

If the document frequency of either term x, term y or term x and y is 0, then chi-square would be set to $Infinity$. The closer a set of terms are to 0, the more feasible it is that they are related to the subject of the articles in the return set. Chi-Squared is a method that quantifies the lack of independence between terms. In other words, pairs of terms getting a high Chi-Squared value are very independent terms and do not relate to each other [16].

A difference in our implementation (compared to the usual implementation of this method), is that we have opted to calculate mutual information on only the return set of articles and not the collection as a whole. This is because, in theory, terms that score well using this approach should be used often together in the articles relating to a specific subject. It can therefore be theorised that these terms also are related to the subject of a set of articles that should be related to a user's query. To differentiate our implementation of this from the classic approach, we will refer to our approach as CSMI, or Chi-Squared Mutual Information.

5.3.2 Neighbouring Terms

This approach is based on the concept of term co-occurrence within a certain window (i.e. a given maximum distance between the two terms). Only in this case, the window is very narrow. While other approaches perform this kind of analysis on terms that are at most i terms apart in a sentence, this approach only looks at terms that occur together as neighbouring terms. A pair of terms are scored by the following:

$$
\begin{aligned}
D_j &= \{t_{j1}, t_{j2} \ldots, t_{j|D_j|}\} \\
R &= \{D_1, D_2, \ldots, D_{|R|}\} \\
NT_{x,y} &= \sum_{j=1}^{|R|} \sum_{i=1}^{|D_j|} t_{ji} \times t_{j(i+1)}, \\
\text{where } t_{ji} &= \begin{cases} 0 & \text{if } t_{ji} \nexists (x) \\ 1 & \text{if } t_{ji} \exists (x) \end{cases} \\
\text{and where } t_{j(i+1)} &= \begin{cases} 0 & \text{if } t_{j(i+1)} \nexists (y) \\ 1 & \text{if } t_{j(i+1)} \exists (y) \end{cases}
\end{aligned}
$$

where R is the result set containing documents D, and t_j is one specific term within D_j. t_{ji} can become either 0 or 1, depending on whether or not the term it represents exists either as the term x, or for $t_{j(i+1)}$, the term y.

This equation will calculate the number of occurrences a pair of terms has together in a set of documents R. For example, given a set R with

$$
\begin{aligned}
D_1 &= \{\text{A B C}\} \\
D_2 &= \{\text{B C D}\} \\
D_3 &= \{\text{A D C}\}
\end{aligned}
$$

Running $NT(B, C)$ would result in the algorithm first picking up D_1 and first running the first and second term. $t_{1,1} \times t_{1,2} = 0$, because $t_{1,1} = 0$ as it does not match the given term defined as x (in this iteration "A"), and $t_{1,2} = 0$ as it does not match the given term defined as y (in this iteration "B").

In the next iteration, $t_{1,2} \times t_{1,3} = 1$, because $t_{1,2} = 1$, and $t_{1,3} = 1$, as both terms matches their corresponding term defined as x (in this iteration "B") and y (in this iteration "C"). In the cases where t_{ji} matches term x, but $t_{j(i+1)}$ does not match term y, the result is still 0 as $1 \times 0 = 0$.

The end result is that after iterating through all pairs of terms in all documents in R, we have a value corresponding to the number of times "B C" occurs together.

A weakness in this approach is that pairs such as "A C" would not be calculated and stored, even though they co-appear in the same documents just as much as "B C" does. This is something CSMI takes into account so the question then is whether or not this feature will result in a better result or not.

5.4 Program Build

The system is written using Java. The main reason for this is to easier enable the use of the Apache Lucene search engine [24] which is built on Java.

Figure 1 shows an overview of the system architecture and what components communicate with each other.

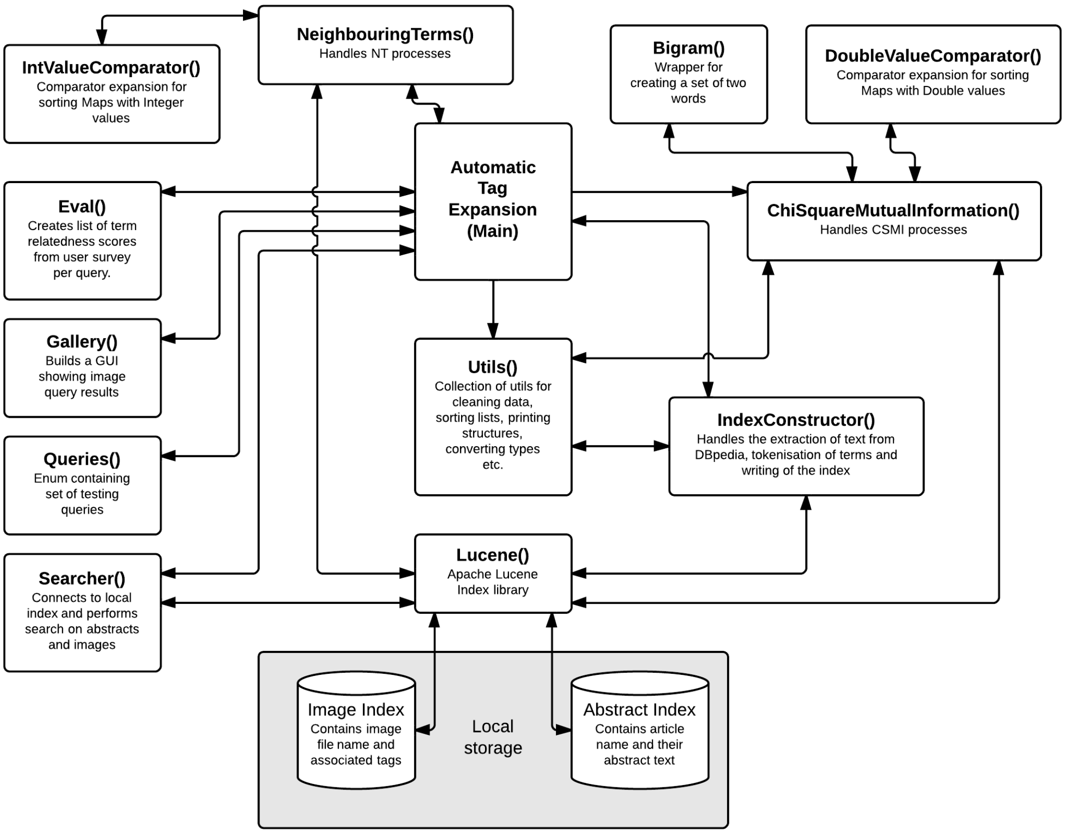

Figure 1: *System architecture overview*

5.4.1 Building Local Index

Lucene indexing works by taking a document, or a collection of documents, and tokenises the contents. This means it breaks up the terms in the document into single terms based on a specific set of rules. These rules depend on what kind of "analyser" we choose for the job. In this thesis we have chosen to use an analyser called "WhitespaceAnalyzer". A commonly used analyser is the StandardAnalyzer, but using this results in some impractical effects. The StandardAnalyzer breaks down a stream of text into single terms wherever any hyphen, special characters, punctuation, commas, white spaces etc. occur. This is impractical because terms like the ethernet cable standard "CAT6" would be broken down to the term "cat". Obviously, an ethernet cable has nothing to do with the feline species, so we need to avoid this. WhitespaceAnalyzer fixes this by only separating terms only wherever there is a white space (and typical sentence ending and/or pausing characters like punctuation, comma, exclamation marks etc) [25].

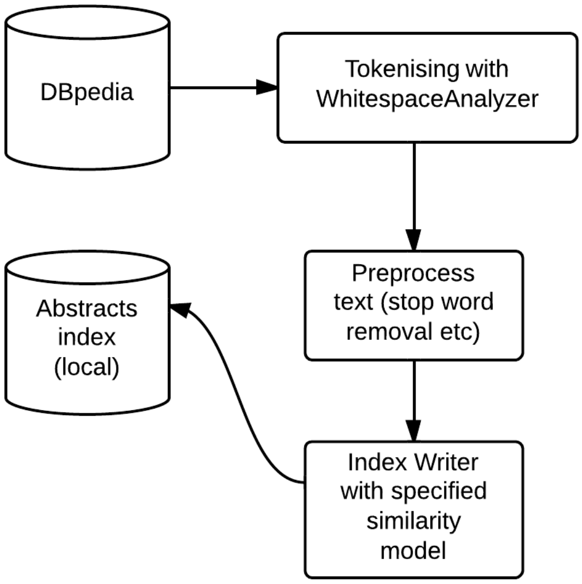

Figure 2: *Building the index*

The process of building the index is shown in Figure 2. The local index is constructed using data from DBpedia as the document collection. Each article element is fetched from the collection, tokenised using the analyser, and preprocessed with stop word removal and other cleanup processes to remove any non-lexical elements. The dataset contains several instances of HTML-code, metadata tags and other text that do not have any semantic value.

When all text for one article has been processed, a similarity model is selected for the index writer, and the content is written to an index using Lucene. When all articles are processed and added to the index, the index is finalised and stored locally on the computer, ready to be read later on.

The system also creates a second index for the images in the collection and their respective tag set. Each image has their own .txt-file containing a list of tags affiliated with the image. There are, however, some images whose corresponding .txt-file is empty. This is because these images from Flickr simply has not been tagged with anything. This image index is not pre-processed in any way and simply store the image name along with its corresponding tags.

Whenever the program is run, the indexes are loaded into the program. Because of the demanding process of building this index, the indexes are built only once, after which the code to trigger the index building is commented out. Any updates to the index can be simply added by running the builder again, where the contents of the input dataset is completely new. Using a a dataset that contains some of the same elements as the one originally used will result in duplicate elements. There is redundancy check, as this would increase the runtime significantly.

6 Evaluation

In this chapter we present the collections we selected for this thesis. We describe the implementations of CSMI and NT, introduce and describe different evaluation metrics, and present the survey conducted for this thesis. We also present and discuss the evaluation results from testing the different variables previously introduced.

6.1 Experimental Setup

6.1.1 Data Collections

I selected an image collection containing 1 million images. The image collection is based on images from Flickr [26], and is maintained by a group of people calling the collection the "MIRFlickr" [12]. All images are under the Creative Commons license. All images have a corresponding txt-file containing all tags each image has, if any. It is worth noting that, as stated earlier, some of the images has no tags, and their corresponding .txt tag files are empty.

The abstract definitions are fetched from DBpedia.com [22]. This web page offers a wide range of data collections from collections containing only body texts to collections containing just the abstracts, all of which are fetched from Wikipedia [7]. Since we want only a concentrated collection of terms with a high probability of them being relevant to the article, we selected a dataset containing just over 4.000.000 abstracts along with the article name they are fetched from.

6.1.2 Processes of Finding Related Terms

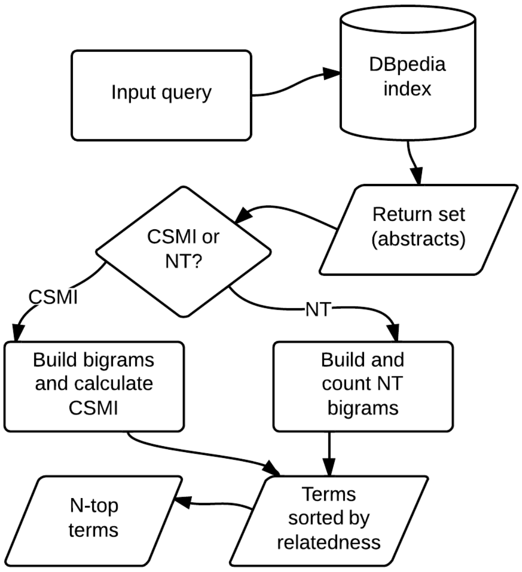

Figure 3: *Processes involved in a query*

In Figure 3 we present the feature extraction process in our implementation and how the most related terms are found. The figure shows a decision where either CSMI or NT is selected. In other words; the implementation does not run both extraction methods when executed. Each implementation work as follows:

Chi-Square Mutual Information (CSMI): The program collects all terms from the return set in a list. All terms are then counted to obtain their frequency which is then stored in a Map where the keys are the terms, and a given key's corresponding value is the frequency of that term. The Map is then traversed, discarding any term with a frequency below a threshold t.

This discard process is motivated by a couple of reasons: 1: Creating bigrams of every term combination possible often results in a list so large that java heap space cannot handle it and the program crashes. 2: More importantly: As stated by Smrž and Rychlý, research on using approaches like MI (Mutual Information) and Chi-Square for locating related terms have shown that these approaches are not very good for low-frequency entities. It is therefore necessary to drop the low-frequency terms [23].

The processed Map is then used to build bigrams in another Map so that all terms that passed the frequency threshold are paired with all the other terms into bigrams. A bigram is a pair of terms combined into one element of term x and term y. For example, a bigram(x,y) can be "computer network", where $x = computer$ and $y = network$. All bigrams are then used to calculate their corresponding chi-square value using the return set as document collection.

In the resulting list of bigrams, sorted by their chi-squared value, we find the terms deemed most related to the subject closest to 0. The higher the value, the less statistical chance that the bigrams are related to the subject of the return set.

We then extract the N elements closest to 0 and add them to the original query before using the new and expanded query to search for images.

Neighbouring Terms(NT): The process of locating these pairs is done by looking at all the text from a document in the return set. Every pair of terms are first picked out, sorted alphabetically, and then stored in a HashMap along with an iterator that is increased by one every time another equal (sorted) pair of terms is located. In the end, after going through all documents in the return set, the HashMap is sorted on the values found after counting and the i most reoccurring pairs are selected as expansion terms.

The idea is to count the number of times two terms appear next to each other, regardless of the order they appear in. The reason for alphabetically sorting the order of the pairs is to have one single point of reference in the HashMap. For example, the pair "computer networking" and "networking computer" would be stored as "computer networking" as this is the alphabetical order of the pair. The idea is that when encountering these two terms together, it does not matter what order they are in. If they both occur together quite frequently, it is still very likely that they are related to the subject of whatever the return set is about. Therefore, if one article contains "computer networking" and another contains "networking computer", they would both increase the counter in the HashMap addressed with the key "computer networking".

For example: If the query is "computer", the program will find articles matching this query. The text of these articles is then traversed, building commonly occurring bigrams like "computer program", "computer instructions", "computer network", "cpu instructions" etc. The program will then traverse these bigrams starting from the most common, selecting unique terms. I.e. a term that is already selected is not selected again. The end result would then be a set of terms to be used for query expansion with terms like "computer, program, cpu, instructions, network" etc.

The N most frequently occurring terms are then added to the original query before using the new and expanded query to search for images.

6.2 Evaluation Metrics

In this thesis, a problem of evaluating results is how to determine relevance of retrieval results. All images also contain tags, and the success of a retrieval process can both be measured on the tags contained in each image, and the image's contents. Say for example we perform a search containing the query term "boarding" while thinking of the act of going onto a ship. We then hypothetically get a picture of a skateboarder, a snowboarder, a boarding school and people boarding a ship. All of which contains the tag "boarding". Are then all images related because they contain the tag "boarding", or is only one related because only one is of what we intended to find?

Because the goal is to view the relatedness of images from a query using the expanded query, and because images can be tagged erroneously (either as a result of a mistake, or by the user misunderstanding what a tag may mean) we will in this thesis look at the contents of the images and whether or not they depict something related to the original query.

6.2.1 Precision

Precision is a calculated value that tells us something about how much of the returned result set R is in fact relevant to our search. The value is found by:

$$P_k = \frac{A_k}{(A_k + B_k)},$$

where P_k is the precision value of a search k, A_k is the number of relevant hits in R, and B_k is the number of non-relevant hits in R. The resulting value will be somewhere between 0 and 1, where the closer it is to 1, the better the result is [2].

For example, say we retrieve 10 hits from a collection, where we determine that 6 of the hits are in fact related. Then the precision of this retrieval is $\frac{6}{6+4} = \frac{6}{10} = 0.6$. If we attempted to increase the number of hits from the query to 20, and we this time only found 8 hits, then the new precision would be $\frac{8}{8+12} = \frac{8}{20} = 0.4$. In other words, we found more hits by increasing the size, but because the returned set now contains even more unrelated hits, the precision dropped.

6.2.2 Recall

While precision tells us how relevant the result set from our search was, recall tells us something about to what extent our search found every relevant item possible in the complete collection of documents.

$$R_k = \frac{A_k}{A_k + C_k},$$

where R_k is the recall value of a search k, A_k is the number of relevant hits in R, and C_k is the number of relevant items *not* in R (i.e. items not found by the search) [2].

Example: Say we know for a fact that for a specific search there are 10 relevant documents in a collection. When we perform a search we recover 7 relevant items. Filling out the equation we get:

$$R_k = \frac{7}{7+3} = \frac{7}{10} = .7$$

Increasing the number of hits retrieved does nothing to the recall value itself, unless the expansion means including more relevant hits.

6.2.3 Precision vs. Recall

Precision vs. recall tends to be a difficult matter. We can increase recall values by including more items in our result set R, but doing so will most likely have a negative impact on precision. Say our collection C contains 10 relevant documents and R is set to a maximum of 10 hits. We perform a search and get 5 of the relevant documents in R. This gives us a precision of 0.5 and a recall of 0.5.

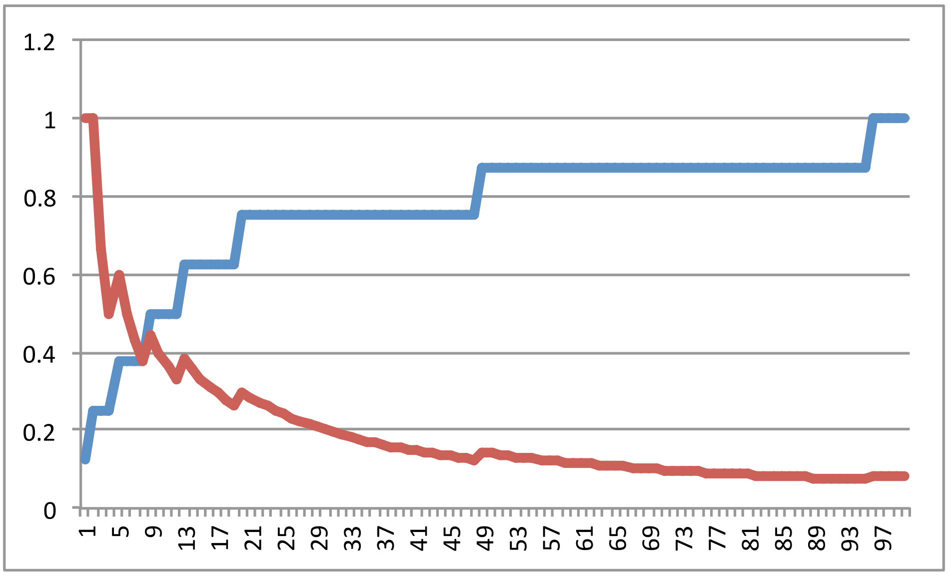

Figure 4: *Example of a precision vs recall graph where the X-axis (horizontal line) is sizes of R, and Y-axis (vertical line) is the value. Blue line: recall, red line: precision*

Next we try to increase the size of R to 15 and redo our search. We now find 1 more relevant documents, thus increasing recall to 0.6. The problem is that there are now 4 irrelevant documents more in R, which gives us a precision of 0.4 An example of the values of precision and recall as the size of R increases can be seen in Figure 4. (This graph has nothing to do with the data in this thesis. It is merely a demonstration of the precision to recall relationship).

It is theoretically possible to get both a high recall *and* precision. Say all 10 items in R was all 10 relevant items in C. This would give a scoring value of 1 on both precision and recall. This is however statistically speaking extremely rare, if not impossible.

What would have been best between precision and recall would also depend on the end user. The common web user would prefer to have as many relevant hits in their search results as possible, i.e. a high precision, while people like paralegals and intelligence analysts would prefer a higher recall value [15].

Another issue is calculating recall with a huge collection of documents. Especially when using a pre-assembled one as we are doing in this thesis. The problem boils down to the fact that there is no way of knowing the specific number of relevant hits to a specific search in a collection of 1 million random images and their tags. The only way of calculating the exact recall value for this search would be to manually look through all images in the entire collection and count hits for any given search. That would entail manually looking through 1 million images for each query performed. This is not a feasible solution. We will come back to this problem later in the thesis.

6.2.4 P@n

P@n is a very simple measurement and is used to show the increase or decrease in precision as the numbers of returned documents increase. Say we return 50 documents from a search. First we calculate *P@10* for this set and find that among the first 10 documents in the set of 50, there are 8 relevant documents. Hence $P@10 = 8/10 = 0.8$, even though the return set is actually 50 documents. In the end we have a list of calculations that show any changes as the size of n increases. For example:

$$P@10 = \frac{8}{10} = 0.8$$
$$P@25 = \frac{16}{25} = 0.64$$
$$P@50 = \frac{30}{50} = 0.6$$

6.2.5 Mean Reciprocal Rank

Mean Reciprocal Rank, or MRR, is a method for giving statistical metrics to a list of searches (amongst other things). The value given after calculating tells us something about to what degree a set of searches gives us a related answer as the first hit. The equation looks as follows:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i},$$

where $|Q|$ is the number of queries, and $rank_i$ is at which hit in $query_i$ the first related hit appears [2].

For example, in Table 1 we list a set of queries and their results. We pick out the first related hit and calculate its rank.

Query	Results	First correct	Rank	Reciprocal rank
computer	flower, cpu, ram	cpu	2	1/2
cat	feline, dog, cigar	feline	1	1
football	network, bottle, player	player	3	1/3
food	fish, paper, book	fish	1	1

Table 1: Queries for a MRR calculation

Given this information we could calculate the MRR as $\frac{(1/2+1+1/3+1)}{4} = 0.7$.

This approach does not give any specifics as to what to do if none of the proposed hits are in fact related, or if there are multiple correct answers in the list (which in this thesis is highly probable). In the cases where no proposed hits are related, it is common to simply set MRR to 0. That takes care of the first issue, but regarding the second problem; because this program is highly probable to give search results with more than one related hit per query, a different metric approach may be more useful for this task, like for example MAP [9].

6.2.6 Mean Average Precision

Mean Average Precision, or MAP, is an approach to calculate the average precision for a set of queries given a specific search algorithm.

$$MAP = \frac{\sum_{q=1}^{Q} AP(q)}{|Q|},$$

where Q is the set of queries, $|Q|$ is the number of queries, and $AP(q)$ is the average precision of query q, i.e. the $P@n$ for all values of n summed up and divided by n [2].

For example, a query has $P@1 = 1$, $P@2 = 0.6$, $P@3 = 0.5$, $P@4 = 0.5$ and $P@5 = 0.4$. AP for this query would then be $\frac{(1+0.6+0.5+0.5+0.4)}{5} = 0.6$

The MAP value would then give an average score of a certain algorithm, showing an overall measurement for several different queries where all related hits per query is taken into account. Say we perform 5 queries, each of which has an AP of 0.6, 0.3, 0.5, 0.8 and 0.2. We would then find MAP by $\frac{(0.6+0.3+0.5+0.8+0.2)}{5} = 0.48$

6.2.7 Sampling for Determining Relevance Objectively

One possible solution to the aforementioned problem of determining relevancy, is to create a small subset of images, called a sample, and perform searches on this collection instead of the original collection. By doing this, we must also take into consideration how well the sample may represent the rest of the collection. This is a problem with the concept of precision levels (not to be confused with precision as in precision and recall) and confidence levels.

The confidence level is based on ideas within the *Central Limit Theorem*. Summed up, this theorem states that for a population that is repeatedly sampled, the average value of the attribute obtained by those samples is equal to the true population value. Furthermore, the values obtained by these samples are distributed normally about the true value. In other words, in a normal distribution, approximately 95% of the sample values will be within two standard deviations of the true population [13].

The precision level is a percentage of a form of accuracy. Roughly speaking we say that the percentage given by the precision level is how sure we can be that our results are close to 95% confidence level. In other words, how much we can trust that our findings represent the complete population, or collection [13].

Typical precision levels are ±3%, ±5%, ±7% and ±10%. For our collection of 1 million images, we would respectively need to create sample sizes of 1 111, 400, 204 and 100 images [13]. Since we would need the most precise numbers to be more or less sure of our results, this would still entail manually traversing over 1000 images for every search performed. Given a test base of 20 queries, that would mean that for this thesis to have a good survey of this nature, the respondents would need to look through and evaluate 1000 images for 20 queries = 20.000 images. This is not a feasible goal given our resources with this thesis.

Fortunately, the majority of searches does not require high recall. Most users require just a few relevant documents as the top hits to their search [2]. With this in mind, and considering the complexity of getting a good base to perform recall analysis on, we will for this thesis put focus on the precision of our search results and not recall.

6.2.8 Survey

It is useful to get an objective view of whether or not tags suggested by the system are in fact related to the query or not. Especially considering that one person's view of relatedness may not be shared by other people. To get a more general and objective opinion regarding what is related and not, it would be good to get other people's views on the matter.

To accomplish this we conducted a survey. The survey gather data on 20 queries along with the terms suggested by the system for all similarity models. For example, if the query is "computer" and the Vector Space model (hypothetically) suggests "computer, network, process", the Okapi BM25 suggests "computer, cable, code", and the Language Model suggests "computer, key, peripheral", then the participants would see in the survey for the query computer: "computer, network, process, cable, code, key, peripheral". They then rank the terms according to how they feel the terms are related to "computer".

Each respondent can evaluate a term as either *"not related"*, *"maybe related"* or *"definitely related"*. The three choices are weighted in that order as 0, 0.5 and 1. The score of a term t, for that given query q, is then calculated by

$$Score_{(t,q)} = \frac{1}{n} \sum_{j=1}^{n} S_{(t,j)},$$

where n is the number of respondents and j is the answer of one respondent.

This score can then be used to calculate $P@k$ using the following equation

$$P@k = \frac{1}{k} \sum_{i=1}^{k} Score_{(t,q)},$$

where k is the number from which point we want to calculate precision P for a given query. (I.e. if we want the precision after 5 terms in a result set of 10 hits, k would be 5). This approach is a suggested approach suitable for this kind of surveys regarding precision by Ruocco and Ramampiaro [21].

One problem with performing a survey like this is that it has to be of limited size. We would have liked to collect a survey to more thoroughly check what number of articles would give the best results. We would also have liked to find whether or not searching through titles were in fact superior to searching through abstracts (or vice versa). However, covering all these factors makes for a gigantic survey that would be very time consuming for people to answer. (Roughly 150 terms $\times$ 20 queries $\times$ 3 similarity models $\times$ 2 co-occurrence approaches).

We will hence perform preliminary decisions in regards to what works best based on our own findings and opinions. These findings will be used to limit the number of terms that users will have to form an opinion about in the survey.

To begin with, both CSMI as well as NT had 6 different return set sizes and 3 similarity models. It also had different results based on whether we searched DBpedia-titles or abstracts. Each query also returned 20 terms. In other words; 6 sizes $\times$ 3 similarity models $\times$ 20 queries $\times$ 20 resulting terms $\times$ 2 results based on article title and abstract $\times$ 2 methods of finding related terms = 28.800 terms in total. The actual number was severely reduced as duplicates of one term per query was removed. It was however still way too many to efficiently create a survey.

After the preliminary evaluations we had 3 sizes of return sets containing 20 queries, all of which returned 20 terms, but only for results from searching titles. (We will come back to the reasons for selecting only the titles later in this thesis). All these were done both for CSMI as well as NT, summing up to 2.400 terms in total. A number which was also reduced from not including duplicate terms.

The final number after all these steps was still very large per query which is why we decided to split the survey up into lesser parts. This resulted in 8 different surveys. One survey covered 5 queries with half of the suggested terms, another survey had the same 5 queries, but with the remaining suggested terms. This way the surveys would not be considered too large for the respondents. In total, 4 people took each of the 8 surveys, totalling up to 32 respondents.

6.2.9 Sizes of N

The number of articles, or N, returned from querying DBpedia titles is likely to change at least some of the terms selected as related terms. This is a likely outcome because for larger sizes of N there is more text to base the statistics on. However, the more articles that are included, the less likely they are to be directly related to the query. This means that there exists some size of N which may (or may not) be counter productive. In this thesis we will experiment with the sizes 10, 75 and 150.

6.2.10 Tag Ambiguity and Relevance

One problem with determining the precision of the outcomes from this program is that the precision can be evaluated in several ways. One way is to view images retrieved based on the expanded terms and see if the images retrieved are relevant or not. This would give a more "hands on" result that could show that the search was indeed efficient in fetching relevant images. One problem however is the ambiguity of terms. A "ball" could just as easily be a spherical object as it could be a formal party for example. Say that a query for "prom" returned amongst others the term "ball". Term-wise, this is a relevant term when viewed in context of "prom", but when looking for images of "ball", many of these images could potentially be of a spherical object. The question then is: Is the latter still related to the query?

With this in mind, we have selected a set of queries to perform testing with, where most of the queries are non-ambiguous, but also some that are. The reason for this is to see how the system performs for both ambiguous terms as well as non-ambiguous.

The queries selected to use for testing in this paper are: *"Cloud"*, *"cat"*, *"weapon"*, *"school"*, *"police"*, *"fruit"*, *"meat"*, *"telephone"*, *"machine"*, *"computer"*, *"space"*, *"game"*, *"power"*, *"piano"*, *"moon"*, *"screen"*, *"car"*, *"book"*, *"bike"* and *"soldier"*.

The ambiguous terms in this set are *"space"*, *"power"*, *"screen"* and *"bike"*. "Space" can be both the concept of outer space with stars, moon, planets and so on, and space as in for example an empty room. "Power" can be both electricity, as well as the concept of a force or might. "Screen" can be a monitor or a canvas to project images on, and it can be for example a set of curtains on a rack used for example in a doctor's office where a patient can change clothes. Finally there is the term "bike", which can be both a bicycle, and a motorcycle.

6.3 Evaluating results

In this section of the thesis; we will write our evaluations of the different possible variables in our approach to find the best way of automatically expanding tag queries.

6.4 Evaluation of Approaches

6.4.1 Removing Stop Words vs. Not Removing

I experimented with removing stop words from the abstracts fetched from DBpedia. As suggested by Becker, Naaman and Gravano, it is not guaranteed that removing stop words will be beneficial for the search [3]. It depends on the dataset. In Becker, Naaman and Gravano's paper, they argue that short descriptions will likely not contain many stop words at all. That said, the Flickr descriptions are hardly comparable to those of a Wikipedia text. Additionally, current literature agrees that removing stop words from free text is beneficial for searching. Since Becker, Naaman and Gravano however had found that removing stop words might prove either unnecessary or even give poorer results, we had to test our own data set to be sure.

To conduct our own experiment on this we created a new index in where we did not remove stop words from the abstracts. We also made an index that had no stemming, as well as an index without both stemming and stop word removal. Indexes without stop word removal were - as expected - devastating to the results, regardless of the similarity model used, and regardless of whether or not stemming was present. Any search, regardless of query, resulted in selecting basic stop words as expansion tags.

NT performed slightly better than CSMI without removing stop words, but on an average, the tag precision (based on the number of tags related to the query) was as low as 0.2–0.3.

6.4.2 Stemming vs. Non-Stemming

As stated previously, stemming is known to usually harm precision while benefiting recall. In this thesis we are looking primarily at precision values. Performing queries on a stemmed index, the result sets were still dominated by stop words, regardless of the similarity method. Simply stemming and not removing stop words did not prove a good solution.

When using indexes with both stemming and stop word removal, the results were - as previously found - better than using no stop word removal. The result sets now had more or less only terms with semantic meaning, but they were not on average more closely related to the query than when using only stop word removal. In other words; stemming harmed precision.

The known facts regarding stop words and stemming, in addition to our own findings when implementing, proves that for this thesis the best approach is to perform stop word removal and avoid stemming.

6.4.3 Searching Abstracts vs. Article titles

The indexes consist of documents with two fields: One field for the article title, and one for the abstract (i.e. summary text). Searching through the abstracts vs. searching through the titles also yielded different results.

The results were varied in terms of what terms were selected for expansion, but our subjectively calculated precision of the results showed that the results were on the whole very similar. For example when searching using the BM25-model for 100 articles using the Mutual Information approach, the results are presented in the graph shown in Figure 5.

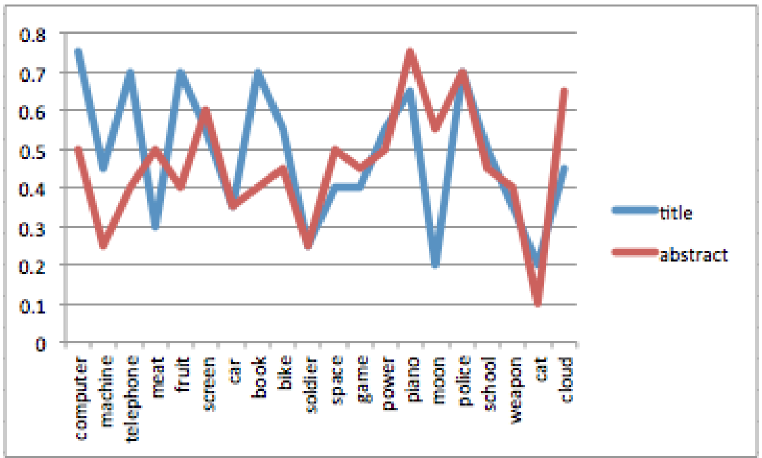

Figure 5: *BM25 model, results based on 100 articles (title vs. abstract text).*

Some queries had somewhat better results when searching through the wikipedia abstracts, but on the whole it was better to search for titles, allthough only slightly better. For example, on average, the BM25 model with 100 articles performed 0.0275 points better with title search than with abstract search. This was a repeated fact (give or take some difference in performance score) for the other similarity models as well, therefore all searches in the final evaluation are based on title queries.

6.4.4 Sizes of N

On looking into what size of N (i.e. what number of articles in the return set) would be the best, we calculated the average relatedness score (based on the survey results) on all 20 terms resulting from running the program on all queries. Then we compared the average value for all queries on all sizes of N and marked the highest average. Finally, counted up how many times each size of N was on average the best for all queries and summed up the results as seen in Table 2.

Model	Vector			BM25			LM		
N	10	75	150	10	75	150	10	75	150
CS	6	8	6	6	7	7	5	8	7
NT	5	7	8	5	7	8	5	7	8
Sum	11	15	14	11	14	15	10	15	15

Table 2: *Number of times different sizes of N gave the best results.*

As evident, $N = 10$ did sometimes result in the best results, but were never as good as either $N = 75$ or $N = 150$. As for the matter of $N = 75$ vs. $N = 150$, this approach was inconclusive in terms of one being best for both methods, as they, in terms of total average, performed equally well. With Vector Space model, $N = 75$ was the best size 15 times, with BM25 $N = 150$ was the best size 15 times, and with Language Model, $N = 75$ and $N = 150$ scored a total of 15 times as the best size of N. Looking closer at the individual results for CSMI and NT it was clear that one size was not necessarily the best size for both approaches.

The average relatedness for all queries for $N = 75$ using CSMI was 0.511, and for NT it was 0.547; summing up to a total average of 0.53. The same for $N = 150$ using CSMI was 0.535, and for NT it was 0.532. Calculating an average then gives us:

$$\text{Avg}(\text{N}_{75}) = \frac{(0.51125 + 0.5471875)}{2} = 0.53$$

$$\text{Avg}(\text{N}_{150}) = \frac{(0.5346875 + 0.531875)}{2} = 0.533$$

Both results are practically identical. However, when viewing Table 2 we can also see that for CSMI, $N = 75$ performs better than $N = 150$ in both Vector and LM, and as good as $N = 150$ for BM25. For NT, $N = 150$ performs better than $N = 75$ in all similarity models. We will therefore conclude that for CSMI, the optimal size of N is 75, and for NT it is 150.

All relatedness values for all queries and all sizes of N can be found in Appendix A and B.

6.5 Analysis of Vector vs. BM25 vs. LM

This section focuses on the results from performing queries with the program and look into the $P@n$ values for the different similarity models. The precision values here are based as earlier specified on the results of the survey. In other words;

when calculating the precision of any given term, they can have any "relatedness" value ranging from 0 to 1 (i.e. including decimal values in between the two).

6.5.1 Results of Queries Using CSMI

In this section we post the results of performing the set of queries with the program, where CSMI, or Chi-Square Mutual Information is the method of determining what terms are related for a given query. The data posted here are graphs showing how the $P@n$ changes for each similarity model as more and more terms deemed related by the program are included. In other words; the far left of a graph is when only one term is fetched, and as it progresses to the right, more and more, up to a total of 20 terms, are included.

Additionally, for each graph there are four tables that can be found in Appendix C. These tables show the specific numeric $P@n$ values, for $n = 1$, $n = 5$, $n = 10$, $n = 15$ and $n = 20$. All three similarity models for the four queries are also given in the tables, as shown in each set of graphs.

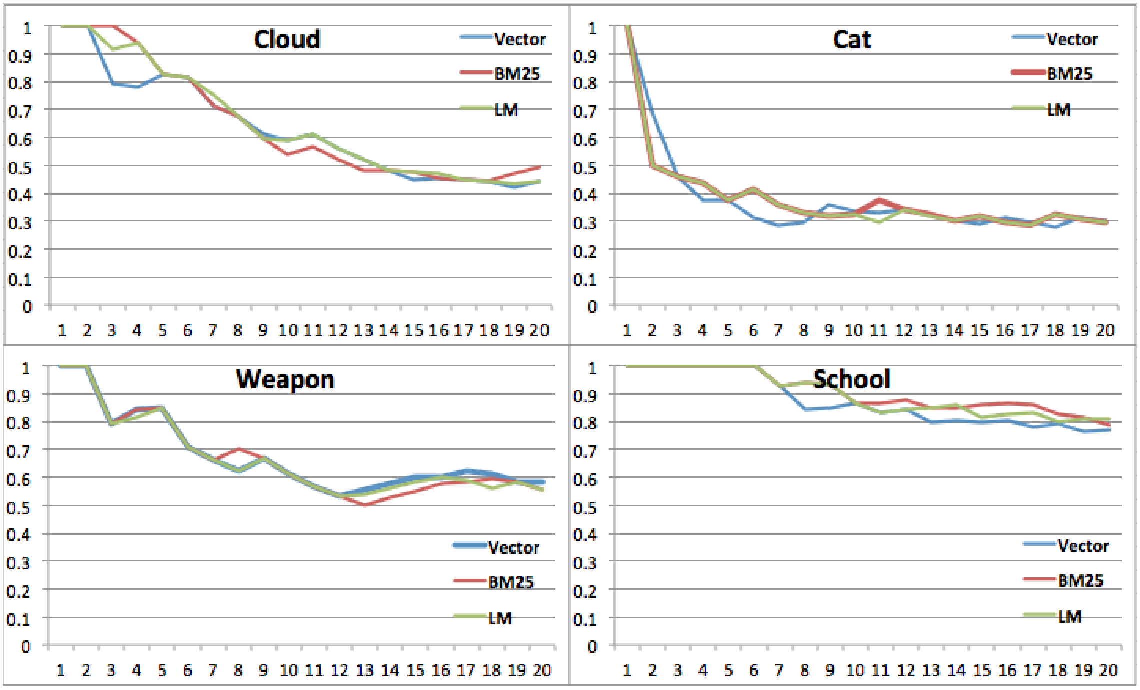

Figure 6: *Graph showing the CSMI* P@n *values as* n *rises to 20.*

In the set of queries shown in Figure 6 we get some varied results. All queries start off with at least one strongly related term, but from there the results start varying. We see that for "cloud", the BM25 model manages to have more strongly related terms on top a bit longer than any of the other models. The vector model drops at this point, but all models converge later and stay more or less together for the rest of the graph.

The query "cat" is a deviation from what can be viewed as normal in these tests. Almost immediately, the graph plummets down to around 0.4-0.5 and never gets any better. It just drops down, and the graph for all similarity models stay somewhere between 0.3 and 0.4. One reason why our approach does not receive many highly related terms here may be a lack of articles containing "cat" in their titles. Many articles regarding different cat races will likely have the race name as a title and not just "cat". These articles are then not likely to be picked up in the first step when searching for related articles to base the term expansion on.

The term "school" on the other hand is performing very nicely. All similarity models produce strongly related terms for as many as the six first terms, and when they start dropping, they never move far away from relatedness values of 0.8. This may be because articles about schools will usually be titled by the name of the school, and most school names do contain the term "school". (Like "London Central Secondary School", "Woodside High School" and so on). This means that many articles about specific schools will get picked out for CSMI and NT processing. Besides, most of these articles will also talk about students, education and other related topics, making it easy to find related terms.

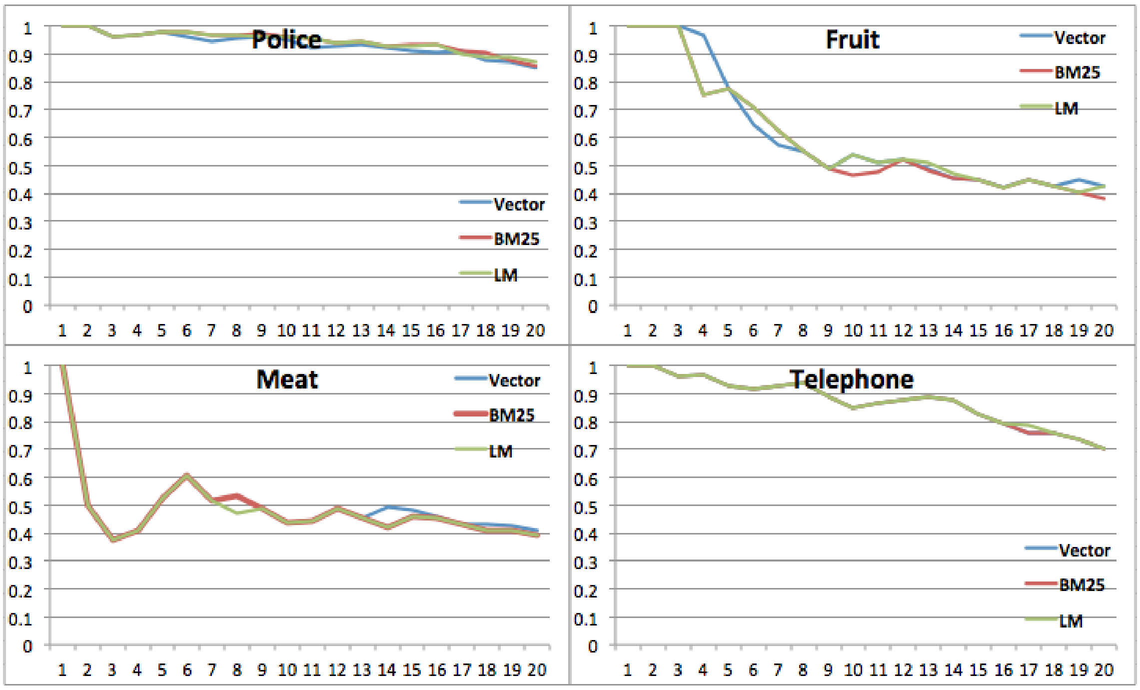

Figure 7: *Graph showing the CSMI* P@n *values as* n *rises to 20.*

In the four queries shown in Figure 7, we see that both queries "police" and "telephone" gives rather good precision values even as n grows larger. Especially so for "police".

"Fruit" has a good beginning, but soon drops down to 0.5, but then evens out and does not drop further down. It can be argued to be one of the queries in this set that keeps more or less to the "norm" for these tests.

"Meat" is the query that performs the more poorly of these four. It almost instantly plummets down to an average of 0.45. While "fruit" comes down to this point as well, it does have several highly related objects as its first hits.

It is somewhat strange that "meat" performs so poorly compared to "fruit". Both are grouping terms that are used to describe a whole range of foods. It may be a bit more difficult to obtain related terms for "meat", simply because it is a grouping term that may not be commonly used as article titles. Since the system is querying titles it is likely not many articles that get hits on the query "meat", as most articles regarding this would likely be more specific. They would likely use more terms like "beef", "pork", "veal" and so on.

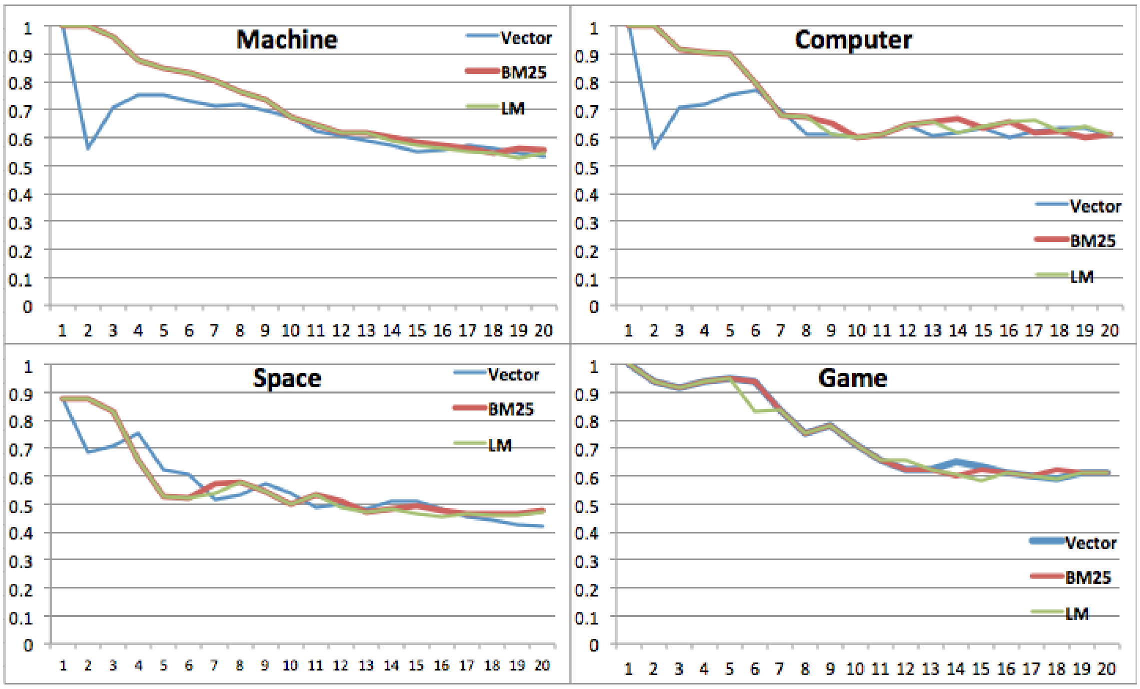

Figure 8: *Graph showing the CSMI* P@n *values as* n *rises to 20.*

In Figure 8 we see four new queries where mostly all of them performs more or less "normally". Notable events are that "space" is performing just a little bit poorer than the rest, while Vector Space similarity takes a dive early on in both queries "machine" and "computer". It ends up performing roughly as well as both BM25 and LM as n rises, but is without a doubt the poorest performer in these cases.

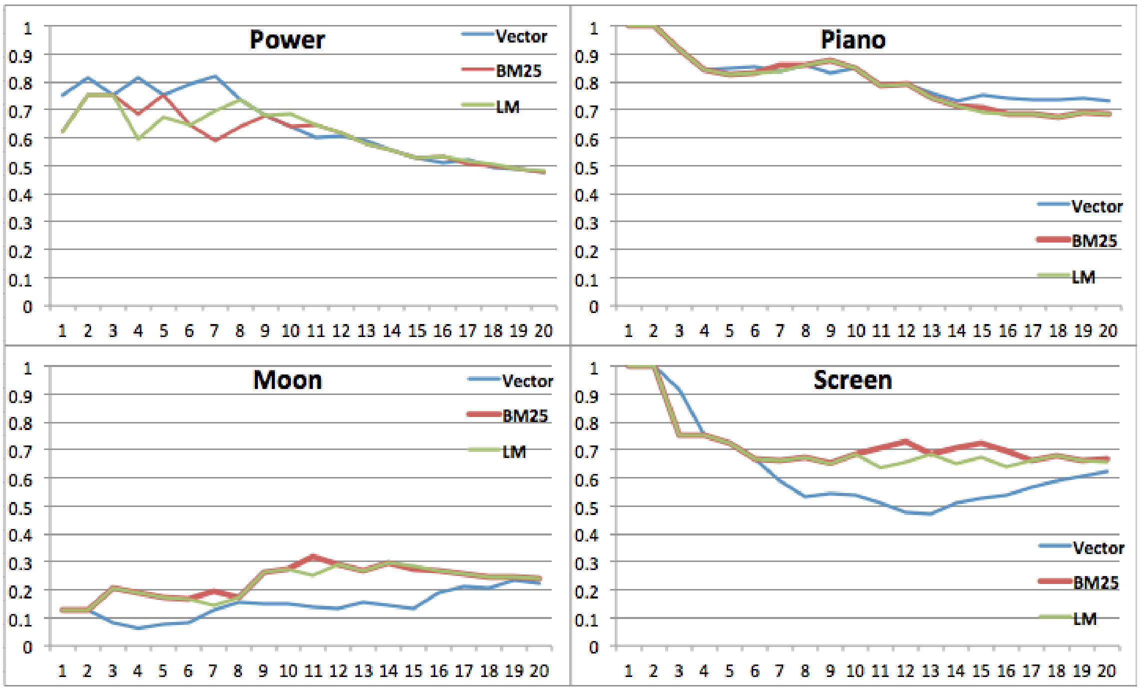

Figure 9: *Graph showing the CSMI* P@n *values as* n *rises to 20.*

In the set of queries shown in Figure 9 we see the first real deviance from what most of the queries show us on average. "Moon" starts off poorly and does perform better as n rises, but never really gets much over 0.3, which is definitively the lowest score of all queries in this test. Again, Vector Space similarity performs on on average worse than the other two almost throughout the entire span. The same can be said for "screen", even though Vector Space does begin at the same scores, and at one point even does better than the other two.

Reasons for why "moon" performs so badly can be hard to explain, but when looking at the terms suggested (presented in Appendix B) we see terms like "united", "states", "america", "released", "album", "record" and "music". "United States of America" is a likely set of terms that appear often. It could easily be because they were the first to land on the moon, and is therefore not that far fetched as a related term. Each of the terms by themselves are however not very related, which is probably why they scored poorly. The remainder of the terms can also be linked to music albums. The moon is often used in album names released by musicians, which have likely become articles in DBpedia which again are picked out as a related hit when searching through the abstracts here.

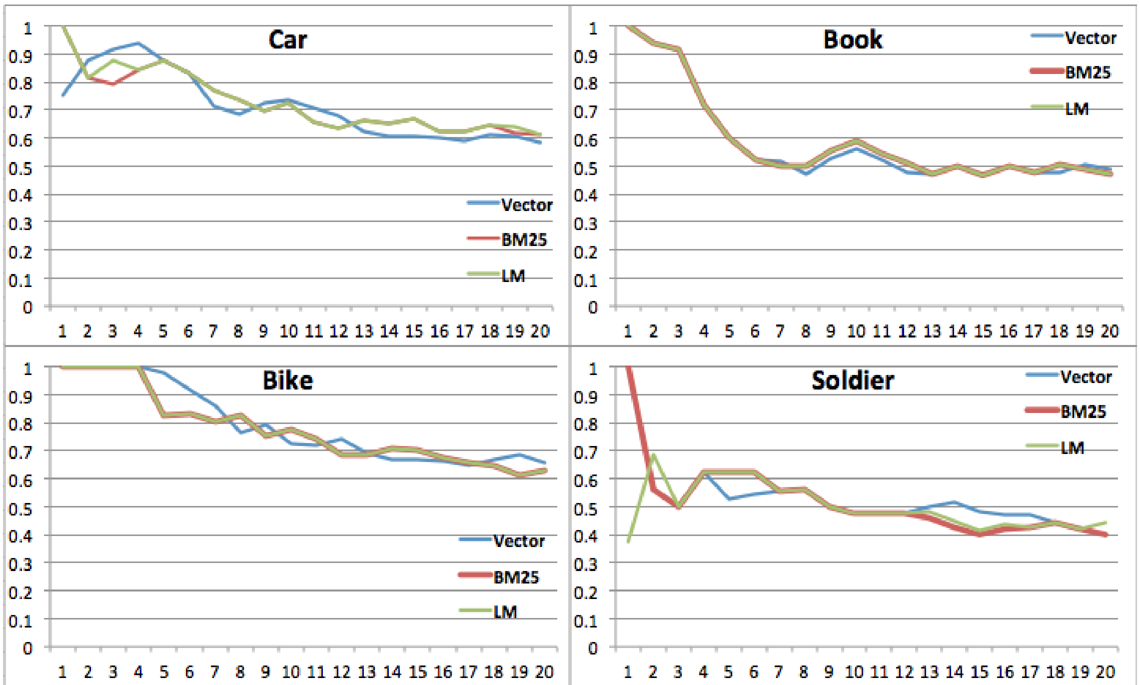

Figure 10: *Graph showing the CSMI* P@n *values as* n *rises to 20.*

In Figure 10 we see another set of terms where both the terms "car" and "bike" perform rather well. "Book" seems to perform well, though not convincing either. "Soldier" however takes a dive quite early on and reaches an average of around 0.55 almost instantly. Again, looking at the terms suggested (see Appendix B), we see a tendency of terms that suggests that articles about musical albums have made it into the set of articles used to extrapolate terms.

6.5.2 Results of Queries Using NT

This section covers the same data as the previous (regarding results of queries using CSMI). Again, the tables showing the specific numeric $P@n$ values can be found in Appendix C.

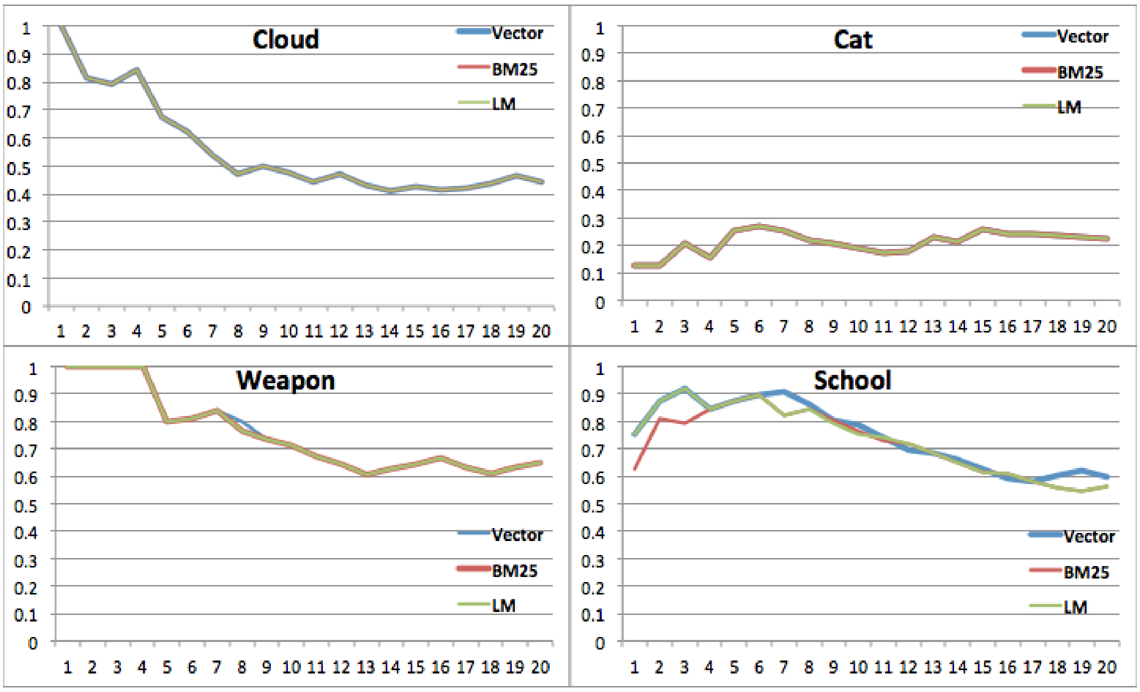

Figure 11: *Graph showing the NT* P@n *values as* n *rises to 20.*

In Figure 11 we can immediately see a difference from the CSMI approach. The most notable change is that the three similarity models seem to perform more or less the same for all queries. Both queries "cloud" and "weapon" perform rather as expected. They start off with strongly related hits and deteriorate as more terms are included.

The query "cat" however performs very poorly. It starts off with rather unrelated terms, and improves somewhat as terms are included, but never goes above 0.3. When looking at terms suggested (see Appendix A) for "cat" we see terms as "hanna", "barbera", "jerry", "produced", "directed", "fred" and "quimby". All of which are strongly related to the cartoon "Tom & Jerry", which is about a cat and a mouse. (Hanna Barbera (William Hanna and Joseph Barbera) and Fred Quimby are considered the creators of the series). Since these terms are related to a film, the terms "produced" and "directed" can also be seen as related to the rest. In other words, many of the terms suggested are in fact related to each other, but are unfortunately not strongly related to the concept of a cat.

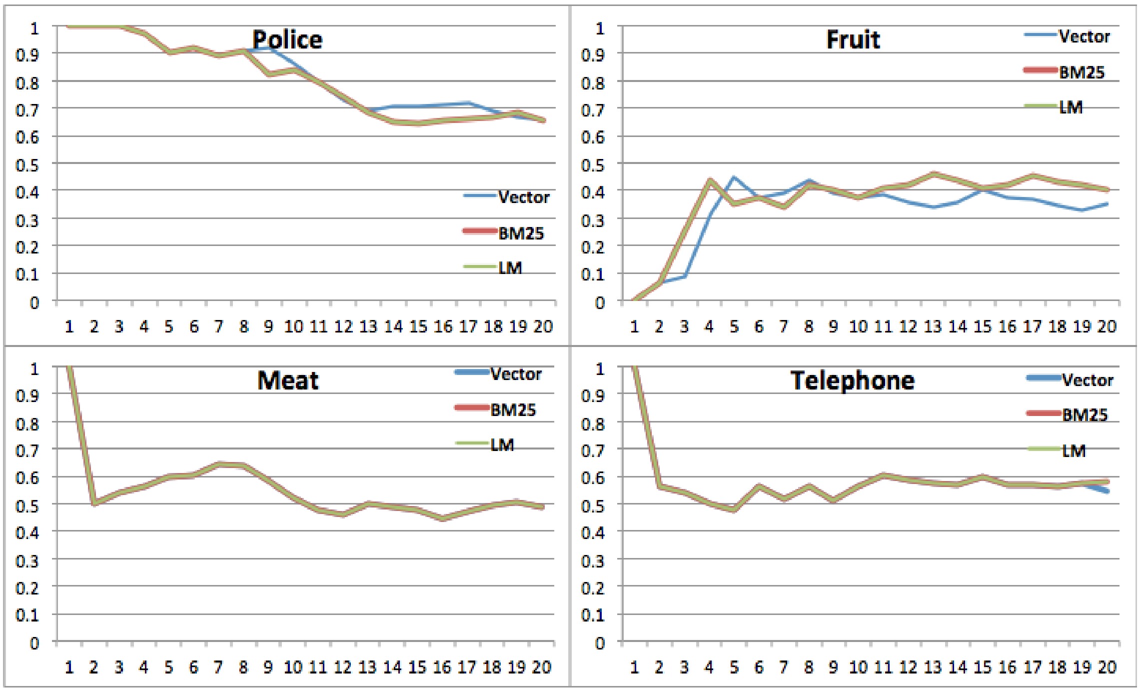

Figure 12: *Graph showing the NT* P@n *values as* n *rises to 20.*

Of all the queries in Figure 12, "fruit" is the one performing the poorest. It is the only query of all performed, both with CSMI as well as NT, where the first hit scores 0 in relatedness. The first suggested term here is "dove", which really has very little to do with fruit. The explanation here is most likely that the term "dove" has made it into this because of the dove variation called "fruit dove". Another term that has 0 relatedness score, but in the right context can be related is "ninja", which combined with "fruit" matches the iOS game "Fruit Ninja". Both of these are included into the base on which we extrapolate terms because their titles contain the term "fruit". Much like "meat" however, most articles about different kinds of fruits do not have "fruit" in the title. They often rather contain the specific name for a certain fruit, like "apple", "banana" and "pear".

"Telephone" and "meat" are both very similar here, giving a rather sudden drop in relatedness already on the second suggested term. After this it manages to stay more or less the same for the remainder of the list of terms. Another interesting event in these two queries is that for almost all numbers of added tags, the results are the same regardless of the similarity model used.

Of the four, "police" is the query performing the best here, with Vector Space performing slightly better than BM25 and LM.

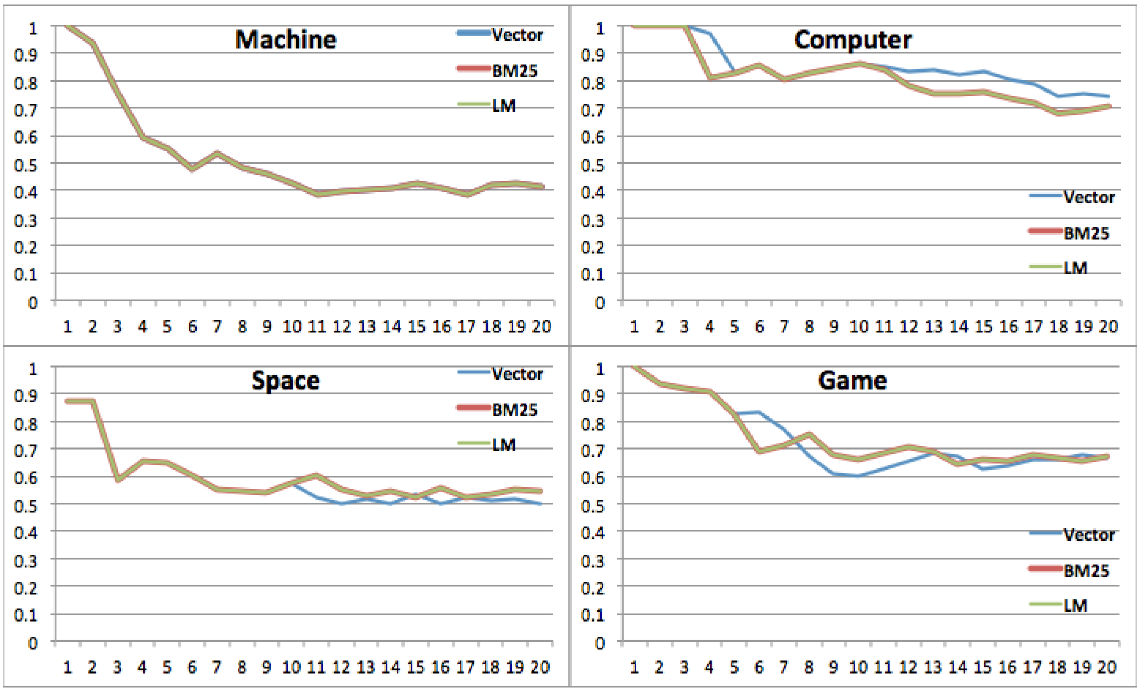

Figure 13: *Graph showing the NT* P@n *values as* n *rises to 20.*

The graph for "computer" in Figure 13 shows a fairly good result, with the Vector Space model averaging as the best result. "Game" in the same figure also performs quite well, though no single similarity model really stands out. Vector Space model breaks from the line formed by BM25 and LM, but on average it is more or less the same.

Of all the terms in this figure, "machine" performs the poorest. That said, it is still better than many of the other terms in this test, averaging at around 0.5. In fact, it is difficult to understand why some of the terms suggested by NT have appeared. Terms like "anaesthetic" and "darwin" seem very far away from the concept of machinery.

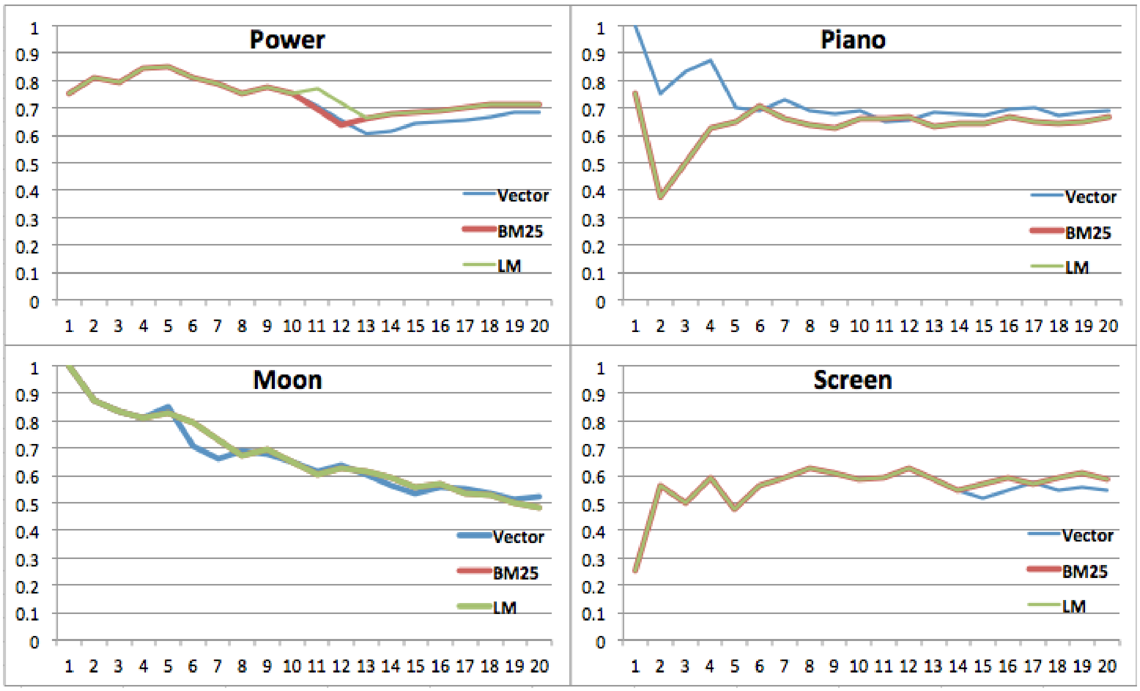

Figure 14: *Graph showing the NT* P@n *values as* n *rises to 20.*

Fig. 14 presents another set of four queries. Here, both "power" and "piano" performs perform fairly similar, however the Vector Space model seems to do much better than the other two similarity models for the query "piano". While the Vector Space model is able to keep most of its terms for "piano" related to a better degree, the other similarity models tend to include terms that are related to music and instruments, but not necessarily a piano. For example terms like "cello" and "violin".

"Screen" is probably the least scoring query of the four here, and starts off quite poorly. It evens out to a result that is not too bad after a few terms, but the poor start pulls the average score down quite a lot. It is rather hard to see how some of the terms are related; like "scottish", "bubble" and "gulf". Others can be seen as related with some context. The term "guild" is the first hit, which scores quite low for obvious reasons. This term has likely made its appearance because of the "Screen Actors Guild". Other terms by themselves are not very related, but joined together with another makes more sense. For example "motion" and "picture" where they individually do not score too well, but would likely score a lot better if they were presented as a whole (ie. "motion picture").

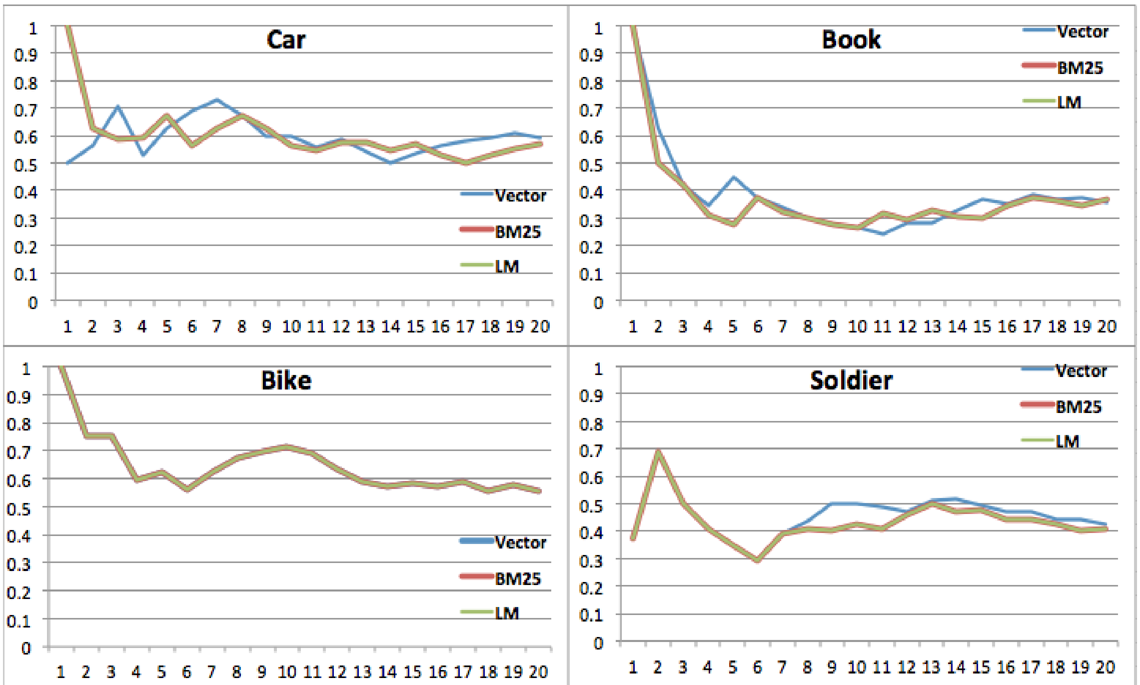

Figure 15: *Graph showing the NT* P@n *values as* n *rises to 20.*

The final set of queries can be seen in Figure 15. Here we see the queries for both "car" and "bike" performing more or less as expected. They start off with strongly related terms, and drop rather quickly down to a rough average of around 0.6 and 0.7. "Book" and "soldier" on the other hand do not perform quite so well. "Book" starts off strongly, but quickly drops to a rather low 0.4 and even 0.3.

"Soldier" starts off weak, then quickly grows before dropping again on the next term. This term too seems riddled with suggested terms that implies that many of the articles fetched are based on text regarding a music album or even a film. Suggested terms like "released", "game", "written", "song", and "buffalo" (likely from "buffalo soldier" by Bob Marley) tells us that this is very likely.

6.5.3 MAPs of CSMI and NT

In Table 3 we count how many times each similarity model has the best (or is tied for the best) value for each query. This means that for any given query, the similarity model with the best AP (Average Precision, all $P@n$ scores summed up and divided by n) gets a point. If two, or even three, similarity models has the same best score, then all two, or three, get a point.

Approach	Vector	BM25	LM
CSMI	7	10	5
NT	12	12	13

Table 3: *Times each similarity model has the best AP per query*

In the Table 4 we have gathered the calculated MAP values for each similarity model. We see that for CSMI, the best similarity model is the BM25 model (on average), while for NT the best is the Vector Space model (again on average).

Approach	Vector	BM25	LM
CSMI	0.647	0.661	0.659
NT	0.598	0.593	0.594

Table 4: MAP values for CSMI and NT

When viewing Tables 3 and 4 together, we see that BM25 was the best ranking similarity model for CSMI on average. Both in the Table 3, and in Table 4. However, for NT, the results are somewhat unexpected. While Vector Space scored the best in terms of MAP values, it was still tied last as the one with the least number of highest scoring APs for the set of queries. BM25 had the same number of highest scoring APs, but still also ranked at the bottom along with Vector Space. This indicates strongly that while Vector Space may not have been the best in terms of how many times it ranked highest, the times it did so it scored a high enough MAP to still rank on top.

Comparing CSMI's best (BM25) MAP of 0.661 with NT's best (Vector Space) MAP of 0.598 we see that CSMI also performs 0.063 points better. The difference is not remarkably large, but still better. That said, this is likely because of a deviance in difference values from the query "moon" specifically.

In Table 5 we can see the specific AP of each query using NT and CSMI for numerical comparison.

Query	NT (Vector)	CSMI (BM25)	Difference (NT - CSMI)
Cloud	0.555	0.647	-0.092
Cat	0.211	0.385	-0.174
Weapon	0.754	0.671	0.083
School	0.746	0.905	-0.159
Police	0.822	0.946	-0.124
Fruit	0.324	0.591	-0.267
Meat	0.55	0.487	0.063
Telephone	0.577	0.872	-0.295
Machine	0.514	0.717	-0.203
Computer	0.85	0.722	0.128
Space	0.58	0.566	0.014
Game	0.732	0.748	-0.016
Power	0.724	0.609	0.115
Piano	0.721	0.801	-0.08
Moon	0.67	0.23	0.44
Screen	0.55	0.724	-0.174
Car	0.593	0.724	-0.131
Book	0.386	0.589	-0.203
Bike	0.645	0.777	-0.132
Soldier	0.458	0.52	-0.062

Table 5: NT vs CSMI; AP values per query

If we look at the AP values in Table 5 for the query "moon", we see that the AP values for CSMI is 0.23, while NT's AP value is 0.675, giving a difference of 0.440 points. This is a rather significant difference that seem to stand out from the rest of the differences. As a comparison; the biggest difference in CSMI's favour is for "telephone" where CSMI is "only" 0.295 better.

Other queries where NT has a better AP than CSMI (besides "moon") are "weapon", "meat", "computer", "space" and "power". If we look at the cases where NT gives a better result than CSMI we see that they are on average 0.141 better. Vice versa, CSMI has an average value of 0.151 better than NT.

We can view "moon" as a statistical outlier [10], and could potentially ignore it, crediting chance as the main reason why it performed better. Outliers should however not be ignored as they can easily be correct as well [4].

However, even if we include "moon" in the comparison, CSMI does perform better than NT, if only just a little.

In Figure 16 and Figure 17 we see the $P@n$ values for all queries where vector space similarity has been used for NT, and BM25 has been used for CSMI.

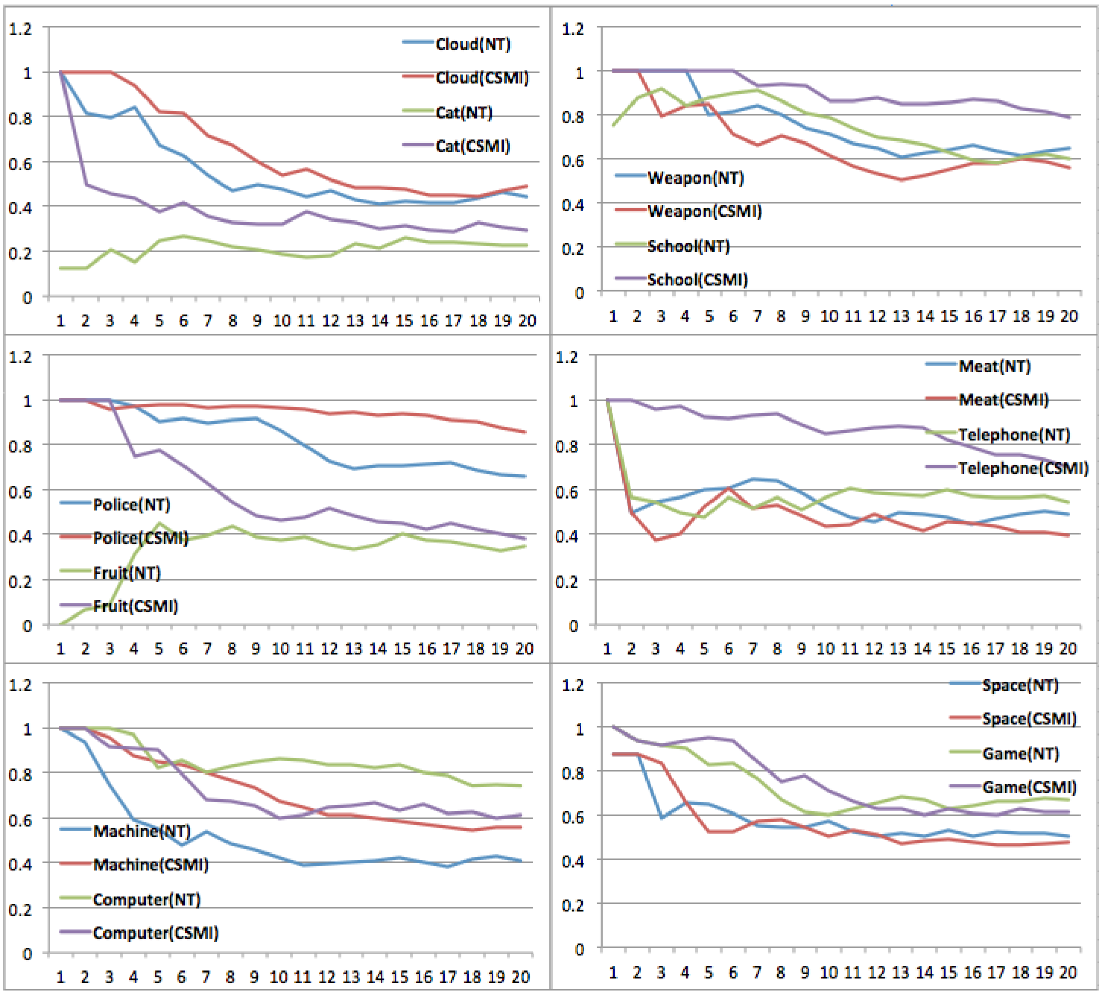

Figure 16: P@n values for queries using NT and CSMI approaches, part 1

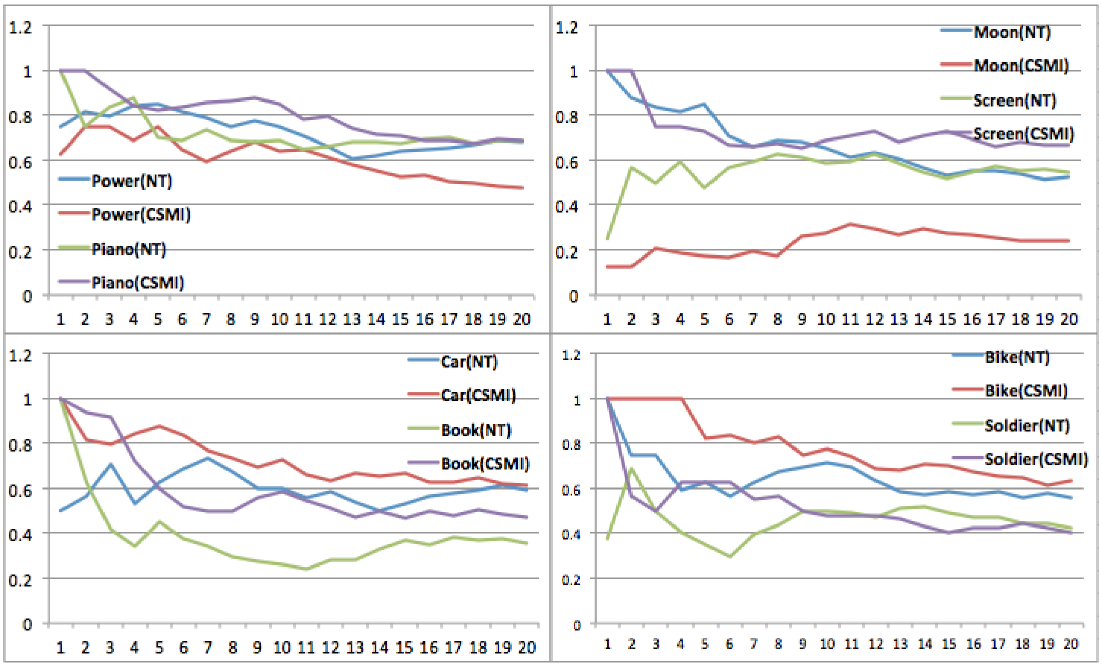

Figure 17: P@n values for queries using NT and CSMI approaches, part 2

These figures do not really tell us much new that was not already apparent, but they do give a good overview of how each query performs compared to the other, using their best case similarity models.

6.5.4 MAP@n: How Many Expansion Terms?

So far in this thesis we have operated with 20 expansion terms. This is not necessary, fruitful nor optimal in a practical solution. It is likely that having too many terms with insufficient relatedness will result in a poorly scoring query. This again may result in a poor image retrieval result. We gathered all AP (average precision) given at points $n = 5$, $n = 10$, $n = 15$ and $n = 20$, show in Table 6 and 7.

Query	AP@5	AP@10	AP@15	AP@20
Cloud	0.953	0.810	0.708	0.647
Cat	0.554	0.452	0.412	0.385
Weapon	0.897	0.784	0.701	0.671
School	1	0.966	0.930	0.905
Police	0.980	0.975	0.963	0.946
Fruit	0.905	0.735	0.649	0.591
Meat	0.561	0.538	0.510	0.487
Telephone	0.970	0.937	0.913	0.872
Machine	0.937	0.850	0.770	0.717
Computer	0.945	0.812	0.756	0.722
Space	0.753	0.648	0.598	0.566
Game	0.948	0.876	0.793	0.748
Power	0.713	0.676	0.645	0.609
Piano	0.917	0.886	0.840	0.801
Moon	0.164	0.189	0.223	0.230
Screen	0.845	0.756	0.741	0.724
Car	0.865	0.808	0.757	0.724
Book	0.835	0.684	0.622	0.589
Bike	0.965	0.882	0.822	0.777
Soldier	0.663	0.603	0.552	0.519
MAP@n	0.818	0.743	0.695	0.661

Table 6: CSMI; MAP@n values

Query	AP@5	AP@10	AP@15	AP@20
Cloud	0.825	0.673	0.594	0.554
Cat	0.173	0.200	0.203	0.211
Weapon	0.960	0.870	0.792	0.754
School	0.852	0.852	0.795	0.746
Police	0.974	0.936	0.866	0.822
Fruit	0.182	0.288	0.314	0.324
Meat	0.641	0.620	0.573	0.550
Telephone	0.616	0.580	0.582	0.577
Machine	0.766	0.621	0.549	0.514
Computer	0.959	0.899	0.878	0.850
Space	0.728	0.646	0.602	0.580
Game	0.917	0.807	0.755	0.732
Power	0.810	0.792	0.743	0.724
Piano	0.832	0.763	0.732	0.721
Moon	0.874	0.776	0.714	0.669
Screen	0.476	0.536	0.548	0.550
Car	0.585	0.622	0.595	0.593
Book	0.567	0.439	0.392	0.386
Bike	0.744	0.699	0.670	0.645
Soldier	0.464	0.444	0.461	0.458
MAP@n	0.697	0.653	0.618	0.598

Table 7: NT MAP@n values per query

First, when examining the table for CSMI (Table 6), we see the MAP values for query results with 5, 10, 15 and 20 expansion terms. As expected, the MAP values go lower the more expansion terms are included. Because we want a good precision (which can also be viewed as the general relatedness of the expanded terms), while also actually expanding the query, it comes down to a question of using a threshold that includes enough terms to be expansive, and still maintain a certain threshold of relatedness.

Because we do not have the resources to perform a large scale analysis of any image retrieval tests, we have opted to go for expanding queries with 5 terms, which will serve as a demonstration of what a search implementing this solution could work like.

6.6 Image Retrieval Using CSMI and NT

In this section we post some results using CSMI and NT as a searching system for looking for images based on tags. Each query is expanded by 5 extra terms, and returns 10 images and their tags. Tables for each query are presented in Appendix D, where each table show which terms were selected for a given query, a list of the ten retrieved images, whether or not the image is related to the query, and which of the queried tags the image itself is tagged with.

The following graphs are visualisations of these tables.

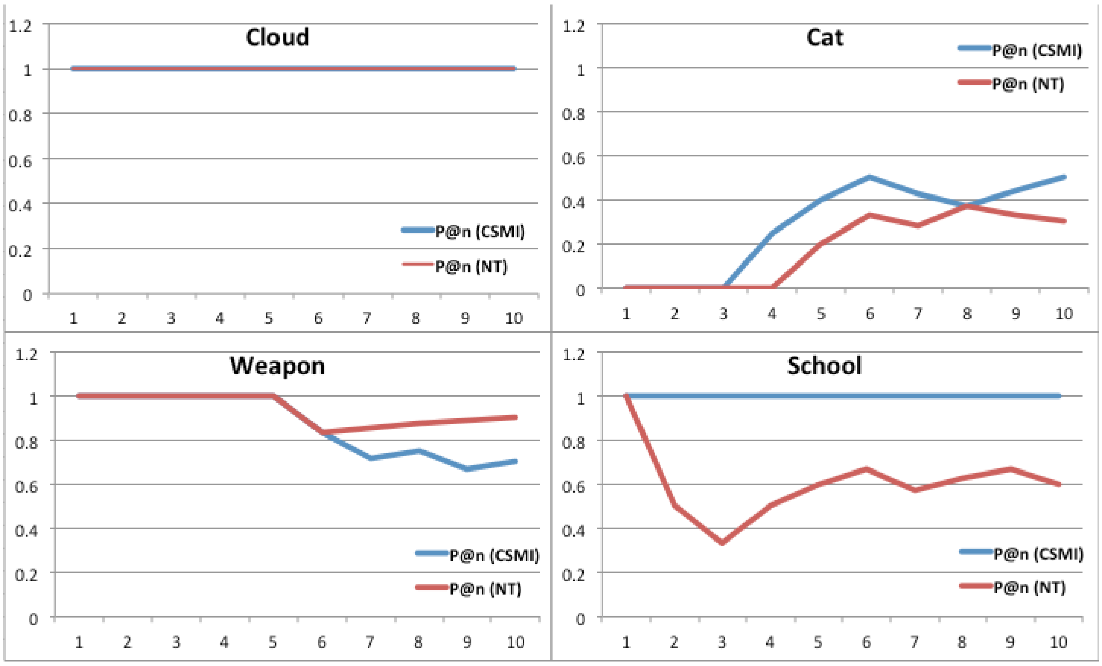

Figure 18: Visual representation of image retrieval queries using CSMI and NT

In the graphs shown in Figure 18 there are some rather diverse results. Both NT and CSMI perform perfectly for the query "Cloud". CSMI also performed perfectly with the query "school", but NT did not manage to find as many related images for that query. If we look at the images returned using NT the images returned did contain the tag "school", but not all images were taken at or of a school. Examples of such images were for example of a spanish sign on an abandoned building, as well as an image of kids on skateboards. The skateboarder image was tagged with "public" and "boarding", both of which can be combined with the term "school" to be about a type of school. The terms themselves are therefore related, but the ambiguity of the individual terms can cause non-relevancies like this.

"Cat" also performed rather poorly. No images of cats were given in the start, but more appeared as more images were included. Both CSMI and NT managed to pick out several images, but CSMI performed marginally better.

The query "weapon" was the only query among the four here which resulted in NT giving better results in the long run, compared to when using CSMI. The relevancy of the results can be debated as some of the images returned for both CSMI and NT were of toy weapons. We chose to define these images as related as they do represent a weapon, even though they are not lethal or harmful. NT returned only one unrelated image, which was tagged "weapons" and "nuclear", but was actually of a child with the peace icon painted on his face. An image clearly from an anti-WMD (Weapons of Mass Destruction) demonstration.

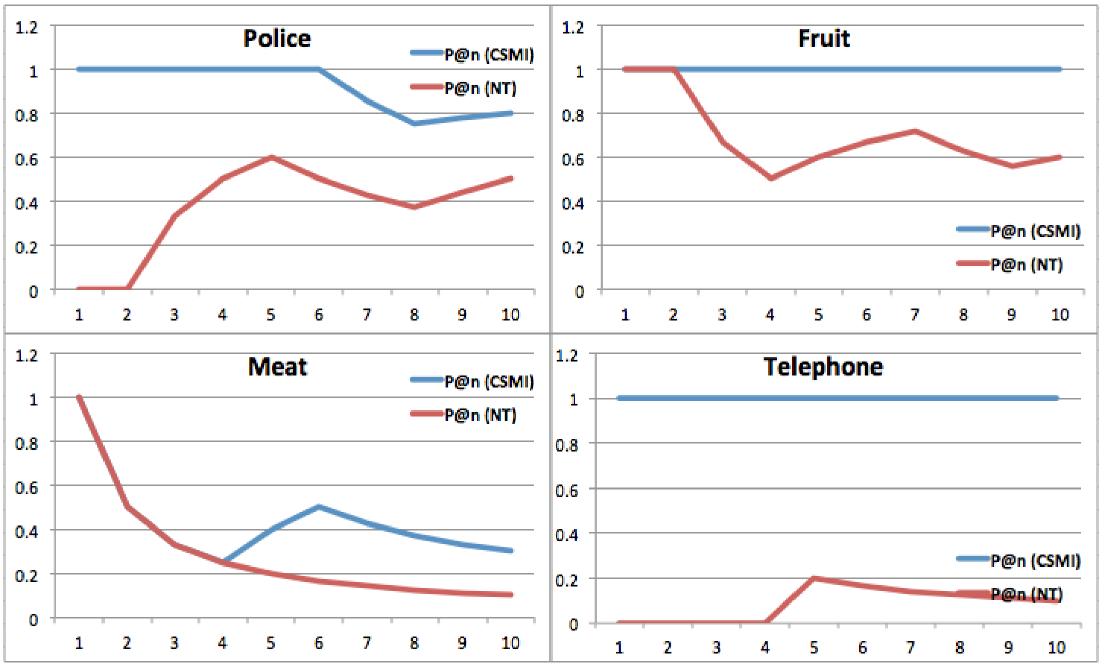

Figure 19: Visual representation of image retrieval queries using CSMI and NT

In Figure 19 we can see that CSMI generally speaking has outperformed NT in all four queries. CSMI perform perfectly for the query "fruit", while NT drops quite fast down to an average of around 0.6. CSMI also performs perfectly for the query "telephone". On this query however, NT performed very poorly with almost no related images. Many of the images were of different numbers on walls, doors and signs, but only one image was actually of a telephone. When looking at the level of relatedness the terms generated for "telephone" by NT, this is not very surprising as it did not score very well on many terms.

For the query "police" CSMI starts off at the top and only yielded a few images unrelated to the query. NT begins at the very bottom, despite the relatedness score of the expanded terms had scored high. This was in our opinion due to peculiar tagging by the user. The first two images were images of the same fire truck (however not a duplicate of the same picture), but was nevertheless tagged "police". Tag-wise it should have been related, but it was not.

"Meat" did not really perform well for any of the approaches. Both NT and CSMI managed to find a related image as the first image, but after this the relevance dropped, where only CSMI managed to pick out more related images later. This query may be a difficult query for any system as many would likely choose to tag an image of meat with something more specific. Something like "beef", "sausage" or even "dinner". None of which were generated as related terms.

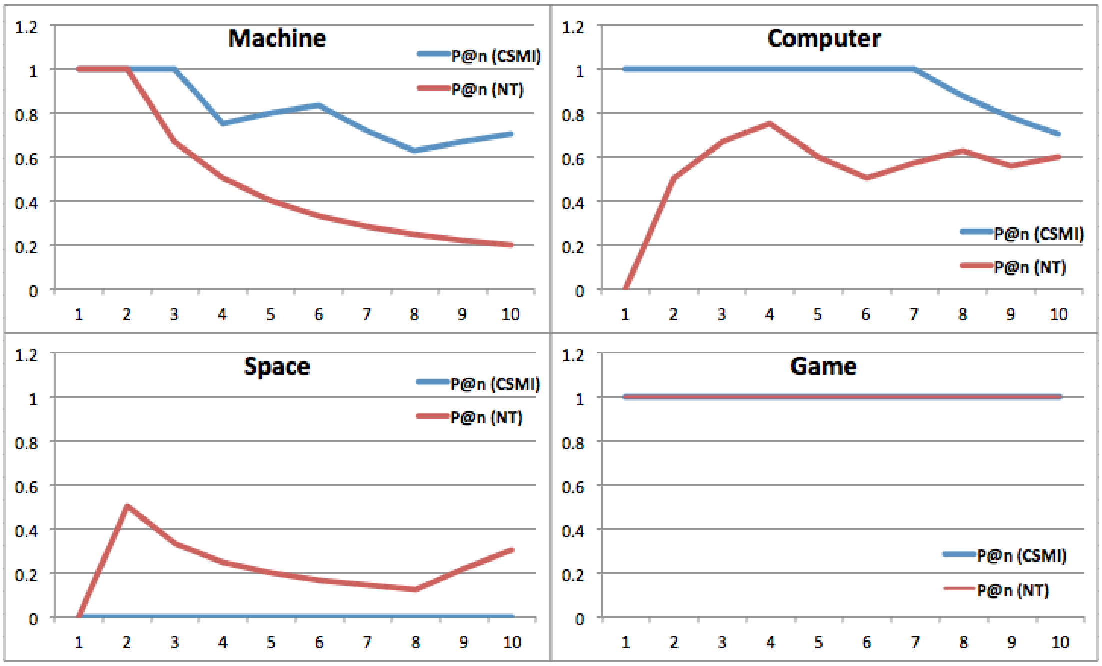

Figure 20: Visual representation of image retrieval queries using CSMI and NT

In Figure 20 there are not many new cases, but one change is for query "space". In this query, NT outperforms CSMI on all accounts. It does not perform very well, but compared to CSMI's average $P@n$s of 0 throughout the graph, it is far superior.

The query "machine" also has a notable differences, this time in benefit of CSMI. NT drops down and never really recovers after a few images, while CSMI continues to give related images on and off throughout the returned set of images.

"Computer" performs quite well when using CSMI, but even if NT manages to find several hits, it is pretty much outperformed by CSMI. Largely because of a poor start. The first image given by NT was an old portrait of a lady who, judging from the tags given, was an early mathematician and programmer. Her occupation would therefore be related, but there was nothing in the image directly related to a computer.

It is of course also worth noting that both NT and CSMI performed perfectly for the query "game".

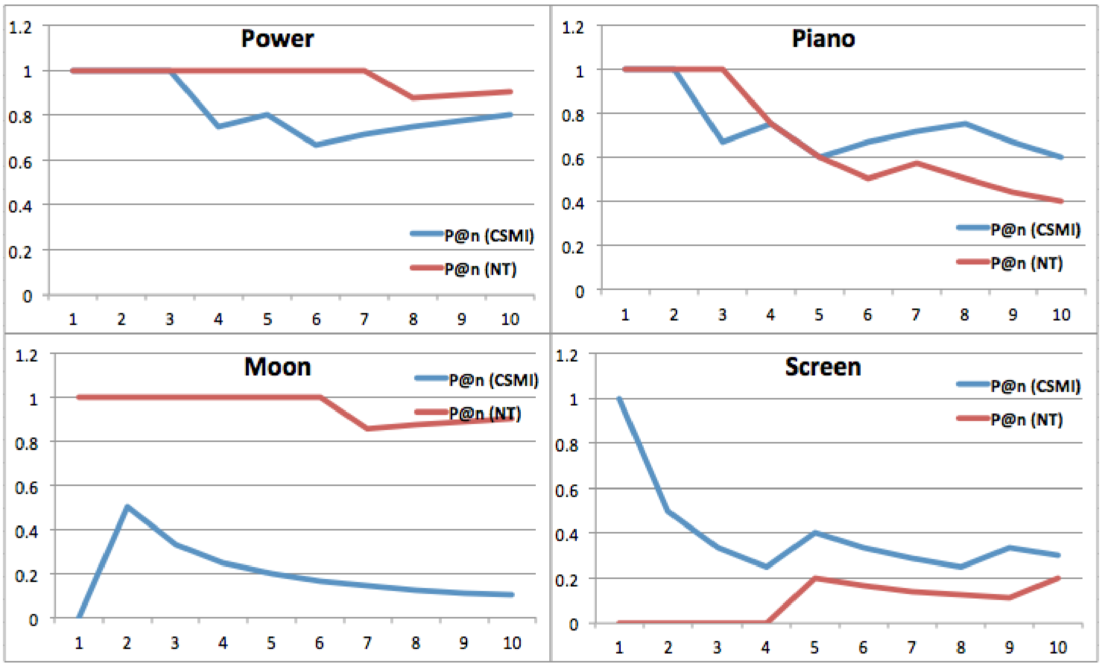

Figure 21: Visual representation of image retrieval queries using CSMI and NT

Figure 21 contains the first and only case of an NT based query completely outperforming CSMI. The query "moon" using CSMI only gives one related image, while with using NT we get an almost perfect result with only one unrelated image. This is also not very surprising judging from the difference in relatedness of the terms generated from NT and CSMI, which was about 0.8 for P@5.

The query "power" performs quite well for both NT and CSMI, but NT still gives more related images. The results for the query "piano" seems to be quite even, though CSMI stays on top the most. "Screen" is the case among the four where

CSMI performs quite better than NT, even though it still drops to a rather low score, averaging at around 0.3.

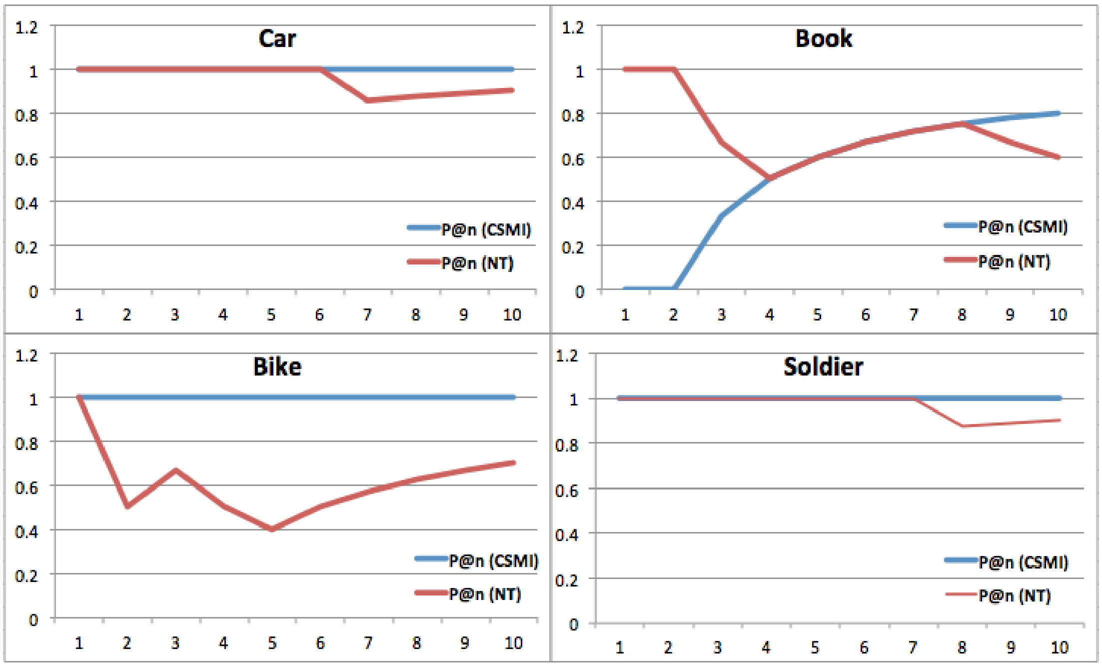

Figure 22: Visual representation of image retrieval queries using CSMI and NT

Of the graphs in Figure 22 the most notable one is "bike", where CSMI performs perfectly, while NT can be said to be unstable, averaging at around 0.6.

Both "car" and "soldier" perform quite well for both CSMI and NT. CSMI performs perfectly for both, while NT's score is stumped just short by one single image. The single unrelated image for "car" was again also an image that could be credited to poor tagging from the photographer. The image was of three girls at a car show. The event at which the photo was taken was in other words related, but because the image was tagged with "car" and "cars" despite not showing any cars, it was catalogued as an unrelated hit.

"Book" starts off very poorly for CSMI, but it continues to give related hits as more images are included and ends up actually outperforming NT despite its head start. The difference is admittedly slight.

When counting the number of relevant images in each of these searches presented in this section, we find the MAP values presented in Table 8.

Approach	MAP
CSMI	0.675
NT	0.585

Table 8: NT vs CSMI; MAP of Image Retrieval requests

We see that, as expected, CSMI performs better than NT. The queries where NT performed better was for "weapon", "space", "moon" and "power". These were also the terms that had better related terms after processing. Some terms are clearly more difficult than others, as is to be expected. All in all, after calculating scores, CSMI stands out as the best approach between the two, even if the difference is small.

6.7 Discussion

Summed up we find that queries for images, using CSMI, give better results than when using NT. There are some of tests resulting in exceptions in the test results, some of which must be said to be significant. One deviated so much it can be viewed as a statistical outlier, but not enough to disturb the overall result.

When viewing the results of using the top 5 generated terms in addition to the query in an image retrieval attempt, we also saw indications that this was again the case. The results from terms extrapolated using CSMI were generally speaking better than for NT with a few exceptions. As expected; in the cases where NT performed better in extrapolating terms, these were also the cases where the image retrieval performed better.

There does not seem to be much correlation between the cases where NT performed better than CSMI. Some queries were clearly more difficult than others, which may be contributed to by several factors. There may be few directly related articles that come up as hits when searching for texts to base the extrapolation on. There are uncertainties as some articles may be a match to the query, but regarding some other context. For example, terms commonly used in titles of music albums and movies, (like "power", "moon", "soldier" and "space"), are sources that seem to disrupt the relevance of the intension of a query.

Another problem that arises is one that seems impossible to solve. Certain images are tagged with terms that seem erroneous. The user tagging the image follows his/her own logic. E.g. "police" providing images of a fire truck). In could be, in the quest for visibility and hits, that users tag their images with a variety of terms that can make it reach the surface in the sea of images that is the Flickr database.

If we were to redo this thesis we would have done mostly everything the same way as we already did, but would perhaps have looked more thoroughly on certain variables like searching for titles vs. searching for abstracts. We would also have looked more into the problem of evaluating article contents before they are included into the set on which we performed related term extraction.

7 Conclusion

7.1 Summary of the Main Findings and Main Contributions

In this thesis we suggested two approaches (CSMI (Chi-Square Mutual Information) and NT (Neighbouring Terms)) to automatically generate terms that should be related to a given query. The user query would first be used to fetch a set of related documents from DBpedia. The text within these documents would then constitute the base on which CSMI and NT were used. The idea was to automatically define a list of terms that could help in locating images related to a query that would otherwise be overlooked in a tag based image retrieval system.

I evaluated many variables: For one we established what similarity model gave better results. In this analysis we found that for CSMI, the Okapi BM25 performed the most efficient. For NT, the best was the Vector Space model. It is worth noting that both BM25 and vector space only won by slim margins, but they were still the best performers for their respective similarity models.

I also looked into how many articles should constitute the base data set for term extraction. To do this we performed tests with different sizes of sets and found that CSMI worked best with 75 articles, and NT with 150 articles. Again, the differences were small, but still one size performed better than the others.

Another variable we looked into was whether it was best to search based on the title of the articles, or the contents of each article. My subjective analysis of tests on this again showed somewhat small differences, but querying the article titles performed better than querying the abstract texts of the articles.

Our conclusion regarding which approach was better of CSMI and NT: CSMI performed best with Okapi BM25 similarity model, basing the term extraction on the top 75 most relevant articles from DBpedia. NT performed best with

Vector Space similarity, basing the term extraction on the top 150 most relevant articles from DBpedia. Both approaches performed better when querying article titles.

When comparing the relatedness of the terms suggested by CSMI and NT, the MAP values show that, except for one case where NT performed significantly better, and some cases where NT performed somewhat better, CSMI was in general the best approach.

When we later used the best settings for each approach to do image retrieval, we found that the results from evaluating the term extraction was apparent in the image retrieval attempts as well. The queries that resulted in NT giving the most related terms, also resulted in NT giving more related images compared to that of CSMI. This was also the case the other way around. In the end, CSMI scored better on both extracted term relatedness as well as image retrieval precision.

Using the rather simple approaches described in this thesis, particularly using CSMI, we were able to see good results for IR queries that had been expanded by tags with high relatedness. Because it is very common for users of photo collections such as Flickr to tag their images with a wide variety of tags, it is advantageous to be able to include more than one related tag automatically. Our findings in this thesis show us that this is a viable approach that does work.

7.2 Future Work

7.2.1 Ignoring Similar Terms

One implementation that could potentially add to the versatility of term selection would be to ignore terms that are "too similar". For example, when the term "computer" has been selected as an expansion term, then the program could ignore terms like "computers". This would give room in the i most common terms for a different term. There is a risk that doing this could result in bringing a less related term into the set of expansion terms. It could, on the other hand, mean that the set of expansion terms would be more diverse and thus bring a wider set of results back from the search.

For example: Say that $i = 2$, and the two most common terms are "computer" and "computers". By removing the term "computers" the program makes space for the third most common term, which could be "laptop", thereby giving the set of expansion terms more diversity, however still relevant.

To try to get around the problem of too similar terms, one possible implementation could be to use a *Levenshtein Distance* measurement of terms, also called *edit distance.*

$$lev_{a,b}(i,j) = \begin{cases} max(i,j) & \text{if min(i,j) = 0,} \\ min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise,} \end{cases}$$

where $1_{(a_i \neq b_j)}$ is the indicator function equal to 0 when $a_i = b_j$ and 1 otherwise [8].

Using this, we could quickly eliminate the occurrence of such similar terms by dropping terms over a certain similarity threshold.

7.2.2 Handling Named Objects and Semantic Ambiguity

The definition database from DBpedia is very extensive. Wikipedia has an article about almost everything, which means that many terms are used in more than one article, and not always about the same thing. This can be both constructive and destructive for the potential efficiency of the system. For example: The term "dog" is a term used in popular culture (film, music, etc). Wikipedia contains specific articles about these things as well, which quickly results in a query set that returns terms like "film", "eat", "release", "band" and "TV". Most people searching for dog will in fact be looking for the animal.

It could be very interesting to create a set of articles about films, music albums etc. that we could use to perform some similarity check on, for example Cosine Similarity, given by

$$CosSim = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(A_i)^2}}$$

If we could find that a comparison between an article K and a set of articles about music albums yielded a similarity value over a certain threshold, we could then potentially avoid including these articles into a set of text that should describe something else.

Another problem is that of ambiguity. One term can mean two things, which can result in a selection that is not relevant to what the user had in mind. For example, in an early stage of creating the algorithm for this system we encountered a problem where the term "deer" resulted in a relevant term "musk". The term is a substance used in perfumes and can be argued to be relevant to "deer", because musk deers are a natural resource for this substance. However, a fairly high number of pictures of street graffiti have also been tagged with "musk", (likely because "musk" is used as a term for "signature" in street graffiti), more or less ruining the precision for searches on the term "deer".

A system for disambiguating terms to exclude terms that are not related *enough* could prove to give better precision.

7.2.3 Wolfram Language & System

3 days prior to this thesis' deadline, Wolfram released a demonstration of their new programming language which is a symbolic language-based system. This system contains a very large amount of processing functions, where output is fetched via a cloud system. One of the functions in this system is called "WordData", and performs a range of automatic term processors. Most common similarity calculations (such as cosine similarity, jaccard and levenschtein distance) are freely available, as well as a function to automatically retrieve broader terms, definitions or even synonyms of a given input [28].

It works by first creating a list of "senses" of the query, for example "fish". This example would return a list of strings, where each string contains the original term, what type of term the term can be viewed as (noun, verb, etc), along with what an alternative sense of the query can be. In this case output would give terms like "aquatic vertebrate", "food", "grab", "search".

Other functions of WordData uses this list in its evaluations. The synonym function takes each of the senses of a term and looks for synonyms for each of them. The previous example of "fish" would tell us that there are no available synonyms for "aquativ vertebrate", "food" or "grab", but would give us "angle" as a synonym for "search". The broader terms function would give a complete list of broader terms for each of the senses of the term. The output would for example tell us that broader terms for "grab" could be "catch", "grab", "take hold of", and for "search" could be; "look for", "search" and "seek".

It is also possible to get inflected forms of a term, so that the input "fish" would again use all senses of the term. The output would tell us that inflected forms of the verb "grab" in the sense of "fish", would be "fished", "fishing" and "fishes".

This seems like a very powerful and fast tool, but it is not yet released for use for the general public. It does however use publicly available data from public knowledge-bases, and performs very similar tasks as the ones we are doing in this thesis. It is very likely that this system can be implemented to perform exactly the tasks performed in this thesis, if not more.

7.2.4 Expanded Descriptions of Objects

I decided on using the abstracts of each article, as opposed to the entire body with the full article texts. This decision was made on the basis that many of the articles were very long. More text could possibly give more precise expansion. However, larger text collections would be less "concentrated" with related terms than the abstracts. It would be interesting to see results based on a much larger dataset per article retrieved.

7.2.5 Thorough Analysis of Image Retrieval

The analysis of the image retrieval in this thesis could be developed further by looking at different results using different numbers of expanded terms from CSMI and/or NT. It would also have been interesting to see if changing the similarity model for the tags as well could give more precise results.

References

[1] Nitish Aggarwal and Paul Buitelaar. "Query Expansion using Wikipedia and DBpedia". In: (2012).

[2] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval: The concepts and technology behind search*. 2nd. Pearson Education Limited, 2011.

[3] Hila Becker, Mor Naaman, and Luis Gravano. "Learning Similarity Metrics for Event Identification in Social Media". In: (2010).

[4] Shaun Burke. "Missing Values, Outliers, Robust Statistics & Non-parametric Methods". In: (2001).

[5] Kenneth Ward Church and Patrick Hanks. "Word Association Norms, Mutual Information, and Lexicography". In: (1990).

[6] W. Bruce Croft, Donald Metzler, and Trevor Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley, 2010.

[7] The Wikimedia Foundation. *Wikipedia*. 2013. URL: http://www.wikipedia.org.

[8] The Wikimedia Foundation. *Wikipedia, Levenshtein Distance*. 2014. URL: http://en.wikipedia.org/wiki/Levenshtein_distance.

[9] The Wikimedia Foundation. *Wikipedia, Mean Reciprocal Rank*. 2014. URL: http://en.wikipedia.org/wiki/Mean_reciprocal_rank.

[10] The Wikimedia Foundation. *Wikipedia, Outlier*. 2014. URL: http://en.wikipedia.org/wiki/Outlier.

[11] Daniel C. Howe. *RiTa.WordNet library*. 2013. URL: http://www.rednoise.org/rita/wordnet/documentation/index.htm.

[12] Mark Huiskes, Bart Thomee, and Michael Lew. *MIRFlickr*. 2013. URL: http://press.liacs.nl/mirflickr/.

[13] Glenn D. Israel. *Determining Sample Size*. 1992. URL: http://edis.ifas.ufl.edu/pd006#TABLE_1.

[14] Bill MacCartney. "NLP Launch Tutorial: Smoothing". In: (2005).

[15] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University, 2008.

[16] Phayung Meesad, Pudsadee Boonrawd, and Vatinee Nuipian. "A Chi-Square-Test for Word Importance Differentiation in Text Classification". In: (2011).

[17] Maciej Pacula. *Semantic <> Link: Automatically Find Related Words*. 2014. URL: http://www.semantic-link.com/how-it-works.html.

[18] Martin F. Porter. "An Algorithm for Suffix Stripping". In: (1980).

[19] Anand Rajaraman and Jeffrey David Ullman. *Mining of Massive Datasets*. Cambridge University Press, New York, 2012.

[20] Stephen Robertson and Hugo Zaragoza. "The Probabilistic Relevance Framework: BM25 and Beyond". In: (2009).

[21] Massimiliano Ruocco and Heri Ramampiaro. "A scalable algorithm for extraction and clustering of event-related pictures". In: (2010).

[22] Christopher Sahnwaldt. *DBPedia*. 2013. URL: http://www.DBpedia.org.

[23] Pavel Smrž and Pavel Rychlý. "Finding Semantically Related Words in Large Corpora". In: (2001).

[24] Apache Lucene Team. *Apache Lucene*. 2013. URL: http://lucene.apache.org/core/.

[25] Apache Lucene Team. *Apeche Lucene API*. 2014. URL: http://lucene.apache.org/core/4_0_0/analyzers-common/org/apache/lucene/analysis/core/WhitespaceTokenizer.html.

[26] The Flickr Team. *Flickr*. 2013. URL: http://www.flickr.com.

[27] webconfs.com. *Stop Words*. 2013. URL: http://www.webconfs.com/stop-words.php.

[28] Wolfram. *Wolfram Language & System, Documentation Center*. 2014. URL: http://reference.wolfram.com/language/ref/WordData.html.

[29] Chengxiang Zhai and John Lafferty. "A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval". In: (2001).

8 Appendix A: Neighbouring Terms: Relatedness of all terms from all queries for all sizes of N

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

VECTOR											
10 articles				**75 articles**				**150 articles**			
Cloud		**Term**	**Relatedness**	**Cloud**		**Term**	**Relatedness**	**Cloud**		**Term**	**Relatedness**
	1	solar	0.75		1	cloud	1		1	cloud	1
	2	system	0.25		2	point	0		2	computing	0.625
	3	cosmic	0.625		3	cirrus	0.75		3	cirrus	0.75
	4	rays	0.25		4	clouds	1		4	clouds	1
	5	bunster	0		5	form	0.375		5	point	0
	6	hill	0.25		6	club	0		6	form	0.375
	7	conjunction	0.25		7	solar	0.75		7	club	0
	8	occur	0.375		8	system	0.25		8	rats	0
	9	local	0.375		9	rats	0		9	solar	0.75
	10	supercluster	0.5		10	peak	0.125		10	system	0.25
	11	belt	0.375		11	actinoform	0.75		11	peak	0.125
	12	scattered	0.75		12	cosmic	0.625		12	actinoform	0.75
	13	cloud	1		13	rays	0.25		13	produce	0
	14	project	0.125		14	liquid	0.75		14	printing	0.125
	15	density	0.875		15	water	1		15	cosmic	0.625
	16	size	0.125		16	produce	0		16	rays	0.25
	17	view	0.5		17	accessory	0		17	communicati	0.5
	18	fire	0.125		18	forcing	0.125		18	liquid	0.75
	19	interstellar	0.375		19	generally	0.125		19	water	1
	20	medium	0.125		20	cover	0.5		20	accessory	0
			0.4				0.41875				0.44375
Cat		**Term**	**Relatedness**	**Cat**		**Term**	**Relatedness**	**Cat**		**Term**	**Relatedness**
	1	cat	0.875		1	population	0.375		1	hanna	0.125
	2	state	0.125		2	size	0		2	joseph	0.125
	3	born	0.125		3	breed	0.875		3	population	0.375
	4	wild	0.75		4	standard	0		4	size	0
	5	felis	0.5		5	effective	0		5	jerry	0.625
	6	silvestris	0.375		6	cat	0.875		6	short	0.375
	7	parts	0		7	prionailurus	0.5		7	barbera	0.125
	8	world	0.125		8	small	0.375		8	produced	0
	9	additional	0		9	wild	0.75		9	harley	0.125
	10	books	0		10	felis	0.5		10	person	0
	11	catus	0.125		11	bay	0		11	directed	0
	12	bay	0		12	black	0.625		12	william	0.25
	13	credits	0		13	created	0.125		13	breed	0.875
	14	cut	0		14	cats	1		14	standard	0
	15	alive	0.875		15	due	0		15	cat	0.875
	16	dead	0.5		16	south	0		16	cay	0
	17	anatomy	0.5		17	similar	0.25		17	meat	0.25
	18	felids	0.5		18	silvestris	0.375		18	fred	0.125
	19	eventually	0		19	concern	0.125		19	quimby	0.125
	20	pig	0		20	iucn	0		20	trees	0.125
			0.26875				0.3375				0.225
Weapon		**Term**	**Relatedness**	**Weapon**		**Term**	**Relatedness**	**Weapon**		**Term**	**Relatedness**
	1	nuclear	1		1	explosive	0.875		1	lethal	1
	2	weapons	1		2	weapons	1		2	weapon	1
	3	ranged	0.875		3	nuclear	1		3	nuclear	1
	4	anti-submari	1		4	assault	1		4	weapons	1
	5	weapon	1		5	weapon	1		5	bit	0
	6	bomb	1		6	conventional	0.25		6	explosive	0.875
	7	code-named	0.5		7	ranged	0.875		7	assault	1
	8	bombings	1		8	secret	0.25		8	fusion	0.5
	9	japan	0.125		9	melee	0.75		9	secret	0.25
	10	condensed	0		10	anti-aircraft	1		10	fission	0.5
	11	explosives	1		11	guns	1		11	conventional	0.25
	12	test	0.375		12	thermonucle	0.875		12	fissile	0.375
	13	ballistic	0.875		13	ceremonial	0		13	material	0.125
	14	missiles	1		14	improvised	0.25		14	ranged	0.875
	15	amount	0		15	edged	0.75		15	thermonucle	0.875
	16	energy	0.375		16	blunt	0.75		16	military	1
	17	considered	0		17	soviet	0.5		17	symbols	0.125
	18	deadly	1		18	union	0		18	air	0.25
	19	combination	0.125		19	cold	0		19	force	1
	20	fission	0.5		20	refer	0.125		20	rifle	1
			0.6375				0.6125				0.65

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

School		Term	Relatedness	School		Term	Relatedness	School		Term	Relatedness
	1	private	1		1	college	1		1	boarding	0.75
	2	schools	1		2	preparatory	0.625		2	school	1
	3	austrian	0		3	high	0.375		3	college	1
	4	school	1		4	school	1		4	preparatory	0.625
	5	primary	0.875		5	private	1		5	public	1
	6	newlyn	0		6	schools	1		6	schools	1
	7	annales	0.375		7	boarding	0.75		7	private	1
	8	jacques	0		8	junior	0.5		8	junior	0.5
	9	le	0.25		9	ranger	0.125		9	good	0.375
	10	cultural	0.125		10	independent	0.5		10	grammar	0.625
	11	history	0.25		11	austrian	0		11	boys	0.25
	12	revel	0		12	boys	0.25		12	girls	0.25
	13	college	1		13	girls	0.25		13	independent	0.5
	14	university	0.75		14	grammar	0.625		14	high	0.375
	15	subjective	0		15	middle	0.375		15	ibadan	0.125
	16	theory	0.5		16	medical	0.625		16	ranger	0.125
	17	public	1		17	han	0		17	united	0.375
	18	economic	0.5		18	porter-gaud	0.5		18	educational	1
	19	secondary	0.875		19	colleges	1		19	institution	0.875
	20	high	0.375		20	accredited	0.5		20	head	0.25
			0.49375				0.55				0.6

Police		Term	Relatedness	Police		Term	Relatedness	Police		Term	Relatedness
	1	force	1		1	police	1		1	police	1
	2	responsible	0.875		2	service	1		2	service	1
	3	fire	0.125		3	force	1		3	force	1
	4	police	1		4	responsible	0.875		4	responsible	0.875
	5	cleveland	0		5	power	0.75		5	chief	0.625
	6	officers	1		6	local	0.75		6	constable	1
	7	department	1		7	quest	0		7	local	0.75
	8	civic	0.125		8	carried	0.125		8	stations	1
	9	guard	0.5		9	kerala	0.125		9	crime	1
	10	command	0.5		10	mounted	0.25		10	prevention	0.375
	11	units	1		11	prosecutors	0.625		11	england	0.125
	12	slang	0		12	federal	0.5		12	wales	0
	13	terms	0		13	government	0.75		13	mounted	0.25
	14	activity	0.25		14	political	0.125		14	public	0.875
	15	policing	0.875		15	repression	0.375		15	safety	0.75
	16	service	1		16	municipal	0.375		16	power	0.75
	17	basic	0		17	station	1		17	services	0.875
	18	called	0.25		18	israel	0.125		18	british	0.125
	19	garda	0.25		19	dead	0.375		19	community	0.25
	20	chemical	0		20	kennedys	0		20	support	0.5
			0.4875				0.50625				0.65625

Fruit		Term	Relatedness	Fruit		Term	Relatedness	Fruit		Term	Relatedness
	1	hong	0		1	fruit	1		1	dove	0
	2	kong	0		2	ninja	0.125		2	ptilinopus	0.125
	3	fruit	1		3	preserves	0.75		3	bizarre	0.125
	4	sours	0.75		4	shoot	0.5		4	fruit	1
	5	accessory	0		5	juice	1		5	tree	1
	6	salad	0.75		6	snacks	0.75		6	iucn	0
	7	"False	0		7	tingles	0.25		7	red	0.5
	8	apples	1		8	gems	0.125		8	preserves	0.75
	9	pears	1		9	hong	0		9	list	0
	10	wines	0.875		10	kong	0		10	threatened	0.25
	11	pseudocarp	0		11	dried	1		11	shoot	0.5
	12	cocktail	0.75		12	fruits	1		12	evaluated	0
	13	fruits	1		13	varieties	0.625		13	ninja	0.125
	14	include	0		14	air	0		14	pie	0.625
	15	bruce	0		15	palaces	0		15	juice	1
	16	lee	0		16	roll-ups	0		16	roll-ups	0
	17	bitter	0.875		17	apple	1		17	tingles	0.25
	18	tangerine	0.625		18	orange	1		18	concern	0
	19	cabel	0		19	jams	1		19	islam	0
	20	hall	0		20	jellies	0.875		20	snacks	0.75
			0.43125				0.55				0.35

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

Meat	Term	Relatedness
1	meat	1
2	puppets	0
3	beef	1
4	salted	0.75
5	states	0
6	united	0
7	arizona	0
8	phoenix	0
9	bat	0
10	hell	0
11	includes	0
12	eaten	0.75
13	food	0.875
14	production	0.75
15	horse	0.625
16	salt	0.5
17	hard	0.25
18	rock	0
19	blending	0.25
20	punk	0
		0.3375

Meat	Term	Relatedness
1	meat	1
2	puppets	0
3	dog	0.25
4	cat	0.375
5	method	0.125
6	packing	0.625
7	raffle	0.25
8	slurry	0.125
9	kai	0
10	owen	0.375
11	production	0.75
12	goat	0.875
13	dried	0.875
14	consumption	1
15	air-dried	0.5
16	salted	0.75
17	slicer	1
18	parts	0.375
19	world	0
20	eating	0.875
		0.50625

Meat	Term	Relatedness
1	meat	1
2	puppets	0
3	products	0.625
4	pie	0.625
5	pies	0.75
6	packing	0.625
7	fresh	0.875
8	product	0.625
9	method	0.125
10	nova	0
11	scotia	0
12	raffle	0.25
13	slicer	1
14	cat	0.375
15	alternative	0.25
16	rock	0
17	goat	0.875
18	bone	0.875
19	meal	0.75
20	slurry	0.125
		0.4875

Telephone	Term	Relatedness
1	base	0.25
2	station	0.625
3	mobile	1
4	network	1
5	cellular	1
6	phone	1
7	cordless	0.875
8	phones	1
9	circuit	0.25
10	telephone	1
11	exchanges	0.25
12	manual	0.125
13	exchange	0.25
14	local	0.375
15	cables	0.5
16	transmission	0.5
17	connected	0.875
18	receiving	0.875
19	hybrid	0.25
20	building	0.125
		0.60625

Telephone	Term	Relatedness
1	mobile	1
2	telephone	1
3	financial	0.125
4	institution	0
5	local	0.375
6	enhanced	0.125
7	switched	0.125
8	public	0.375
9	area	0.5
10	code	0.375
11	bell	0.5
12	system	0.5
13	potomac	0
14	chesapeake	0
15	access	0.5
16	pioneer	0.25
17	field	0.25
18	telephones	1
19	emergency	0.75
20	base	0.25
		0.4

Telephone	Term	Relatedness
1	numbering	1
2	plan	0.125
3	area	0.5
4	codes	0.375
5	local	0.375
6	telephone	1
7	land	0.25
8	line	0.875
9	switched	0.125
10	mobile	1
11	phone	1
12	public	0.375
13	bell	0.5
14	system	0.5
15	call	1
16	prefix	0.125
17	central	0.5
18	office	0.5
19	model	0.75
20	england	0
		0.54375

Machine	Term	Relatedness
1	machine	1
2	pistols	0.375
3	translation	0.25
4	lisp	0
5	machines	1
6	explosive	0.375
7	rounds	0.25
8	pistol	0.375
9	automatic	0.75
10	rifle	0.375
11	state	0.375
12	time	0.25
13	ii	0
14	firearm	0.375
15	number	0.375
16	states	0
17	moving	0.375
18	parts	0.75
19	assembly	1
20	code	0.5
		0.4375

Machine	Term	Relatedness
1	machine	1
2	tools	0.875
3	iron	0.375
4	man	0.125
5	pistols	0.375
6	translation	0.25
7	darwin	0.125
8	adding	0.5
9	lisp	0
10	machines	1
11	artificial	0.5
12	intelligence	0.5
13	power	0.5
14	rangers	0
15	knowledge	0.25
16	american	0.125
17	football	0
18	pistol	0.375
19	flickering	0.5
20	ganzfeld	0
		0.36875

Machine	Term	Relatedness
1	machine	1
2	tools	0.875
3	iron	0.375
4	man	0.125
5	pistols	0.375
6	darwin	0.125
7	tool	0.875
8	anaesthetic	0.125
9	translation	0.25
10	cigarette	0.125
11	joint	0
12	adding	0.5
13	artificial	0.5
14	intelligence	0.5
15	models	0.625
16	abstract	0.125
17	lisp	0
18	machines	1
19	output	0.625
20	values	0.125
		0.4125

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

Computer	Term	Relatedness	Computer	Term	Relatedness	Computer	Term	Relatedness
1	computer	1	1	desktop	1	1	desktop	1
2	vision	0.25	2	software	1	2	software	1
3	programs	1	3	computer	1	3	computer	1
4	code	1	4	vision	0.25	4	systems	0.875
5	written	0.5	5	programs	1	5	vision	0.25
6	machine	0.875	6	analog	0.5	6	programs	1
7	bletchley	0.25	7	computers	1	7	analog	0.5
8	park	0.125	8	world	0.125	8	computers	1
9	analog	0.5	9	lab	0.375	9	data	1
10	computers	1	10	literacy	0.5	10	processing	1
11	application	0.875	11	systems	0.875	11	science	0.75
12	software	1	12	program	0.875	12	corporation	0.625
13	automated	0.875	13	engineers	0.625	13	access	0.875
14	image	0.75	14	data	1	14	information	0.625
15	applications	0.875	15	processing	1	15	technology	1
16	programming	1	16	science	0.75	16	lab	0.375
17	computing	1	17	forensics	0.375	17	literacy	0.5
18	science	0.75	18	icons	0.875	18	warriors	0
19	efficient	0.75	19	bletchley	0.25	19	program	0.875
20	evolvable	0.125	20	park	0.125	20	engineers	0.625
		0.725			0.675			0.74375

Space	Term	Relatedness	Space	Term	Relatedness	Space	Term	Relatedness
1	dual	0.25	1	sample	0	1	outer	0.875
2	space	0.875	2	space	0.875	2	space	0.875
3	euclidean	0.375	3	personal	0.625	3	sample	0
4	banach	0.25	4	dual	0.25	4	station	0.875
5	spaces	0.75	5	outer	0.875	5	personal	0.625
6	arbitrarily	0.25	6	metric	0.5	6	phase	0.375
7	close	0.25	7	warfare	0.25	7	dual	0.25
8	cotangent	0.5	8	euclidean	0.375	8	topological	0.5
9	continuous	0.625	9	station	0.875	9	vector	0.5
10	separated	0.125	10	spaces	0.75	10	international	0.875
11	metric	0.5	11	topological	0.5	11	tribe	0
12	topology	0.25	12	vector	0.5	12	banach	0.25
13	hilbert	0.25	13	probability	0.125	13	spaces	0.75
14	awareness	0.75	14	games	0.375	14	warfare	0.25
15	group	0.625	15	workshop	0.125	15	public	1
16	defined	0.375	16	contractible	0.25	16	hardy	0
17	real	0.625	17	homotopy	0.25	17	mathematics	0.875
18	alexandrov	0	18	type	0.125	18	fundamental	0.375
19	pavel	0.125	19	geodesic	0.125	19	group	0.625
20	man-kzin	0	20	triangle	0.25	20	branches	0.125
		0.3875			0.4			0.5

Game	Term	Relatedness	Game	Term	Relatedness	Game	Term	Relatedness
1	game	1	1	game	1	1	game	1
2	theory	0.625	2	players	0.875	2	players	0.875
3	card	0.875	3	games	0.875	3	games	0.875
4	casino	0.875	4	played	0.875	4	played	0.875
5	games	0.875	5	require	0.125	5	complete	0.5
6	person	0.5	6	electronic	0.75	6	genre	0.875
7	drinking	0.875	7	ender	0	7	combinatoria	0.375
8	board	1	8	center	0.5	8	tree	0
9	conclusion	0.375	9	arcade	0.875	9	require	0.125
10	rhyme	0.375	10	generally	0.25	10	involves	0.5
11	rules	0.875	11	algebraic	0.625	11	role-playing	0.875
12	genre	0.875	12	notation	0.125	12	video	1
13	computer	1	13	role-playing	0.875	13	player	1
14	played	0.875	14	video	1	14	genie	0.5
15	arcade	0.875	15	chessgames	0.875	15	ender	0
16	traditional	0.75	16	opening	0.5	16	arcade	0.875
17	combination:	0.75	17	perfect	0.375	17	board	1
18	outcomes	0.5	18	play	1	18	college	0.625
19	counting-out	0.5	19	genre	0.875	19	football	1
20	require	0.125	20	dr	0	20	family	0.5
		0.725			0.61875			0.66875

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

Power	Term	Relatedness	Power	Term	Relatedness	Power	Term	Relatedness
1	nuclear	1	1	market	0.75	1	electric	0.75
2	power	0.875	2	power	0.875	2	power	0.875
3	detector	1	3	vacuum	0.625	3	market	0.75
4	system	0.625	4	supply	1	4	supply	1
5	set	0.5	5	series	0.5	5	rating	0.875
6	love	0.625	6	electric	0.75	6	vacuum	0.625
7	number-one	0.375	7	nominal	0.5	7	transmission	0.625
8	r&b	0	8	nb	0	8	series	0.5
9	refers	0.375	9	steering	0.75	9	tools	1
10	signal	1	10	process	0.5	10	nominal	0.5
11	averaging	0.375	11	police	0.75	11	optical	0.25
12	snr	0.25	12	instruction	0.5	12	trios	0.125
13	electrical	1	13	set	0.5	13	nb	0
14	nep	0.25	14	transmission	0.625	14	motors	0.75
15	energy	1	15	projection	0.5	15	source	1
16	unit	1	16	hong	0	16	international	0.75
17	powerset	1	17	kong	0.125	17	relations	0.75
18	daiichi	0.125	18	tower	0.625	18	maximum	0.875
19	expressed	0.625	19	stone	0	19	energy	1
20	units	1	20	labour	0	20	stations	0.625
		0.65			0.49375			0.68125

Piano	Term	Relatedness	Piano	Term	Relatedness	Piano	Term	Relatedness
1	piano	1	1	piano	1	1	piano	1
2	tuning	0.875	2	roll	0.5	2	roll	0.5
3	systems	0.25	3	novelty	0.625	3	grand	1
4	attic	0.25	4	rolls	0.375	4	minor	1
5	floor	0.5	5	electric	1	5	op	0
6	larger	0.25	6	pianos	0.625	6	novelty	0.625
7	windows	0.125	7	grand	1	7	sonata	1
8	pianos	0.625	8	chamber	0.625	8	rolls	0.375
9	tuned	1	9	music	1	9	concerto	0.625
10	organ	0.75	10	sextet	0.625	10	quintet	0.75
11	reasons	0	11	digital	0.75	11	robert	0.25
12	rooms	0.25	12	tuning	0.875	12	schumann	0.75
13	classical	1	13	wind	0	13	electric	1
14	styles	0.625	14	electronic	0.875	14	pianos	0.625
15	sonatas	0.75	15	played	0.75	15	chamber	0.625
16	percussive	0.5	16	pianists	0.875	16	music	1
17	sound	1	17	playing	1	17	italian	0.75
18	noble	0.5	18	quintet	0.75	18	region	0.25
19	strings	0.875	19	produced	0.75	19	strings	0.875
20	spring	0.375	20	live	0.875	20	major	0.75
		0.575			0.74375			0.6875

Moon	Term	Relatedness	Moon	Term	Relatedness	Moon	Term	Relatedness
1	eclipses	0.875	1	moon	1	1	moon	1
2	occur	0.625	2	occurs	0.75	2	occurs	0.75
3	moon	1	3	eclipses	0.875	3	natural	0.75
4	occurs	0.75	4	occur	0.625	4	satellite	0.75
5	dark	1	5	blue	0.75	5	full	1
6	solar	0.875	6	apollo	0.875	6	short	0
7	system	0.5	7	mission	0.875	7	story	0.375
8	visible	0.75	8	aka	0	8	eclipses	0.875
9	sailor	0.375	9	held	0	9	occur	0.625
10	senshi	0.25	10	record	0.375	10	dean	0.375
11	eclipse	0.875	11	book	0.375	11	billion	0.25
12	astronomy	1	12	written	0.125	12	years	0.875
13	fuku	0.25	13	natural	0.75	13	dead	0.25
14	hemisphere	0.875	14	satellite	0.75	14	aka	0
15	full	1	15	exploration	1	15	wrasses	0.125
16	centered	0.375	16	research	0.875	16	solar	0.875
17	film	0.625	17	dark	1	17	system	0.5
18	industry	0.25	18	term	0.25	18	bbc	0.25
19	io	0.375	19	appearance	0.375	19	soap	0.125
20	satellite	0.75	20	december	0.5	20	blue	0.75
		0.66875			0.60625			0.525

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

Screen	Term	Relatedness	Screen	Term	Relatedness	Screen	Term	Relatedness
	1 rood	0.25		1 bubble	0.125		1 guild	0.25
	2 screens	1		2 screen	1		2 theater	0.875
	3 enlarged	0.625		3 projection	1		3 clear	0.375
	4 portion	0.375		4 scottish	0		4 view	0.875
	5 operating	0.5		5 sharing	0.125		5 fire	0
	6 systems	0.375		6 front	0.125		6 screen	1
	7 functional	0.5		7 snapshots	0.75		7 motion	0.75
	8 vision	0.75		8 ralph	0.125		8 picture	0.875
	9 screen	1		9 staub	0		9 operating	0.5
	10 smoke	0		10 screens	1		10 systems	0.375
	11 magnificatio	0.375		11 folding	0.25		11 type	0.625
	12 techniques	0.125		12 type	0.625		12 projection	1
	13 choir	0		13 film	0.875		13 bubble	0.125
	14 material	0.125		14 television	1		14 scottish	0
	15 mesh	0.25		15 recall	0.125		15 sharing	0.125
	16 openings	0.25		16 task	0.25		16 computer	1
	17 interest	0.375		17 independent	0.125		17 display	1
	18 user	0		18 production	0.5		18 front	0.125
	19 theory	0		19 motion	0.75		19 snapshots	0.75
	20 attached	0		20 picture	0.875		20 gulf	0.25
		0.34375			0.48125			0.54375

Car	Term	Relatedness	Car	Term	Relatedness	Car	Term	Relatedness
	1 baggage-dor	0.75		1 car	1		1 body	0.5
	2 cars	1		2 type	0.5		2 style	0.625
	3 cadillac	1		3 drag	1		3 cars	1
	4 fleetwood	0		4 racing	1		4 include	0
	5 car	1		5 phone	0.25		5 drag	1
	6 talk	0		6 body	0.5		6 racing	1
	7 amtrak	0.25		7 style	0.625		7 car	1
	8 received	0.125		8 automated	0.875		8 phone	0.25
	9 wheels	0.875		9 vehicles	1		9 surfing	0
	10 covered	0.125		10 cars	1		10 energy	0.625
	11 hoppers	0		11 include	0		11 water	0.125
	12 iron	0.375		12 caution	0.5		12 burning	0.875
	13 range	0.875		13 period	0		13 bait	0
	14 configured	0.5		14 motor	1		14 intended	0
	15 general	0.375		15 combustion	0.75		15 motor	1
	16 motors	0.875		16 engine	1		16 vehicles	1
	17 introduced	0		17 cadillac	1		17 rental	0.875
	18 lancia	0.5		18 fleetwood	0		18 combustion	0.75
	19 ledged	0.375		19 talk	0		19 engine	1
	20 rudimentum	0.125		20 mobile	0.75		20 leon	0.25
		0.45625			0.6375			0.59375

Book	Term	Relatedness	Book	Term	Relatedness	Book	Term	Relatedness
	1 book	1		1 book	1		1 book	1
	2 music	0.25		2 music	0.25		2 music	0.25
	3 print	0.875		3 review	0.75		3 lungs	0
	4 run	0		4 lungs	0		4 tokens	0.125
	5 sense	0.125		5 print	0.875		5 print	0.875
	6 tv	0.25		6 run	0		6 run	0
	7 series	0.875		7 high	0		7 closure	0.125
	8 reprinted	0.875		8 school	0.75		8 boldon	0
	9 additional	0.125		9 curse	0.125		9 curse	0.125
	10 copies	0.625		10 block	0.25		10 building	0.125
	11 a.m	0		11 books	1		11 high	0
	12 eastern	0.25		12 curses	0.125		12 school	0.75
	13 bookseller	1		13 sense	0.125		13 block	0.25
	14 art	0.625		14 burning	0.25		14 books	1
	15 form	0.125		15 collecting	0.5		15 series	0.875
	16 publisher	1		16 series	0.875		16 date	0.125
	17 time	0.125		17 cover	1		17 published	0.875
	18 called	0		18 considered	0		18 sense	0.125
	19 towns	0		19 elected	0.125		19 collecting	0.5
	20 books	1		20 artists	0.25		20 born	0
		0.45625			0.4125			0.35625

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

Bike	Term	Relatedness	Bike	Term	Relatedness	Bike	Term	Relatedness
1	star	0	1	bike	1	1	bike	1
2	wars	0	2	trail	1	2	park	0.5
3	bike	1	3	star	0	3	path	0.75
4	ms	0	4	wars	0	4	week	0.125
5	society	0.25	5	bus	0.375	5	paths	0.75
6	power	0.375	6	build	0	6	bar	0.25
7	speed	0.75	7	racing	0.875	7	bikes	1
8	multiple	0.25	8	paths	0.75	8	trail	1
9	sclerosis	0.125	9	affordable	0.875	9	racing	0.875
10	friday	0	10	housing	0	10	motorcycle	0.875
11	drive	0.625	11	ms	0	11	rally	0.5
12	train	0.375	12	arcata	0	12	star	0
13	events	0.125	13	club	0	13	wars	0
14	organized	0	14	head	0.125	14	train	0.375
15	logging	0	15	tube	0.5	15	road	0.75
16	roads	0.625	16	shop	0.625	16	bus	0.375
17	acceleration	0.75	17	path	0.75	17	ride	0.875
18	braking	1	18	society	0.25	18	build	0
19	frame	0.25	19	philly	0	19	bicycle	1
20	large	0.125	20	creative	0	20	sharing	0.125
		0.33125			0.35625			0.33125

Soldier	Term	Relatedness	Soldier	Term	Relatedness	Soldier	Term	Relatedness
1	cavalry	1	1	lone	0.375	1	lone	0.375
2	regiment	0.875	2	soldiers	1	2	soldiers	1
3	infantry	1	3	soldier	1	3	game	0.125
4	glen	0.125	4	action	0.5	4	released	0.125
5	matlock	0.125	5	figures	0.375	5	song	0.125
6	african-amer	0.125	6	good	0.25	6	written	0
7	regiments	0.625	7	century	0.125	7	soldier	1
8	cross-countr	0.125	8	toys	0.5	8	civil	0.75
9	skiing	0.125	9	vietnam	1	9	war	1
10	fictional	0	10	war	1	10	action	0.5
11	iggy	0	11	military	1	11	figures	0.375
12	pop	0	12	personnel	0.75	12	good	0.25
13	buffalo	0.125	13	cavalry	1	13	vietnam	1
14	soldiers	1	14	regiment	0.875	14	film	0.625
15	dead	0.75	15	set	0	15	starring	0.125
16	prez	0	16	world	0.375	16	century	0.125
17	army	1	17	buffalo	0.125	17	toys	0.5
18	domestic	0	18	korean	0.375	18	set	0
19	album	0	19	studio	0	19	world	0.375
20	lack	0	20	game	0.125	20	buffalo	0.125
		0.35			0.5375			0.425
		5			**7**			**8**

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

BM25

10 articles				75 articles				150 articles			
Cloud		**Term**	**Relatedness**	**Cloud**		**Term**	**Relatedness**	**Cloud**		**Term**	**Relatedness**
	1	solar	0.75		1	cloud	1		1	cloud	1
	2	system	0.25		2	point	0		2	computing	0.625
	3	cosmic	0.625		3	cirrus	0.75		3	cirrus	0.75
	4	rays	0.25		4	clouds	1		4	clouds	1
	5	bunster	0		5	form	0.375		5	point	0
	6	hill	0.25		6	solar	0.75		6	form	0.375
	7	conjunction	0.25		7	system	0.25		7	club	0
	8	occur	0.375		8	peak	0.125		8	rats	0
	9	local	0.375		9	actinoform	0.75		9	solar	0.75
	10	supercluster	0.5		10	cosmic	0.625		10	system	0.25
	11	belt	0.375		11	rays	0.25		11	peak	0.125
	12	scattered	0.75		12	liquid	0.75		12	actinoform	0.75
	13	cloud	1		13	water	1		13	produce	0
	14	project	0.125		14	produce	0		14	printing	0.125
	15	density	0.875		15	accessory	0		15	cosmic	0.625
	16	size	0.125		16	forcing	0.125		16	rays	0.25
	17	view	0.5		17	generally	0.125		17	communicati	0.5
	18	fire	0.125		18	cover	0.5		18	liquid	0.75
	19	interstellar	0.375		19	forms	0.5		19	water	1
	20	medium	0.125		20	fantasy	0.25		20	accessory	0
			0.4				0.45625				0.44375
Cat		**Term**	**Relatedness**	**Cat**		**Term**	**Relatedness**	**Cat**		**Term**	**Relatedness**
	1	cat	0.875		1	population	0.375		1	hanna	0.125
	2	state	0.125		2	size	0		2	joseph	0.125
	3	felis	0.5		3	breed	0.875		3	population	0.375
	4	silvestris	0.375		4	standard	0		4	size	0
	5	added	0		5	effective	0		5	jerry	0.625
	6	manufacture	0		6	cat	0.875		6	short	0.375
	7	species	1		7	prionailurus	0.5		7	barbera	0.125
	8	additional	0		8	small	0.375		8	produced	0
	9	books	0		9	wild	0.75		9	harley	0.125
	10	catus	0.125		10	felis	0.5		10	person	0
	11	alive	0.875		11	bay	0		11	directed	0
	12	dead	0.5		12	black	0.625		12	william	0.25
	13	harry	0.125		13	created	0.125		13	breed	0.875
	14	robert	0		14	cats	1		14	standard	0
	15	anatomy	0.5		15	due	0		15	cat	0.875
	16	felids	0.5		16	south	0		16	cay	0
	17	central	0		17	similar	0.25		17	meat	0.25
	18	regions	0		18	silvestris	0.375		18	fred	0.125
	19	fink	0		19	concern	0.125		19	quimby	0.125
	20	julian	0.125		20	iucn	0		20	trees	0.125
			0.28125				0.3375				0.225
Weapon		**Term**	**Relatedness**	**Weapon**		**Term**	**Relatedness**	**Weapon**		**Term**	**Relatedness**
	1	nuclear	1		1	explosive	0.875		1	lethal	1
	2	weapons	1		2	weapons	1		2	weapon	1
	3	ranged	0.875		3	nuclear	1		3	nuclear	1
	4	anti-submari	1		4	assault	1		4	weapons	1
	5	weapon	1		5	weapon	1		5	bit	0
	6	bomb	1		6	conventional	0.25		6	explosive	0.875
	7	code-named	0.5		7	ranged	0.875		7	assault	1
	8	bombings	1		8	secret	0.25		8	secret	0.25
	9	japan	0.125		9	melee	0.75		9	fusion	0.5
	10	condensed	0		10	anti-aircraft	1		10	fission	0.5
	11	explosives	1		11	guns	1		11	conventional	0.25
	12	test	0.375		12	thermonucle	0.875		12	fissile	0.375
	13	ballistic	0.875		13	ceremonial	0		13	material	0.125
	14	missiles	1		14	improvised	0.25		14	ranged	0.875
	15	amount	0		15	edged	0.75		15	thermonucle	0.875
	16	energy	0.375		16	blunt	0.75		16	military	1
	17	considered	0		17	soviet	0.5		17	symbols	0.125
	18	deadly	1		18	union	0		18	air	0.25
	19	combination	0.125		19	cold	0		19	force	1
	20	fission	0.5		20	refer	0.125		20	rifle	1
			0.6375				0.6125				0.65

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

School	Term	Relatedness	School	Term	Relatedness	School	Term	Relatedness
1	preparatory	0.625	1	junior	0.5	1	preparatory	0.625
2	school	1	2	school	1	2	school	1
3	junior	0.5	3	preparatory	0.625	3	boarding	0.75
4	private	1	4	high	0.375	4	college	1
5	schools	1	5	college	1	5	private	1
6	austrian	0	6	private	1	6	schools	1
7	grammar	0.625	7	schools	1	7	middle	0.375
8	newlyn	0	8	independent	0.5	8	public	1
9	annales	0.375	9	austrian	0	9	junior	0.5
10	jacques	0	10	boarding	0.75	10	high	0.375
11	le	0.25	11	middle	0.375	11	good	0.375
12	cultural	0.125	12	medical	0.625	12	grammar	0.625
13	history	0.25	13	han	0	13	boys	0.25
14	revel	0	14	porter-gaud	0.5	14	girls	0.25
15	college	1	15	academy	1	15	ibadan	0.125
16	university	0.75	16	status	0.125	16	prep	0.5
17	subjective	0	17	men	0	17	day	0.125
18	theory	0.5	18	women	0	18	ranger	0.125
19	economic	0.5	19	sixth	0.25	19	united	0.375
20	boarding	0.75	20	grammar	0.625	20	coeducationa	0.875
		0.4625			0.5125			0.5625

Police	Term	Relatedness	Police	Term	Relatedness	Police	Term	Relatedness
1	police	1	1	police	1	1	police	1
2	service	1	2	service	1	2	service	1
3	fire	0.125	3	force	1	3	force	1
4	response	0.25	4	responsible	0.875	4	responsible	0.875
5	unit	0.75	5	power	0.75	5	chief	0.625
6	separate	0.25	6	local	0.75	6	constable	1
7	units	1	7	quest	0	7	local	0.75
8	civic	0.125	8	carried	0.125	8	stations	1
9	guard	0.5	9	kerala	0.125	9	administratic	0.125
10	australian	0	10	british	0.125	10	crime	1
11	state	1	11	prosecutors	0.625	11	prevention	0.375
12	slang	0	12	rescue	0.375	12	england	0.125
13	terms	0	13	federal	0.5	13	wales	0
14	activity	0.25	14	government	0.75	14	mounted	0.25
15	policing	0.875	15	matlock	0.25	15	national	0.5
16	called	0.25	16	mounted	0.25	16	public	0.875
17	garda	0.25	17	political	0.125	17	safety	0.75
18	situations	0.25	18	repression	0.375	18	power	0.75
19	variety	0	19	municipal	0.375	19	station	1
20	coordinate	0.25	20	station	1	20	british	0.125
		0.40625			0.51875			0.65625

Fruit	Term	Relatedness	Fruit	Term	Relatedness	Fruit	Term	Relatedness
1	hong	0	1	fruit	1	1	dove	0
2	kong	0	2	ninja	0.125	2	ptilinopus	0.125
3	dried	1	3	preserves	0.75	3	subtropical	0.625
4	fruits	1	4	shoot	0.5	4	tropical	1
5	fruit	1	5	juice	1	5	columbidae	0
6	sours	0.75	6	tingles	0.25	6	family	0.5
7	apples	1	7	gems	0.125	7	bizarre	0.125
8	pears	1	8	hong	0	8	fruit	1
9	accessory	0	9	kong	0	9	habitat	0.25
10	salad	0.75	10	dried	1	10	loss	0.125
11	FALSE	0	11	fruits	1	11	preserves	0.75
12	pseudocarp	0	12	varieties	0.625	12	shoot	0.5
13	cocktail	0.75	13	air	0	13	tree	1
14	include	0	14	palaces	0	14	ninja	0.125
15	bruce	0	15	roll-ups	0	15	bird	0
16	lee	0	16	jams	1	16	pie	0.625
17	bitter	0.875	17	jellies	0.875	17	juice	1
18	tangerine	0.625	18	spring	0.625	18	roll-ups	0
19	cabel	0	19	water	0.125	19	tingles	0.25
20	hall	0	20	compound	0.25	20	islam	0
		0.4375			0.4625			0.4

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

Meat	Term	Relatedness	Meat	Term	Relatedness	Meat	Term	Relatedness
1	meat	1	1	meat	1	1	meat	1
2	puppets	0	2	puppets	0	2	puppets	0
3	products	0.625	3	dog	0.25	3	products	0.625
4	beef	1	4	cat	0.375	4	pie	0.625
5	salted	0.75	5	method	0.125	5	pies	0.75
6	states	0	6	packing	0.625	6	packing	0.625
7	united	0	7	raffle	0.25	7	fresh	0.875
8	arizona	0	8	slurry	0.125	8	product	0.625
9	phoenix	0	9	kai	0	9	method	0.125
10	bat	0	10	owen	0.375	10	nova	0
11	hell	0	11	production	0.75	11	scotia	0
12	includes	0	12	goat	0.875	12	raffle	0.25
13	cuts	0.625	13	dried	0.875	13	slicer	1
14	eaten	0.75	14	consumption	1	14	cat	0.375
15	food	0.875	15	air-dried	0.5	15	alternative	0.25
16	production	0.75	16	salted	0.75	16	rock	0
17	horse	0.625	17	slicer	1	17	goat	0.875
18	salt	0.5	18	parts	0.375	18	bone	0.875
19	hard	0.25	19	world	0	19	meal	0.75
20	rock	0	20	eating	0.875	20	slurry	0.125
		0.3875			0.50625			0.4875

Telephone	Term	Relatedness	Telephone	Term	Relatedness	Telephone	Term	Relatedness
1	cellular	1	1	mobile	1	1	numbering	1
2	phone	1	2	telephone	1	2	plan	0.125
3	circuit	0.25	3	financial	0.125	3	area	0.5
4	telephone	1	4	institution	0	4	codes	0.375
5	exchanges	0.25	5	local	0.375	5	local	0.375
6	manual	0.125	6	enhanced	0.125	6	telephone	1
7	mobile	1	7	switched	0.125	7	land	0.25
8	network	1	8	public	0.375	8	line	0.875
9	exchange	0.25	9	area	0.5	9	switched	0.125
10	local	0.375	10	code	0.375	10	mobile	1
11	cables	0.5	11	bell	0.5	11	phone	1
12	transmission	0.5	12	system	0.5	12	public	0.375
13	connected	0.875	13	potomac	0	13	bell	0.5
14	receiving	0.875	14	chesapeake	0	14	system	0.5
15	hybrid	0.25	15	access	0.5	15	call	1
16	building	0.125	16	pioneer	0.25	16	prefix	0.125
17	wires	0.625	17	field	0.25	17	central	0.5
18	converts	0.125	18	telephones	1	18	office	0.5
19	sound	0.625	19	emergency	0.75	19	emergency	0.75
20	directory	0	20	base	0.25	20	services	0.75
		0.5375			0.4			0.58125

Machine	Term	Relatedness	Machine	Term	Relatedness	Machine	Term	Relatedness
1	machine	1	1	machine	1	1	machine	1
2	pistols	0.375	2	tools	0.875	2	tools	0.875
3	translation	0.25	3	iron	0.375	3	iron	0.375
4	lisp	0	4	man	0.125	4	man	0.125
5	machines	1	5	pistols	0.375	5	pistols	0.375
6	explosive	0.375	6	translation	0.25	6	darwin	0.125
7	rounds	0.25	7	darwin	0.125	7	tool	0.875
8	pistol	0.375	8	adding	0.5	8	anaesthetic	0.125
9	automatic	0.75	9	lisp	0	9	translation	0.25
10	rifle	0.375	10	machines	1	10	cigarette	0.125
11	state	0.375	11	output	0.625	11	joint	0
12	time	0.25	12	artificial	0.5	12	adding	0.5
13	ii	0	13	intelligence	0.5	13	artificial	0.5
14	firearm	0.375	14	values	0.125	14	intelligence	0.5
15	number	0.375	15	current	0.25	15	models	0.625
16	states	0	16	state	0.375	16	abstract	0.125
17	moving	0.375	17	power	0.5	17	lisp	0
18	parts	0.75	18	rangers	0	18	machines	1
19	assembly	1	19	knowledge	0.25	19	output	0.625
20	code	0.5	20	american	0.125	20	values	0.125
		0.4375			0.39375			0.4125

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

Computer	Term	Relatedness
1	computer	1
2	vision	0.25
3	programs	1
4	code	1
5	written	0.5
6	machine	0.875
7	bletchley	0.25
8	park	0.125
9	analog	0.5
10	computers	1
11	application	0.875
12	software	1
13	automated	0.875
14	image	0.75
15	applications	0.875
16	programming	1
17	computing	1
18	science	0.75
19	efficient	0.75
20	evolvable	0.125
		0.725

Computer	Term	Relatedness
1	desktop	1
2	software	1
3	computer	1
4	vision	0.25
5	programs	1
6	analog	0.5
7	computers	1
8	world	0.125
9	lab	0.375
10	literacy	0.5
11	systems	0.875
12	program	0.875
13	engineers	0.625
14	data	1
15	processing	1
16	user	1
17	science	0.75
18	forensics	0.375
19	icons	0.875
20	bletchley	0.25
		0.71875

Computer	Term	Relatedness
1	desktop	1
2	software	1
3	computer	1
4	vision	0.25
5	systems	0.875
6	programs	1
7	analog	0.5
8	computers	1
9	data	1
10	processing	1
11	corporation	0.625
12	real	0.125
13	time	0.375
14	science	0.75
15	access	0.875
16	lab	0.375
17	literacy	0.5
18	warriors	0
19	program	0.875
20	network	1
		0.70625

Space	Term	Relatedness
1	dual	0.25
2	space	0.875
3	euclidean	0.375
4	banach	0.25
5	spaces	0.75
6	arbitrarily	0.25
7	close	0.25
8	cotangent	0.5
9	continuous	0.625
10	separated	0.125
11	metric	0.5
12	topology	0.25
13	hilbert	0.25
14	awareness	0.75
15	group	0.625
16	defined	0.375
17	real	0.625
18	alexandrov	0
19	pavel	0.125
20	man-kzin	0
		0.3875

Space	Term	Relatedness
1	sample	0
2	space	0.875
3	dual	0.25
4	outer	0.875
5	tribe	0
6	metric	0.5
7	warfare	0.25
8	station	0.875
9	spaces	0.75
10	intermembra	0.25
11	topological	0.5
12	vector	0.5
13	probability	0.125
14	games	0.375
15	workshop	0.125
16	earth	0.875
17	orbit	0.625
18	euclidean	0.375
19	contractible	0.25
20	geodesic	0.125
		0.425

Space	Term	Relatedness
1	outer	0.875
2	space	0.875
3	sample	0
4	station	0.875
5	personal	0.625
6	phase	0.375
7	dual	0.25
8	topological	0.5
9	vector	0.5
10	international	0.875
11	mathematics	0.875
12	tribe	0
13	banach	0.25
14	spaces	0.75
15	warfare	0.25
16	public	1
17	hardy	0
18	solar	0.75
19	system	0.875
20	fundamental	0.375
		0.54375

Game	Term	Relatedness
1	game	1
2	theory	0.625
3	card	0.875
4	casino	0.875
5	games	0.875
6	person	0.5
7	drinking	0.875
8	board	1
9	conclusion	0.375
10	rhyme	0.375
11	rules	0.875
12	genre	0.875
13	computer	1
14	played	0.875
15	arcade	0.875
16	traditional	0.75
17	combination	0.75
18	outcomes	0.5
19	counting-out	0.5
20	require	0.125
		0.725

Game	Term	Relatedness
1	game	1
2	players	0.875
3	games	0.875
4	played	0.875
5	require	0.125
6	electronic	0.75
7	ender	0
8	center	0.5
9	arcade	0.875
10	generally	0.25
11	algebraic	0.625
12	notation	0.125
13	role-playing	0.875
14	video	1
15	chessgames	0.875
16	opening	0.5
17	board	1
18	perfect	0.375
19	play	1
20	genre	0.875
		0.66875

Game	Term	Relatedness
1	game	1
2	players	0.875
3	games	0.875
4	played	0.875
5	complete	0.5
6	tree	0
7	genre	0.875
8	board	1
9	require	0.125
10	involves	0.5
11	role-playing	0.875
12	video	1
13	genie	0.5
14	ender	0
15	arcade	0.875
16	college	0.625
17	football	1
18	family	0.5
19	center	0.5
20	player	1
		0.675

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

Power	Term	Relatedness	Power	Term	Relatedness	Power	Term	Relatedness
	1 nuclear	1		1 market	0.75		1 market	0.75
	2 power	0.875		2 power	0.875		2 power	0.875
	3 detector	1		3 series	0.5		3 electric	0.75
	4 system	0.625		4 nominal	0.5		4 supply	1
	5 set	0.5		5 supply	1		5 rating	0.875
	6 love	0.625		6 nb	0		6 vacuum	0.625
	7 number-one	0.375		7 steering	0.75		7 transmission	0.625
	8 r&b	0		8 electric	0.75		8 series	0.5
	9 refers	0.375		9 process	0.5		9 tools	1
	10 signal	1		10 tower	0.625		10 nominal	0.5
	11 averaging	0.375		11 police	0.75		11 trios	0.125
	12 snr	0.25		12 instruction	0.5		12 nb	0
	13 electrical	1		13 set	0.5		13 source	1
	14 nep	0.25		14 stations	0.625		14 supplies	0.875
	15 energy	1		15 projection	0.5		15 international	0.75
	16 unit	1		16 hong	0		16 relations	0.75
	17 powerset	1		17 kong	0.125		17 maximum	0.875
	18 daiichi	0.125		18 stone	0		18 energy	1
	19 expressed	0.625		19 labour	0		19 stations	0.625
	20 units	1		20 energy	1		20 steering	0.75
		0.65			0.5125			0.7125

Piano	Term	Relatedness	Piano	Term	Relatedness	Piano	Term	Relatedness
	1 piano	1		1 piano	1		1 major	0.75
	2 tuning	0.875		2 roll	0.5		2 op	0
	3 systems	0.25		3 novelty	0.625		3 flat	0.75
	4 attic	0.25		4 rolls	0.375		4 piano	1
	5 floor	0.5		5 electric	1		5 trio	0.75
	6 larger	0.25		6 pianos	0.625		6 minor	1
	7 windows	0.125		7 grand	1		7 cello	0.375
	8 pianos	0.625		8 chamber	0.625		8 violin	0.5
	9 tuned	1		9 music	1		9 roll	0.5
	10 organ	0.75		10 sextet	0.625		10 grand	1
	11 reasons	0		11 digital	0.75		11 novelty	0.625
	12 rooms	0.25		12 tuning	0.875		12 trios	0.75
	13 classical	1		13 played	0.75		13 robert	0.25
	14 styles	0.625		14 wind	0		14 schumann	0.75
	15 sonatas	0.75		15 electronic	0.875		15 chamber	0.625
	16 percussive	0.5		16 classical	1		16 music	1
	17 sound	1		17 pianists	0.875		17 rolls	0.375
	18 noble	0.5		18 playing	1		18 concerto	0.625
	19 strings	0.875		19 quintet	0.75		19 quintet	0.75
	20 spring	0.375		20 produced	0.75		20 electric	1
		0.575			0.75			0.66875

Moon	Term	Relatedness	Moon	Term	Relatedness	Moon	Term	Relatedness
	1 eclipses	0.875		1 moon	1		1 moon	1
	2 occur	0.625		2 occurs	0.75		2 occurs	0.75
	3 moon	1		3 eclipses	0.875		3 natural	0.75
	4 occurs	0.75		4 occur	0.625		4 satellite	0.75
	5 dark	1		5 billion	0.25		5 eclipses	0.875
	6 solar	0.875		6 years	0.875		6 occur	0.625
	7 system	0.5		7 blue	0.75		7 dean	0.375
	8 visible	0.75		8 aka	0		8 billion	0.25
	9 sailor	0.375		9 term	0.25		9 years	0.875
	10 senshi	0.25		10 held	0		10 dead	0.25
	11 eclipse	0.875		11 record	0.375		11 wrasses	0.125
	12 astronomy	1		12 bbc	0.25		12 solar	0.875
	13 fuku	0.25		13 soap	0.125		13 system	0.5
	14 hemisphere	0.875		14 dark	1		14 bbc	0.25
	15 full	1		15 short	0		15 soap	0.125
	16 centered	0.375		16 story	0.375		16 blue	0.75
	17 film	0.625		17 eastenders	0		17 short	0
	18 industry	0.25		18 played	0.125		18 story	0.375
	19 io	0.375		19 apollo	0.875		19 eastenders	0
	20 satellite	0.75		20 mission	0.875		20 played	0.125
		0.66875			0.46875			0.48125

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

Screen	Term	Relatedness	Screen	Term	Relatedness	Screen	Term	Relatedness
1	rood	0.25	1	bubble	0.125	1	guild	0.25
2	screens	1	2	screen	1	2	theater	0.875
3	enlarged	0.625	3	film	0.875	3	clear	0.375
4	portion	0.375	4	television	1	4	view	0.875
5	functional	0.5	5	projection	1	5	fire	0
6	vision	0.75	6	scottish	0	6	screen	1
7	screen	1	7	sharing	0.125	7	motion	0.75
8	smoke	0	8	front	0.125	8	picture	0.875
9	magnificatio	0.375	9	screens	1	9	operating	0.5
10	techniques	0.125	10	type	0.625	10	systems	0.375
11	choir	0	11	recall	0.125	11	type	0.625
12	material	0.125	12	task	0.25	12	projection	1
13	mesh	0.25	13	loading	0.75	13	bubble	0.125
14	openings	0.25	14	motion	0.75	14	scottish	0
15	interest	0.375	15	picture	0.875	15	film	0.875
16	user	0	16	enlarged	0.625	16	television	1
17	theory	0	17	portion	0.375	17	sharing	0.125
18	attached	0	18	rood	0.25	18	computer	1
19	stencil	0	19	moving	0.625	19	display	1
20	print	0.5	20	middle	0.375	20	front	0.125
		0.325			0.54375			0.5875

Car	Term	Relatedness	Car	Term	Relatedness	Car	Term	Relatedness
1	body	0.5	1	body	0.5	1	car	1
2	style	0.625	2	style	0.625	2	finance	0.25
3	baggage-dor	0.75	3	car	1	3	body	0.5
4	cars	1	4	phone	0.25	4	style	0.625
5	cadillac	1	5	type	0.5	5	cars	1
6	fleetwood	0	6	bait	0	6	include	0
7	car	1	7	automated	0.875	7	drag	1
8	talk	0	8	vehicles	1	8	racing	1
9	amtrak	0.25	9	cars	1	9	phone	0.25
10	received	0.125	10	include	0	10	surfing	0
11	wheels	0.875	11	caution	0.5	11	general	0.375
12	carry	0.5	12	period	0	12	motors	0.875
13	town	0.125	13	motor	1	13	energy	0.625
14	configured	0.5	14	drag	1	14	water	0.125
15	general	0.375	15	racing	1	15	burning	0.875
16	motors	0.875	16	combustion	0.75	16	bait	0
17	introduced	0	17	engine	1	17	intended	0
18	lancia	0.5	18	model	0.875	18	motor	1
19	ledged	0.375	19	year	0.375	19	vehicles	1
20	rudimentum	0.125	20	area	0.125	20	rental	0.875
		0.475			0.61875			0.56875

Book	Term	Relatedness	Book	Term	Relatedness	Book	Term	Relatedness
1	book	1	1	book	1	1	book	1
2	music	0.25	2	music	0.25	2	soup	0
3	print	0.875	3	tokens	0.125	3	music	0.25
4	run	0	4	review	0.75	4	lungs	0
5	sense	0.125	5	lungs	0	5	tokens	0.125
6	series	0.875	6	print	0.875	6	print	0.875
7	reprinted	0.875	7	run	0	7	run	0
8	additional	0.125	8	curse	0.125	8	closure	0.125
9	copies	0.625	9	high	0	9	curse	0.125
10	publisher	1	10	school	0.75	10	building	0.125
11	bookseller	1	11	block	0.25	11	printed	0.875
12	art	0.625	12	books	1	12	high	0
13	form	0.125	13	curses	0.125	13	school	0.75
14	called	0	14	sense	0.125	14	born	0
15	towns	0	15	burning	0.25	15	block	0.25
16	books	1	16	collecting	0.5	16	books	1
17	mechanical	0.125	17	printed	0.875	17	series	0.875
18	organs	0	18	series	0.875	18	date	0.125
19	antiquarian	0.375	19	cover	1	19	generally	0
20	initial	0	20	considered	0	20	published	0.875
		0.45			0.44375			0.36875

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

Bike	Term	Relatedness	Bike	Term	Relatedness	Bike	Term	Relatedness
1	star	0	1	star	0	1	bike	1
2	wars	0	2	wars	0	2	park	0.5
3	bike	1	3	bike	1	3	path	0.75
4	ms	0	4	bus	0.375	4	week	0.125
5	society	0.25	5	build	0	5	paths	0.75
6	multiple	0.25	6	racing	0.875	6	bar	0.25
7	sclerosis	0.125	7	affordable	0.875	7	bikes	1
8	friday	0	8	housing	0	8	trail	1
9	drive	0.625	9	park	0.5	9	racing	0.875
10	train	0.375	10	ms	0	10	motorcycle	0.875
11	events	0.125	11	club	0	11	rally	0.5
12	organized	0	12	head	0.125	12	star	0
13	logging	0	13	tube	0.5	13	wars	0
14	roads	0.625	14	shop	0.625	14	train	0.375
15	frame	0.25	15	path	0.75	15	road	0.75
16	large	0.125	16	society	0.25	16	bus	0.375
17	maintaining	0.75	17	cycling	1	17	ride	0.875
18	dismounting	0.25	18	event	0.125	18	build	0
19	legs	0.75	19	philly	0	19	bicycle	1
20	track	0.875	20	creative	0	20	sharing	0.125
		0.31875			0.35			0.55625

Soldier	Term	Relatedness	Soldier	Term	Relatedness	Soldier	Term	Relatedness
1	cavalry	1	1	lone	0.375	1	lone	0.375
2	regiment	0.875	2	soldiers	1	2	soldiers	1
3	infantry	1	3	soldier	1	3	game	0.125
4	glen	0.125	4	action	0.5	4	released	0.125
5	matlock	0.125	5	figures	0.375	5	song	0.125
6	african-amer	0.125	6	good	0.25	6	written	0
7	regiments	0.625	7	century	0.125	7	soldier	1
8	cross-countr	0.125	8	toys	0.5	8	action	0.5
9	skiing	0.125	9	vietnam	1	9	figures	0.375
10	fictional	0	10	war	1	10	iron	0.625
11	iggy	0	11	military	1	11	good	0.25
12	pop	0	12	personnel	0.75	12	vietnam	1
13	buffalo	0.125	13	cavalry	1	13	war	1
14	soldiers	1	14	regiment	0.875	14	century	0.125
15	dead	0.75	15	set	0	15	toys	0.5
16	prez	0	16	world	0.375	16	set	0
17	army	1	17	buffalo	0.125	17	world	0.375
18	domestic	0	18	studio	0	18	buffalo	0.125
19	album	0	19	game	0.125	19	studio	0
20	lack	0	20	released	0.125	20	future	0.5
		0.35			0.525			0.40625
		5			**7**			**8**

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

LM												
10 articles				75 articles					150 articles			
Cloud		Term	Relatedness	Cloud		Term	Relatedness		Cloud		Term	Relatedness
	1	solar	0.75		1	cloud	1			1	cloud	1
	2	system	0.25		2	point	0			2	computing	0.625
	3	cosmic	0.625		3	cirrus	0.75			3	cirrus	0.75
	4	rays	0.25		4	clouds	1			4	clouds	1
	5	bunster	0		5	form	0.375			5	point	0
	6	hill	0.25		6	solar	0.75			6	form	0.375
	7	conjunction	0.25		7	system	0.25			7	club	0
	8	occur	0.375		8	peak	0.125			8	rats	0
	9	local	0.375		9	actinoform	0.75			9	solar	0.75
	10	supercluster	0.5		10	cosmic	0.625			10	system	0.25
	11	belt	0.375		11	rays	0.25			11	peak	0.125
	12	scattered	0.75		12	liquid	0.75			12	actinoform	0.75
	13	cloud	1		13	water	1			13	produce	0
	14	project	0.125		14	produce	0			14	printing	0.125
	15	density	0.875		15	accessory	0			15	cosmic	0.625
	16	size	0.125		16	forcing	0.125			16	rays	0.25
	17	view	0.5		17	generally	0.125			17	communicati	0.5
	18	fire	0.125		18	cover	0.5			18	liquid	0.75
	19	interstellar	0.375		19	forms	0.5			19	water	1
	20	medium	0.125		20	fantasy	0.25			20	accessory	0
			0.4				0.45625	0.45625				0.44375
Cat		Term	Relatedness	Cat		Term	Relatedness		Cat		Term	Relatedness
	1	cat	0.875		1	population	0.375			1	hanna	0.125
	2	state	0.125		2	size	0			2	joseph	0.125
	3	felis	0.5		3	breed	0.875			3	population	0.375
	4	silvestris	0.375		4	standard	0			4	size	0
	5	added	0		5	effective	0			5	jerry	0.625
	6	manufacture	0		6	cat	0.875			6	short	0.375
	7	species	1		7	prionailurus	0.5			7	barbera	0.125
	8	additional	0		8	small	0.375			8	produced	0
	9	books	0		9	wild	0.75			9	harley	0.125
	10	catus	0.125		10	felis	0.5			10	person	0
	11	alive	0.875		11	bay	0			11	directed	0
	12	dead	0.5		12	black	0.625			12	william	0.25
	13	harry	0.125		13	created	0.125			13	breed	0.875
	14	robert	0		14	cats	1			14	standard	0
	15	anatomy	0.5		15	due	0			15	cat	0.875
	16	felids	0.5		16	south	0			16	cay	0
	17	central	0		17	similar	0.25			17	meat	0.25
	18	regions	0		18	silvestris	0.375			18	fred	0.125
	19	fink	0		19	concern	0.125			19	quimby	0.125
	20	julian	0.125		20	iucn	0			20	trees	0.125
			0.28125				0.3375	0.3375				0.225
Weapon		Term	Relatedness	Weapon		Term	Relatedness		Weapon		Term	Relatedness
	1	nuclear	1		1	explosive	0.875			1	lethal	1
	2	weapons	1		2	weapons	1			2	weapon	1
	3	ranged	0.875		3	nuclear	1			3	nuclear	1
	4	anti-submari	1		4	assault	1			4	weapons	1
	5	weapon	1		5	weapon	1			5	bit	0
	6	bomb	1		6	conventional	0.25			6	explosive	0.875
	7	code-named	0.5		7	ranged	0.875			7	assault	1
	8	bombings	1		8	secret	0.25			8	secret	0.25
	9	japan	0.125		9	melee	0.75			9	fusion	0.5
	10	condensed	0		10	anti-aircraft	1			10	fission	0.5
	11	explosives	1		11	guns	1			11	conventional	0.25
	12	test	0.375		12	thermonucle	0.875			12	fissile	0.375
	13	ballistic	0.875		13	ceremonial	0			13	material	0.125
	14	missiles	1		14	improvised	0.25			14	ranged	0.875
	15	amount	0		15	edged	0.75			15	thermonucle	0.875
	16	energy	0.375		16	blunt	0.75			16	military	1
	17	considered	0		17	soviet	0.5			17	symbols	0.125
	18	deadly	1		18	union	0			18	air	0.25
	19	combination	0.125		19	cold	0			19	force	1
	20	fission	0.5		20	refer	0.125			20	rifle	1
			0.6375				0.6125	0.6125				0.65

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

School	Term	Relatedness	School	Term	Relatedness		School	Term	Relatedness
1	private	1	1	high	0.375		1	boarding	0.75
2	schools	1	2	school	1		2	school	1
3	austrian	0	3	middle	0.375		3	college	1
4	school	1	4	college	1		4	preparatory	0.625
5	primary	0.875	5	preparatory	0.625		5	private	1
6	newlyn	0	6	junior	0.5		6	schools	1
7	annales	0.375	7	private	1		7	middle	0.375
8	jacques	0	8	schools	1		8	public	1
9	le	0.25	9	boarding	0.75		9	high	0.375
10	cultural	0.125	10	austrian	0		10	good	0.375
11	history	0.25	11	grammar	0.625		11	grammar	0.625
12	revel	0	12	medical	0.625		12	junior	0.5
13	college	1	13	han	0		13	boys	0.25
14	university	0.75	14	porter-gaud	0.5		14	girls	0.25
15	subjective	0	15	academy	1		15	ibadan	0.125
16	theory	0.5	16	status	0.125		16	prep	0.5
17	public	1	17	men	0		17	day	0.125
18	economic	0.5	18	women	0		18	ranger	0.125
19	secondary	0.875	19	year	0.5		19	united	0.375
20	high	0.375	20	sixth	0.25		20	coeducationa	0.875
		0.49375			0.5125	0.5125			0.5625

Police	Term	Relatedness	Police	Term	Relatedness		Police	Term	Relatedness
1	police	1	1	police	1		1	police	1
2	service	1	2	service	1		2	service	1
3	fire	0.125	3	force	1		3	force	1
4	response	0.25	4	responsible	0.875		4	responsible	0.875
5	unit	0.75	5	power	0.75		5	chief	0.625
6	separate	0.25	6	local	0.75		6	constable	1
7	units	1	7	quest	0		7	local	0.75
8	civic	0.125	8	carried	0.125		8	stations	1
9	guard	0.5	9	kerala	0.125		9	administratic	0.125
10	australian	0	10	british	0.125		10	crime	1
11	state	1	11	prosecutors	0.625		11	prevention	0.375
12	slang	0	12	rescue	0.375		12	england	0.125
13	terms	0	13	federal	0.5		13	wales	0
14	activity	0.25	14	government	0.75		14	mounted	0.25
15	policing	0.875	15	matlock	0.25		15	national	0.5
16	called	0.25	16	mounted	0.25		16	public	0.875
17	garda	0.25	17	political	0.125		17	safety	0.75
18	situations	0.25	18	repression	0.375		18	power	0.75
19	variety	0	19	municipal	0.375		19	station	1
20	coordinate	0.25	20	station	1		20	british	0.125
		0.40625			0.51875	0.51875			0.65625

Fruit	Term	Relatedness	Fruit	Term	Relatedness		Fruit	Term	Relatedness
1	hong	0	1	fruit	1		1	dove	0
2	kong	0	2	ninja	0.125		2	ptilinopus	0.125
3	dried	1	3	preserves	0.75		3	subtropical	0.625
4	fruits	1	4	shoot	0.5		4	tropical	1
5	fruit	1	5	juice	1		5	columbidae	0
6	sours	0.75	6	tingles	0.25		6	family	0.5
7	apples	1	7	gems	0.125		7	bizarre	0.125
8	pears	1	8	hong	0		8	fruit	1
9	accessory	0	9	kong	0		9	habitat	0.25
10	salad	0.75	10	dried	1		10	loss	0.125
11	FALSE	0	11	fruits	1		11	preserves	0.75
12	pseudocarp	0	12	varieties	0.625		12	shoot	0.5
13	cocktail	0.75	13	air	0		13	tree	1
14	include	0	14	palaces	0		14	ninja	0.125
15	bruce	0	15	roll-ups	0		15	bird	0
16	lee	0	16	jams	1		16	pie	0.625
17	bitter	0.875	17	jellies	0.875		17	juice	1
18	tangerine	0.625	18	spring	0.625		18	roll-ups	0
19	cabel	0	19	water	0.125		19	tingles	0.25
20	hall	0	20	compound	0.25		20	islam	0
		0.4375			0.4625	0.4625			0.4

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

Meat	Term	Relatedness	Meat	Term	Relatedness	Meat	Term	Relatedness
1	meat	1	1	meat	1	1	meat	1
2	puppets	0	2	puppets	0	2	puppets	0
3	products	0.625	3	dog	0.25	3	products	0.625
4	beef	1	4	cat	0.375	4	pie	0.625
5	salted	0.75	5	method	0.125	5	pies	0.75
6	states	0	6	packing	0.625	6	packing	0.625
7	united	0	7	raffle	0.25	7	fresh	0.875
8	arizona	0	8	slurry	0.125	8	product	0.625
9	phoenix	0	9	kai	0	9	method	0.125
10	bat	0	10	owen	0.375	10	nova	0
11	hell	0	11	production	0.75	11	scotia	0
12	includes	0	12	goat	0.875	12	raffle	0.25
13	cuts	0.625	13	dried	0.875	13	slicer	1
14	eaten	0.75	14	consumption	1	14	cat	0.375
15	food	0.875	15	air-dried	0.5	15	alternative	0.25
16	production	0.75	16	salted	0.75	16	rock	0
17	horse	0.625	17	slicer	1	17	goat	0.875
18	salt	0.5	18	parts	0.375	18	bone	0.875
19	hard	0.25	19	world	0	19	meal	0.75
20	rock	0	20	eating	0.875	20	slurry	0.125
		0.3875			0.50625	0.50625		0.4875

Telephone	Term	Relatedness	Telephone	Term	Relatedness	Telephone	Term	Relatedness
1	cellular	1	1	mobile	1	1	numbering	1
2	phone	1	2	telephone	1	2	plan	0.125
3	circuit	0.25	3	financial	0.125	3	area	0.5
4	telephone	1	4	institution	0	4	codes	0.375
5	exchanges	0.25	5	local	0.375	5	local	0.375
6	manual	0.125	6	enhanced	0.125	6	telephone	1
7	mobile	1	7	switched	0.125	7	land	0.25
8	network	1	8	public	0.375	8	line	0.875
9	exchange	0.25	9	area	0.5	9	switched	0.125
10	local	0.375	10	code	0.375	10	mobile	1
11	cables	0.5	11	bell	0.5	11	phone	1
12	transmission	0.5	12	system	0.5	12	public	0.375
13	connected	0.875	13	potomac	0	13	bell	0.5
14	receiving	0.875	14	chesapeake	0	14	system	0.5
15	hybrid	0.25	15	access	0.5	15	call	1
16	building	0.125	16	pioneer	0.25	16	prefix	0.125
17	wires	0.625	17	field	0.25	17	central	0.5
18	converts	0.125	18	telephones	1	18	office	0.5
19	sound	0.625	19	emergency	0.75	19	emergency	0.75
20	directory	0	20	base	0.25	20	services	0.75
		0.5375			0.4	0.4		0.58125

Machine	Term	Relatedness	Machine	Term	Relatedness	Machine	Term	Relatedness
1	machine	1	1	machine	1	1	machine	1
2	pistols	0.375	2	tools	0.875	2	tools	0.875
3	translation	0.25	3	iron	0.375	3	iron	0.375
4	lisp	0	4	man	0.125	4	man	0.125
5	machines	1	5	pistols	0.375	5	pistols	0.375
6	explosive	0.375	6	translation	0.25	6	darwin	0.125
7	rounds	0.25	7	darwin	0.125	7	tool	0.875
8	pistol	0.375	8	adding	0.5	8	anaesthetic	0.125
9	automatic	0.75	9	lisp	0	9	translation	0.25
10	rifle	0.375	10	machines	1	10	cigarette	0.125
11	state	0.375	11	output	0.625	11	joint	0
12	time	0.25	12	artificial	0.5	12	adding	0.5
13	ii	0	13	intelligence	0.5	13	artificial	0.5
14	firearm	0.375	14	values	0.125	14	intelligence	0.5
15	number	0.375	15	current	0.25	15	models	0.625
16	states	0	16	state	0.375	16	abstract	0.125
17	moving	0.375	17	power	0.5	17	lisp	0
18	parts	0.75	18	rangers	0	18	machines	1
19	assembly	1	19	knowledge	0.25	19	output	0.625
20	code	0.5	20	american	0.125	20	values	0.125
		0.4375			0.39375	0.39375		0.4125

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

Computer	Term	Relatedness		Computer	Term	Relatedness		Computer	Term	Relatedness
1	computer	1		1	desktop	1		1	desktop	1
2	vision	0.25		2	software	1		2	software	1
3	programs	1		3	computer	1		3	computer	1
4	code	1		4	vision	0.25		4	vision	0.25
5	written	0.5		5	programs	1		5	systems	0.875
6	machine	0.875		6	analog	0.5		6	programs	1
7	bletchley	0.25		7	computers	1		7	analog	0.5
8	park	0.125		8	world	0.125		8	computers	1
9	analog	0.5		9	lab	0.375		9	data	1
10	computers	1		10	literacy	0.5		10	processing	1
11	application	0.875		11	systems	0.875		11	corporation	0.625
12	software	1		12	program	0.875		12	real	0.125
13	automated	0.875		13	engineers	0.625		13	time	0.375
14	image	0.75		14	data	1		14	science	0.75
15	applications	0.875		15	processing	1		15	access	0.875
16	programmin	1		16	user	1		16	lab	0.375
17	computing	1		17	science	0.75		17	literacy	0.5
18	science	0.75		18	forensics	0.375		18	warriors	0
19	efficient	0.75		19	icons	0.875		19	program	0.875
20	evolvable	0.125		20	bletchley	0.25		20	network	1
		0.725				0.71875	0.71875			0.70625

Space	Term	Relatedness		Space	Term	Relatedness		Space	Term	Relatedness
1	dual	0.25		1	sample	0		1	outer	0.875
2	space	0.875		2	space	0.875		2	space	0.875
3	euclidean	0.375		3	dual	0.25		3	sample	0
4	banach	0.25		4	outer	0.875		4	station	0.875
5	spaces	0.75		5	tribe	0		5	personal	0.625
6	arbitrarily	0.25		6	metric	0.5		6	phase	0.375
7	close	0.25		7	warfare	0.25		7	dual	0.25
8	cotangent	0.5		8	station	0.875		8	topological	0.5
9	continuous	0.625		9	spaces	0.75		9	vector	0.5
10	separated	0.125		10	intermembra	0.25		10	international	0.875
11	metric	0.5		11	topological	0.5		11	mathematics	0.875
12	topology	0.25		12	vector	0.5		12	tribe	0
13	hilbert	0.25		13	probability	0.125		13	banach	0.25
14	awareness	0.75		14	games	0.375		14	spaces	0.75
15	group	0.625		15	workshop	0.125		15	warfare	0.25
16	defined	0.375		16	earth	0.875		16	public	1
17	real	0.625		17	orbit	0.625		17	hardy	0
18	alexandrov	0		18	euclidean	0.375		18	solar	0.75
19	pavel	0.125		19	contractible	0.25		19	system	0.875
20	man-kzin	0		20	geodesic	0.125		20	fundamental	0.375
		0.3875				0.425	0.425			0.54375

Game	Term	Relatedness		Game	Term	Relatedness		Game	Term	Relatedness
1	game	1		1	game	1		1	game	1
2	theory	0.625		2	players	0.875		2	players	0.875
3	card	0.875		3	games	0.875		3	games	0.875
4	casino	0.875		4	played	0.875		4	played	0.875
5	games	0.875		5	require	0.125		5	complete	0.5
6	person	0.5		6	electronic	0.75		6	tree	0
7	drinking	0.875		7	ender	0		7	genre	0.875
8	board	1		8	center	0.5		8	board	1
9	conclusion	0.375		9	arcade	0.875		9	require	0.125
10	rhyme	0.375		10	generally	0.25		10	involves	0.5
11	rules	0.875		11	algebraic	0.625		11	role-playing	0.875
12	genre	0.875		12	notation	0.125		12	video	1
13	computer	1		13	role-playing	0.875		13	genie	0.5
14	played	0.875		14	video	1		14	ender	0
15	arcade	0.875		15	chessgames	0.875		15	arcade	0.875
16	traditional	0.75		16	opening	0.5		16	college	0.625
17	combination	0.75		17	board	1		17	football	1
18	outcomes	0.5		18	perfect	0.375		18	family	0.5
19	counting-out	0.5		19	play	1		19	center	0.5
20	require	0.125		20	genre	0.875		20	player	1
		0.725				0.66875	0.66875			0.675

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

Power	Term	Relatedness	Power	Term	Relatedness	Power	Term	Relatedness
	1 nuclear	1		1 market	0.75		1 market	0.75
	2 power	0.875		2 power	0.875		2 power	0.875
	3 detector	1		3 series	0.5		3 electric	0.75
	4 system	0.625		4 nominal	0.5		4 supply	1
	5 set	0.5		5 supply	1		5 rating	0.875
	6 love	0.625		6 nb	0		6 vacuum	0.625
	7 number-one	0.375		7 steering	0.75		7 transmission	0.625
	8 r&b	0		8 electric	0.75		8 series	0.5
	9 refers	0.375		9 process	0.5		9 tools	1
	10 signal	1		10 police	0.75		10 nominal	0.5
	11 averaging	0.375		11 instruction	0.5		11 source	1
	12 snr	0.25		12 set	0.5		12 trios	0.125
	13 electrical	1		13 stations	0.625		13 nb	0
	14 nep	0.25		14 projection	0.5		14 supplies	0.875
	15 energy	1		15 hong	0		15 international	0.75
	16 unit	1		16 kong	0.125		16 relations	0.75
	17 powerset	1		17 tower	0.625		17 maximum	0.875
	18 daiichi	0.125		18 stone	0		18 energy	1
	19 expressed	0.625		19 labour	0		19 stations	0.625
	20 units	1		20 energy	1		20 steering	0.75
		0.65			0.5125	0.5125		0.7125

Piano	Term	Relatedness	Piano	Term	Relatedness	Piano	Term	Relatedness
	1 piano	1		1 piano	1		1 major	0.75
	2 tuning	0.875		2 roll	0.5		2 op	0
	3 systems	0.25		3 novelty	0.625		3 flat	0.75
	4 attic	0.25		4 rolls	0.375		4 piano	1
	5 floor	0.5		5 electric	1		5 trio	0.75
	6 larger	0.25		6 pianos	0.625		6 minor	1
	7 windows	0.125		7 grand	1		7 cello	0.375
	8 pianos	0.625		8 chamber	0.625		8 violin	0.5
	9 tuned	1		9 music	1		9 roll	0.5
	10 organ	0.75		10 sextet	0.625		10 grand	1
	11 reasons	0		11 digital	0.75		11 novelty	0.625
	12 rooms	0.25		12 tuning	0.875		12 trios	0.75
	13 classical	1		13 played	0.75		13 robert	0.25
	14 styles	0.625		14 wind	0		14 schumann	0.75
	15 sonatas	0.75		15 electronic	0.875		15 chamber	0.625
	16 percussive	0.5		16 classical	1		16 music	1
	17 sound	1		17 pianists	0.875		17 rolls	0.375
	18 noble	0.5		18 playing	1		18 concerto	0.625
	19 strings	0.875		19 quintet	0.75		19 quintet	0.75
	20 spring	0.375		20 produced	0.75		20 electric	1
		0.575			0.75	0.75		0.66875

Moon	Term	Relatedness	Moon	Term	Relatedness	Moon	Term	Relatedness
	1 eclipses	0.875		1 moon	1		1 moon	1
	2 occur	0.625		2 occurs	0.75		2 occurs	0.75
	3 moon	1		3 eclipses	0.875		3 natural	0.75
	4 occurs	0.75		4 occur	0.625		4 satellite	0.75
	5 dark	1		5 billion	0.25		5 eclipses	0.875
	6 solar	0.875		6 years	0.875		6 occur	0.625
	7 system	0.5		7 blue	0.75		7 dean	0.375
	8 visible	0.75		8 aka	0		8 billion	0.25
	9 sailor	0.375		9 term	0.25		9 years	0.875
	10 senshi	0.25		10 held	0		10 dead	0.25
	11 eclipse	0.875		11 record	0.375		11 wrasses	0.125
	12 astronomy	1		12 bbc	0.25		12 solar	0.875
	13 fuku	0.25		13 soap	0.125		13 system	0.5
	14 hemisphere	0.875		14 dark	1		14 bbc	0.25
	15 full	1		15 short	0		15 soap	0.125
	16 centered	0.375		16 story	0.375		16 blue	0.75
	17 film	0.625		17 eastenders	0		17 short	0
	18 industry	0.25		18 played	0.125		18 story	0.375
	19 io	0.375		19 apollo	0.875		19 eastenders	0
	20 satellite	0.75		20 mission	0.875		20 played	0.125
		0.66875			0.46875	0.46875		0.48125

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

Screen	Term	Relatedness	Screen	Term	Relatedness	Screen	Term	Relatedness
	1 rood	0.25		1 bubble	0.125		1 guild	0.25
	2 screens	1		2 screen	1		2 theater	0.875
	3 enlarged	0.625		3 film	0.875		3 clear	0.375
	4 portion	0.375		4 television	1		4 view	0.875
	5 functional	0.5		5 projection	1		5 fire	0
	6 vision	0.75		6 scottish	0		6 screen	1
	7 screen	1		7 sharing	0.125		7 motion	0.75
	8 smoke	0		8 front	0.125		8 picture	0.875
	9 magnificatio	0.375		9 screens	1		9 operating	0.5
	10 techniques	0.125		10 type	0.625		10 systems	0.375
	11 choir	0		11 recall	0.125		11 type	0.625
	12 material	0.125		12 task	0.25		12 projection	1
	13 mesh	0.25		13 loading	0.75		13 bubble	0.125
	14 openings	0.25		14 motion	0.75		14 scottish	0
	15 interest	0.375		15 picture	0.875		15 film	0.875
	16 user	0		16 enlarged	0.625		16 television	1
	17 theory	0		17 portion	0.375		17 sharing	0.125
	18 attached	0		18 rood	0.25		18 computer	1
	19 stencil	0		19 moving	0.625		19 display	1
	20 print	0.5		20 middle	0.375		20 front	0.125
		0.325			0.54375	0.54375		0.5875

Car	Term	Relatedness	Car	Term	Relatedness	Car	Term	Relatedness
	1 body	0.5		1 body	0.5		1 car	1
	2 style	0.625		2 style	0.625		2 finance	0.25
	3 baggage-dor	0.75		3 car	1		3 body	0.5
	4 cars	1		4 phone	0.25		4 style	0.625
	5 cadillac	1		5 type	0.5		5 cars	1
	6 fleetwood	0		6 bait	0		6 include	0
	7 car	1		7 automated	0.875		7 drag	1
	8 talk	0		8 vehicles	1		8 racing	1
	9 amtrak	0.25		9 cars	1		9 phone	0.25
	10 received	0.125		10 include	0		10 surfing	0
	11 wheels	0.875		11 caution	0.5		11 general	0.375
	12 carry	0.5		12 period	0		12 motors	0.875
	13 town	0.125		13 motor	1		13 energy	0.625
	14 configured	0.5		14 drag	1		14 water	0.125
	15 general	0.375		15 racing	1		15 burning	0.875
	16 motors	0.875		16 combustion	0.75		16 bait	0
	17 introduced	0		17 engine	1		17 intended	0
	18 lancia	0.5		18 model	0.875		18 motor	1
	19 ledged	0.375		19 year	0.375		19 vehicles	1
	20 rudimentum	0.125		20 area	0.125		20 rental	0.875
		0.475			0.61875	0.61875		0.56875

Book	Term	Relatedness	Book	Term	Relatedness	Book	Term	Relatedness
	1 book	1		1 book	1		1 book	1
	2 music	0.25		2 music	0.25		2 soup	0
	3 print	0.875		3 tokens	0.125		3 music	0.25
	4 run	0		4 review	0.75		4 lungs	0
	5 sense	0.125		5 lungs	0		5 tokens	0.125
	6 series	0.875		6 print	0.875		6 print	0.875
	7 reprinted	0.875		7 run	0		7 run	0
	8 additional	0.125		8 curse	0.125		8 closure	0.125
	9 copies	0.625		9 high	0		9 curse	0.125
	10 publisher	1		10 school	0.75		10 building	0.125
	11 bookseller	1		11 block	0.25		11 printed	0.875
	12 art	0.625		12 books	1		12 high	0
	13 form	0.125		13 curses	0.125		13 school	0.75
	14 called	0		14 sense	0.125		14 born	0
	15 towns	0		15 burning	0.25		15 block	0.25
	16 books	1		16 collecting	0.5		16 books	1
	17 mechanical	0.125		17 printed	0.875		17 series	0.875
	18 organs	0		18 series	0.875		18 date	0.125
	19 antiquarian	0.375		19 cover	1		19 generally	0
	20 initial	0		20 considered	0		20 published	0.875
		0.45			0.44375	0.44375		0.36875

Appendix A: Neighbouring Terms: All queries for all Similarity models and sizes of N

Bike		Term	Relatedness	Bike		Term	Relatedness	Bike		Term	Relatedness
	1	star	0		1	star	0		1	bike	1
	2	wars	0		2	wars	0		2	park	0.5
	3	bike	1		3	bike	1		3	path	0.75
	4	ms	0		4	bus	0.375		4	week	0.125
	5	society	0.25		5	build	0		5	paths	0.75
	6	multiple	0.25		6	racing	0.875		6	bar	0.25
	7	sclerosis	0.125		7	affordable	0.875		7	bikes	1
	8	friday	0		8	housing	0		8	trail	1
	9	drive	0.625		9	park	0.5		9	racing	0.875
	10	train	0.375		10	ms	0		10	motorcycle	0.875
	11	events	0.125		11	club	0		11	rally	0.5
	12	organized	0		12	head	0.125		12	star	0
	13	logging	0		13	tube	0.5		13	wars	0
	14	roads	0.625		14	shop	0.625		14	train	0.375
	15	frame	0.25		15	path	0.75		15	road	0.75
	16	large	0.125		16	society	0.25		16	bus	0.375
	17	maintaining	0.75		17	cycling	1		17	ride	0.875
	18	dismounting	0.25		18	event	0.125		18	build	0
	19	legs	0.75		19	philly	0		19	bicycle	1
	20	track	0.875		20	creative	0		20	sharing	0.125
			0.31875				0.35	0.35			0.55625
Soldier		**Term**	**Relatedness**	**Soldier**		**Term**	**Relatedness**	**Soldier**		**Term**	**Relatedness**
	1	cavalry	1		1	lone	0.375		1	lone	0.375
	2	regiment	0.875		2	soldiers	1		2	soldiers	1
	3	infantry	1		3	soldier	1		3	game	0.125
	4	glen	0.125		4	action	0.5		4	released	0.125
	5	matlock	0.125		5	figures	0.375		5	song	0.125
	6	african-amer	0.125		6	good	0.25		6	written	0
	7	regiments	0.625		7	century	0.125		7	soldier	1
	8	cross-countr	0.125		8	toys	0.5		8	action	0.5
	9	skiing	0.125		9	vietnam	1		9	figures	0.375
	10	fictional	0		10	war	1		10	iron	0.625
	11	iggy	0		11	military	1		11	good	0.25
	12	pop	0		12	personnel	0.75		12	vietnam	1
	13	buffalo	0.125		13	cavalry	1		13	war	1
	14	soldiers	1		14	regiment	0.875		14	century	0.125
	15	dead	0.75		15	set	0		15	toys	0.5
	16	prez	0		16	world	0.375		16	set	0
	17	army	1		17	buffalo	0.125		17	world	0.375
	18	domestic	0		18	studio	0		18	buffalo	0.125
	19	album	0		19	game	0.125		19	studio	0
	20	lack	0		20	released	0.125		20	future	0.5
			0.35				0.525	0.525			0.40625
								0.51125			
			5				**7**				**8**

9 Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

VECTOR											
10 articles				**75 articles**				**150 articles**			
Cloud		**Term**	**Relatedness**	**Cloud**		**Term**	**Relatedness**	**Cloud**		**Term**	**Relatedness**
	1	cloud	1		1	clouds	1		1	clouds	1
	2	occur	0.375		2	cloud	1		2	cloud	1
	3	depend	0.25		3	large	0.375		3	large	0.375
	4	limit	0		4	altitude	0.75		4	america	0
	5	water	1		5	water	1		5	altitude	0.75
	6	article	0		6	cumulus	0.75		6	water	1
	7	observation	0.75		7	develop	0.125		7	american	0
	8	clouds	1		8	precipitation	0.375		8	release	0
	9	temperature	0.875		9	generally	0.125		9	light	0.875
	10	observations	0.75		10	level	0.375		10	common	0
	11	served	0.25		11	vapor	0.875		11	cumulus	0.75
	12	local	0.375		12	north	0		12	system	0.25
	13	density	0.875		13	character	0		13	service	0.5
	14	connection	0.125		14	produce	0		14	released	0
	15	ratio	0		15	common	0		15	generally	0.125
	16	outer	0		16	surface	0.5		16	based	0.125
	17	significant	0		17	occur	0.375		17	formation	1
	18	cumulus	0.75		18	small	0.375		18	level	0.375
	19	direct	0		19	ground	0		19	produce	0
	20	aerosol	0		20	light	0.875		20	computing	0.625
			0.41875				0.44375				0.4375
Cat		**Term**	**Relatedness**	**Cat**		**Term**	**Relatedness**	**Cat**		**Term**	**Relatedness**
	1	world	0.125		1	species	1		1	common	0.125
	2	effect	0		2	small	0.375		2	america	0
	3	species	1		3	america	0		3	small	0.375
	4	number	0.125		4	common	0.125		4	released	0.125
	5	early	0		5	origin	0.375		5	short	0.375
	6	popular	0.5		6	called	0		6	american	0
	7	super	0.125		7	world	0.125		7	origin	0.375
	8	language	0		8	population	0.375		8	called	0
	9	population	0.375		9	breed	0.875		9	species	1
	10	genetic	0.625		10	state	0.125		10	produce	0
	11	felids	0.5		11	original	0.25		11	breed	0.875
	12	considered	0.125		12	felis	0.5		12	domestic	1
	13	small	0.375		13	range	0.125		13	character	0.625
	14	called	0		14	series	0		14	include	0
	15	distinguish	0.25		15	number	0.125		15	record	0
	16	state	0.125		16	character	0.625		16	number	0.125
	17	felis	0.5		17	american	0		17	world	0.125
	18	million	0		18	south	0		18	series	0
	19	feral	0.625		19	domestic	1		19	animal	1
	20	domestic	1		20	record	0		20	state	0.125
			0.31875				0.3				0.3125
Weapon		**Term**	**Relatedness**	**Weapon**		**Term**	**Relatedness**	**Weapon**		**Term**	**Relatedness**
	1	weapon	1		1	weapon	1		1	weapon	1
	2	weapons	1		2	weapons	1		2	weapons	1
	3	device	0.875		3	design	0.375		3	design	0.375
	4	modern	0.125		4	military	1		4	develop	0.75
	5	force	1		5	device	0.875		5	system	0
	6	projectile	1		6	include	0		6	state	0.625
	7	include	0		7	designed	0.375		7	large	0.25
	8	purpose	0.375		8	purpose	0.375		8	military	1
	9	implements	0		9	power	1		9	united	0.125
	10	damage	1		10	similar	0.125		10	states	0.625
	11	warfare	1		11	refer	0.125		11	power	1
	12	power	1		12	common	0.125		12	release	0.25
	13	great	0		13	explosive	0.875		13	include	0
	14	missile	1		14	combat	0.875		14	force	1
	15	effect	0.5		15	america	0.875		15	device	0.875
	16	advantage	0.25		16	state	0.625		16	designed	0.375
	17	produce	0.25		17	force	1		17	refer	0.125
	18	combat	0.875		18	intended	0.375		18	based	0
	19	increase	0.25		19	modern	0.125		19	released	0.125
	20	explosive	0.875		20	effect	0.5		20	purpose	0.375
			0.61875				0.58125				0.49375

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

School		Term	Relatedness
	1	school	1
	2	chool	0
	3	schools	1
	4	education	1
	5	private	1
	6	schooling	1
	7	generally	0
	8	public	1
	9	institution	0.875
	10	common	0.625
	11	child	1
	12	attend	1
	13	student	1
	14	government	0.375
	15	building	0.875
	16	economic	0.5
	17	children	1
	18	include	0.625
	19	alternative	0
	20	students	1
			0.74375

School		Term	Relatedness
	1	education	1
	2	school	1
	3	schools	1
	4	educational	1
	5	students	1
	6	student	1
	7	founded	0.5
	8	located	0.25
	9	state	0.875
	10	college	1
	11	independent	0.5
	12	private	1
	13	member	0.25
	14	secondary	0.875
	15	board	0.75
	16	institution	0.875
	17	united	0.375
	18	campus	1
	19	england	0.25
	20	states	0.875
			0.76875

School		Term	Relatedness
	1	education	1
	2	school	1
	3	schools	1
	4	educational	1
	5	student	1
	6	students	1
	7	independent	0.5
	8	located	0.25
	9	state	0.875
	10	founded	0.5
	11	college	1
	12	board	0.75
	13	grade	1
	14	england	0.25
	15	preparatory	0.625
	16	boarding	0.75
	17	private	1
	18	states	0.875
	19	member	0.25
	20	united	0.375
			0.75

Police		Term	Relatedness
	1	force	1
	2	police	1
	3	officers	1
	4	officer	1
	5	territorial	0.5
	6	units	1
	7	policing	0.875
	8	authorities	1
	9	major	0.125
	10	enforcement	0.875
	11	forces	1
	12	traffic	0.875
	13	large	0.125
	14	responsible	0.875
	15	crime	1
	16	state	1
	17	members	0.25
	18	services	0.875
	19	enforce	0.875
	20	service	1
			0.8125

Police		Term	Relatedness
	1	force	1
	2	police	1
	3	office	0.875
	4	officer	1
	5	state	1
	6	enforce	0.875
	7	enforcement	0.875
	8	officers	1
	9	crime	1
	10	policing	0.875
	11	agency	0.625
	12	forces	1
	13	service	1
	14	person	0.75
	15	power	0.75
	16	public	0.875
	17	states	1
	18	member	0.25
	19	united	0.75
	20	territorial	0.5
			0.85

Police		Term	Relatedness
	1	force	1
	2	police	1
	3	office	0.875
	4	officer	1
	5	officers	1
	6	state	1
	7	crime	1
	8	policing	0.875
	9	enforce	0.875
	10	enforcement	0.875
	11	service	1
	12	forces	1
	13	territorial	0.5
	14	agency	0.625
	15	responsible	0.875
	16	public	0.875
	17	united	0.75
	18	states	1
	19	include	0.5
	20	local	0.75
			0.86875

Fruit		Term	Relatedness
	1	seeds	1
	2	fruit	1
	3	include	0
	4	apple	1
	5	fruits	1
	6	orange	1
	7	flavor	0.875
	8	flesh	0.5
	9	produce	0.375
	10	flower	0.875
	11	called	0
	12	sweet	0.875
	13	apples	1
	14	state	0
	15	varies	0.125
	16	language	0
	17	action	0
	18	juicy	0
	19	green	0
	20	plants	0
			0.48125

Fruit		Term	Relatedness
	1	fruits	1
	2	fruit	1
	3	apple	1
	4	sweet	0.875
	5	common	0
	6	include	0
	7	range	0.125
	8	produce	0.375
	9	called	0
	10	orange	1
	11	similar	0.25
	12	sugar	0.625
	13	large	0.125
	14	released	0
	15	product	0.375
	16	album	0
	17	flavor	0.875
	18	based	0
	19	flavour	0.875
	20	version	0
			0.425

Fruit		Term	Relatedness
	1	fruits	1
	2	fruit	1
	3	range	0.125
	4	species	0.125
	5	large	0.125
	6	common	0
	7	small	0.125
	8	produce	0.375
	9	apple	1
	10	release	0.125
	11	sweet	0.875
	12	product	0.375
	13	album	0
	14	orange	1
	15	south	0.125
	16	include	0
	17	america	0
	18	similar	0.25
	19	united	0
	20	version	0
			0.33125

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

Meat	Term	Relatedness
1	style	0
2	album	0
3	albums	0
4	music	0
5	sound	0
6	success	0
7	america	0.375
8	served	0.875
9	world	0
10	refer	0.125
11	process	0.5
12	protein	0.875
13	developed	0.25
14	released	0
15	notable	0.125
16	original	0.25
17	performance	0
18	release	0
19	broke	0
20	vocal	0
		0.16875

Meat	Term	Relatedness
1	animal	1
2	include	0
3	refer	0.125
4	process	0.5
5	animals	1
6	consumption	1
7	world	0
8	product	0.625
9	method	0.125
10	common	0
11	produce	0.5
12	meats	1
13	called	0
14	cooked	1
15	human	0.375
16	including	0.125
17	united	0
18	america	0.375
19	parts	0.375
20	countries	0
		0.40625

Meat	Term	Relatedness
1	animal	1
2	product	0.625
3	process	0.5
4	include	0
5	state	0
6	produce	0.5
7	release	0
8	united	0
9	released	0
10	states	0
11	album	0
12	including	0.125
13	large	0.125
14	america	0.375
15	world	0
16	ground	0.375
17	animals	1
18	called	0
19	products	0.625
20	consumption	1
		0.3125

Telephone	Term	Relatedness
1	telephone	1
2	phone	1
3	connect	0.75
4	phones	1
5	service	0.875
6	telephones	1
7	electric	0.75
8	network	1
9	system	0.5
10	signals	0.875
11	device	0.75
12	current	0.125
13	signal	0.875
14	switch	0.25
15	small	0.5
16	called	0.75
17	access	0.5
18	electrical	0.625
19	telecommun	0
20	connected	0.875
		0.7

Telephone	Term	Relatedness
1	telephone	1
2	phone	1
3	service	0.875
4	phones	1
5	connect	0.75
6	number	0.875
7	telephones	1
8	communicati	1
9	system	0.5
10	company	0.5
11	network	1
12	communicati	1
13	telecommun	1
14	services	0.75
15	common	0.125
16	exchange	0.25
17	switch	0.25
18	electric	0.75
19	place	0.375
20	state	0
		0.7

Telephone	Term	Relatedness
1	telephone	1
2	phone	1
3	number	0.875
4	service	0.875
5	system	0.5
6	communicati	1
7	phones	1
8	network	1
9	communicati	1
10	company	0.5
11	numbers	0.875
12	digit	0.875
13	telecom	1
14	services	0.75
15	telephones	1
16	connect	0.75
17	telecommun	1
18	national	0.125
19	local	0.375
20	exchange	0.25
		0.7875

Machine	Term	Relatedness
1	machine	1
2	abstract	0.125
3	system	0.75
4	computer	0.875
5	number	0.375
6	machines	1
7	systems	0.625
8	general	0.25
9	language	0.125
10	design	0.5
11	large	0.375
12	level	0.375
13	require	0
14	produce	0.75
15	government	0.125
16	personal	0
17	ratio	0.25
18	hardware	0.875
19	effective	0
20	simple	0.125
		0.425

Machine	Term	Relatedness
1	machine	1
2	abstract	0.125
3	machines	1
4	computer	0.875
5	system	0.75
6	process	0.625
7	device	0.625
8	produce	0.75
9	perform	0.5
10	design	0.5
11	called	0.125
12	application	0.375
13	state	0.375
14	number	0.375
15	designed	0.25
16	model	0.625
17	mechanical	0.875
18	large	0.375
19	general	0.25
20	original	0.25
		0.53125

Machine	Term	Relatedness
1	machine	1
2	abstract	0.125
3	machines	1
4	computer	0.875
5	device	0.625
6	produce	0.75
7	system	0.75
8	design	0.5
9	process	0.625
10	state	0.375
11	number	0.375
12	called	0.125
13	designed	0.25
14	common	0
15	mechanical	0.875
16	original	0.25
17	perform	0.5
18	create	0
19	class	0.25
20	model	0.625
		0.49375

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

Computer	Term	Relatedness
1	computer	1
2	abstract	0.125
3	computers	1
4	program	0.875
5	application	0.875
6	system	0.75
7	world	0.125
8	solve	0.375
9	applications	0.875
10	electronic	0.75
11	machine	0.875
12	process	0.625
13	general	0
14	perform	0.5
15	systems	0.875
16	state	0.375
17	problem	0.625
18	design	0.25
19	digital	0.75
20	order	0.25
		0.59375

Computer	Term	Relatedness
1	computer	1
2	abstract	0.125
3	computers	1
4	system	0.75
5	systems	0.875
6	program	0.875
7	design	0.25
8	include	0
9	process	0.625
10	develop	0.5
11	electronic	0.75
12	computing	1
13	based	0.125
14	device	0.75
15	machine	0.875
16	refer	0.125
17	software	1
18	digital	0.75
19	information	0.625
20	require	0.125
		0.60625

Computer	Term	Relatedness
1	computer	1
2	computers	1
3	system	0.75
4	systems	0.875
5	design	0.25
6	include	0
7	machine	0.875
8	process	0.625
9	program	0.875
10	format	0.75
11	based	0.125
12	software	1
13	computing	1
14	hardware	1
15	electronic	0.75
16	digital	0.75
17	develop	0.5
18	device	0.75
19	perform	0.5
20	early	0
		0.66875

Space	Term	Relatedness
1	space	0.875
2	spaces	0.75
3	mathematics	0.875
4	point	0.625
5	topological	0.5
6	finite	0.5
7	define	0.375
8	definition	0.25
9	euclid	0.375
10	euclidean	0.375
11	defined	0.375
12	system	0.875
13	dimension	0.625
14	concept	0.5
15	number	0.375
16	element	0.625
17	infinite	0
18	general	0
19	topology	0.25
20	named	0
		0.45625

Space	Term	Relatedness
1	space	0.875
2	abstract	0.5
3	spaces	0.75
4	mathematics	0.875
5	logic	0.125
6	topological	0.5
7	called	0
8	point	0.625
9	system	0.875
10	topology	0.25
11	general	0
12	element	0.625
13	include	0.25
14	earth	0.875
15	concept	0.5
16	structure	0.125
17	notion	0
18	close	0.25
19	sense	0.125
20	related	0.25
		0.41875

Space	Term	Relatedness
1	space	0.875
2	abstract	0.5
3	spaces	0.75
4	mathematics	0.875
5	called	0
6	point	0.625
7	topological	0.5
8	general	0
9	topology	0.25
10	system	0.875
11	metric	0.5
12	theory	0.625
13	function	0.25
14	element	0.625
15	algebra	0.375
16	related	0.25
17	earth	0.875
18	vector	0.5
19	points	0.375
20	structure	0.125
		0.4875

Game	Term	Relatedness
1	player	1
2	games	0.875
3	players	0.875
4	involve	0.5
5	playing	1
6	person	0.5
7	number	0.25
8	human	0.5
9	board	1
10	computer	1
11	played	0.875
12	early	0.125
13	require	0.125
14	exist	0.375
15	present	0.375
16	purpose	0.5
17	range	0.25
18	rules	0.875
19	distinction	0
20	result	0
		0.55

Game	Term	Relatedness
1	player	1
2	games	0.875
3	players	0.875
4	video	1
5	board	1
6	played	0.875
7	number	0.25
8	called	0.125
9	playing	1
10	general	0.125
11	require	0.125
12	common	0.25
13	action	0.625
14	computer	1
15	include	0.375
16	based	0.25
17	involve	0.5
18	popular	0.375
19	chess	1
20	theory	0.625
		0.6125

Game	Term	Relatedness
1	player	1
2	games	0.875
3	video	1
4	players	0.875
5	played	0.875
6	number	0.25
7	include	0.375
8	board	1
9	general	0.125
10	playing	1
11	popular	0.375
12	called	0.125
13	world	0.5
14	computer	1
15	involve	0.5
16	action	0.625
17	theory	0.625
18	common	0.25
19	typically	0.125
20	require	0.125
		0.58125

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

Power	Term	Relatedness	Power	Term	Relatedness	Power	Term	Relatedness
1	abstract	0.125	1	electric	0.75	1	electric	0.75
2	power	0.875	2	power	0.875	2	power	0.875
3	ratio	0.5	3	system	0.625	3	system	0.625
4	define	0.125	4	electrical	1	4	refer	0.125
5	electric	0.75	5	ratio	0.5	5	large	0.25
6	generate	0.625	6	control	1	6	energy	1
7	system	0.625	7	energy	1	7	state	0.25
8	number	0.375	8	refer	0.125	8	electrical	1
9	electrical	1	9	large	0.25	9	control	1
10	measure	0.5	10	state	0.25	10	source	1
11	defined	0.375	11	ability	0.25	11	general	0.125
12	output	0.5	12	transmission	0.625	12	ability	0.25
13	energy	1	13	require	0.375	13	world	0.25
14	signal	1	14	general	0.125	14	national	0.5
15	frequency	0	15	concept	0.125	15	include	0.25
16	bandwidth	0	16	point	0.25	16	great	0.25
17	generated	0.625	17	force	0.75	17	transmission	0.625
18	noise	0	18	common	0	18	called	0.125
19	distribution	0.5	19	service	0.375	19	america	0.125
20	developer	0.25	20	world	0.25	20	systems	0.625
		0.4875			0.475			0.5

Piano	Term	Relatedness	Piano	Term	Relatedness	Piano	Term	Relatedness
1	strings	0.875	1	music	1	1	music	1
2	piano	1	2	piano	1	2	piano	1
3	string	0.875	3	instrument	0.75	3	instrument	0.75
4	instrument	0.75	4	pianos	0.625	4	pianos	0.625
5	system	0	5	instruments	0.875	5	pianist	1
6	sound	1	6	perform	0.875	6	sound	1
7	board	0.125	7	played	0.75	7	piece	0.5
8	refer	0	8	sound	1	8	instruments	0.875
9	music	1	9	keyboard	0.625	9	written	0.125
10	popular	0.125	10	pianist	1	10	played	0.75
11	italian	0.75	11	popular	0.125	11	string	0.875
12	keyboard	0.625	12	string	0.875	12	released	0.25
13	pianos	0.625	13	feature	0.375	13	keyboard	0.625
14	architect	0	14	record	0.375	14	musical	0.625
15	version	0	15	performance	1	15	album	0.375
16	world	0	16	musician	0.625	16	include	0
17	requires	0	17	musical	0.625	17	performance	1
18	hammer	0.875	18	produced	0.75	18	popular	0.125
19	person	0	19	strings	0.875	19	musician	0.625
20	frequencies	0	20	piece	0.5	20	concert	1
		0.43125			0.73125			0.65625

Moon	Term	Relatedness	Moon	Term	Relatedness	Moon	Term	Relatedness
1	solar	0.875	1	american	0.125	1	american	0.125
2	occur	0.625	2	america	0.125	2	america	0.125
3	occurs	0.75	3	release	0	3	release	0
4	calendar	0.375	4	released	0	4	released	0
5	phases	0.25	5	state	0.125	5	written	0.125
6	precise	0	6	album	0.125	6	fiction	0.125
7	calendars	0.25	7	record	0.375	7	state	0.125
8	plane	0	8	music	0.375	8	story	0.375
9	lunar	1	9	written	0.125	9	album	0.125
10	precisely	0.125	10	states	0.125	10	record	0.375
11	eclipse	0.875	11	character	0	11	music	0.375
12	phase	0.625	12	played	0.125	12	states	0.125
13	earth	0.375	13	story	0.375	13	united	0
14	influence	0.375	14	series	0	14	space	0.75
15	including	0	15	united	0	15	played	0.125
16	released	0	16	lunar	1	16	character	0
17	japan	0	17	occur	0.625	17	original	0.125
18	eclipses	0.875	18	september	0.125	18	published	0
19	resulting	0.125	19	space	0.75	19	series	0
20	planet	0.625	20	based	0	20	based	0
		0.40625			0.225			0.15

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

Screen	Term	Relatedness
1	screen	1
2	screens	1
3	system	0.375
4	content	0.75
5	present	0.625
6	original	0.125
7	technology	0.875
8	common	0.125
9	origin	0
10	visually	1
11	visual	1
12	reading	0
13	assistive	0.25
14	windows	0
15	english	0
16	operating	0.5
17	cover	0.625
18	target	0
19	popular	0.375
20	print	0.5
		0.45625

Screen	Term	Relatedness
1	screen	1
2	screens	1
3	design	0.75
4	include	0.25
5	product	0.625
6	system	0.375
7	material	0.125
8	direct	0.125
9	cover	0.625
10	feature	0.5
11	function	0.25
12	common	0.125
13	create	0.375
14	display	1
15	vision	0.75
16	motion	0.75
17	light	1
18	image	1
19	surface	0.875
20	video	1
		0.625

Screen	Term	Relatedness
1	screen	1
2	screens	1
3	vision	0.75
4	include	0.25
5	television	1
6	product	0.625
7	design	0.75
8	produce	0.5
9	display	1
10	cinema	1
11	america	0
12	system	0.375
13	video	1
14	table	0.375
15	based	0.125
16	production	0.5
17	series	0.5
18	national	0
19	american	0
20	material	0.125
		0.54375

Car	Term	Relatedness
1	automobile	1
2	america	0.625
3	north	0
4	large	0.375
5	american	0.125
6	referred	0.125
7	larger	0.375
8	compact	0.5
9	short	0.25
10	light	0.25
11	united	0
12	vehicle	1
13	states	0.125
14	railroad	0.375
15	motor	1
16	range	0.875
17	turned	0
18	chevrolet	1
19	market	0.625
20	extreme	0.125
		0.4375

Car	Term	Relatedness
1	mobile	0.75
2	vehicle	1
3	automobile	1
4	vehicles	1
5	america	0.625
6	design	0.625
7	include	0
8	sport	0.5
9	drive	1
10	power	0.875
11	general	0.375
12	small	0.375
13	united	0
14	large	0.375
15	designed	0.625
16	built	0.5
17	railroad	0.375
18	engine	1
19	class	0.5
20	generally	0.125
		0.58125

Car	Term	Relatedness
1	vehicles	1
2	vehicle	1
3	automobile	1
4	mobile	0.75
5	america	0.625
6	large	0.375
7	design	0.625
8	sport	0.5
9	include	0
10	drive	1
11	built	0.5
12	small	0.375
13	engine	1
14	general	0.375
15	train	0.5
16	typical	0.125
17	passenger	1
18	designed	0.625
19	common	0.5
20	motor	1
		0.64375

Book	Term	Relatedness
1	large	0.125
2	books	1
3	print	0.875
4	paper	0.875
5	printed	0.875
6	pages	1
7	market	0.5
8	store	0.875
9	cover	1
10	written	1
11	magazine	0.75
12	bookseller	1
13	author	1
14	longer	0.375
15	publisher	1
16	format	0.875
17	series	0.875
18	organ	0
19	order	0
20	electronic	0.625
		0.73125

Book	Term	Relatedness
1	books	1
2	format	0.875
3	print	0.875
4	general	0.125
5	include	0.125
6	large	0.125
7	refer	0.5
8	common	0.125
9	pages	1
10	paper	0.875
11	graph	0.125
12	called	0
13	collect	0.375
14	printed	0.875
15	generally	0
16	cover	1
17	state	0.125
18	record	0.5
19	author	1
20	small	0.125
		0.4875

Book	Term	Relatedness
1	books	1
2	print	0.875
3	format	0.875
4	state	0.125
5	general	0.125
6	large	0.125
7	printed	0.875
8	include	0.125
9	paper	0.875
10	cover	1
11	pages	1
12	united	0
13	common	0.125
14	states	0
15	published	0.875
16	record	0.5
17	called	0
18	refer	0.5
19	public	0.5
20	generally	0
		0.475

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

Bike	Term	Relatedness	Bike	Term	Relatedness	Bike	Term	Relatedness
1	bicycle	1	1	bicycle	1	1	bicycle	1
2	cycle	1	2	cycle	1	2	cycle	1
3	bikes	1	3	bikes	1	3	cycling	1
4	typical	0.125	4	cycling	1	4	bikes	1
5	typically	0.125	5	motorcycle	0.875	5	event	0.125
6	roads	0.625	6	design	0.625	6	mountain	1
7	large	0.125	7	sport	0.5	7	state	0.125
8	motor	0.625	8	state	0.125	8	sport	0.5
9	place	0.125	9	wheel	1	9	design	0.625
10	cycling	1	10	event	0.125	10	rider	1
11	purpose	0.125	11	street	0.625	11	include	0.125
12	cyclist	1	12	rider	1	12	cyclists	1
13	rider	1	13	include	0.125	13	trail	1
14	speed	0.75	14	public	0.375	14	motorcycle	0.875
15	sport	0.5	15	drive	0.625	15	world	0.125
16	country	0.125	16	vehicle	0.625	16	cross	0.375
17	frame	0.25	17	cross	0.375	17	public	0.375
18	cyclists	1	18	cyclists	1	18	south	0
19	power	0.375	19	bicycles	1	19	country	0.125
20	train	0.375	20	world	0.125	20	riders	1
		0.5625			0.65625			0.33125

Soldier	Term	Relatedness	Soldier	Term	Relatedness	Soldier	Term	Relatedness
1	soldier	1	1	soldier	1	1	soldier	1
2	soldiers	1	2	release	0.125	2	release	0.125
3	album	0	3	world	0.375	3	released	0.125
4	produce	0	4	soldiers	1	4	america	0.625
5	state	0.625	5	released	0.125	5	soldiers	1
6	originally	0	6	america	0.625	6	world	0.375
7	group	0.75	7	state	0.625	7	state	0.625
8	number	0.5	8	united	0.625	8	written	0
9	force	0.875	9	original	0	9	american	0.5
10	states	0.5	10	based	0.25	10	series	0.25
11	united	0.625	11	states	0.5	11	album	0
12	fight	1	12	american	0.5	12	based	0.25
13	regiment	0.875	13	national	0.75	13	original	0
14	military	1	14	group	0.75	14	united	0.625
15	record	0	15	including	0	15	title	0.5
16	produced	0.125	16	series	0.25	16	directed	0.5
17	generally	0	17	directed	0.5	17	record	0
18	created	0	18	album	0	18	states	0.5
19	theme	0.25	19	music	0	19	including	0
20	including	0	20	include	0	20	group	0.75
		0.45625			0.4			0.3875
		6			**8**			**6**

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

BM25

10 articles			75 articles			150 articles		
Cloud	**Term**	**Relatedness**	**Cloud**	**Term**	**Relatedness**	**Cloud**	**Term**	**Relatedness**
	1 cloud	1		1 clouds	1		1 clouds	1
	2 occur	0.375		2 cloud	1		2 cloud	1
	3 observation	0.75		3 water	1		3 large	0.375
	4 observations	0.75		4 altitude	0.75		4 america	0
	5 clouds	1		5 large	0.375		5 altitude	0.75
	6 limit	0		6 cumulus	0.75		6 water	1
	7 temperature	0.875		7 generally	0.125		7 release	0
	8 water	1		8 precipitation	0.375		8 american	0
	9 article	0		9 produce	0		9 light	0.875
	10 depend	0.25		10 common	0		10 common	0
	11 served	0.25		11 vapor	0.875		11 cumulus	0.75
	12 confirmed	0		12 character	0		12 service	0.5
	13 medium	0.125		13 ground	0		13 based	0.125
	14 ionized	0		14 surface	0.5		14 released	0
	15 physics	1		15 occur	0.375		15 generally	0.125
	16 altitude	0.75		16 develop	0.125		16 system	0.25
	17 droplets	0		17 level	0.375		17 formation	1
	18 galaxies	0		18 small	0.375		18 produce	0
	19 limited	0		19 formation	1		19 level	0.375
	20 significant	0		20 shape	0.875		20 including	0
		0.40625			0.49375			0.40625

Cat	**Term**	**Relatedness**	**Cat**	**Term**	**Relatedness**	**Cat**	**Term**	**Relatedness**
	1 species	1		1 species	1		1 released	0.125
	2 result	0.25		2 america	0		2 america	0
	3 america	0		3 origin	0.375		3 american	0
	4 common	0.125		4 small	0.375		4 small	0.375
	5 vision	0.375		5 common	0.125		5 common	0.125
	6 early	0		6 character	0.625		6 origin	0.375
	7 popular	0.5		7 called	0		7 short	0.375
	8 super	0.125		8 world	0.125		8 produce	0
	9 domestic	1		9 original	0.25		9 breed	0.875
	10 specific	0.25		10 population	0.375		10 domestic	1
	11 sense	0.75		11 breed	0.875		11 include	0
	12 similar	0.25		12 series	0		12 called	0
	13 language	0		13 state	0.125		13 species	1
	14 artist	0.125		14 american	0		14 character	0.625
	15 number	0.125		15 felis	0.5		15 number	0.125
	16 american	0		16 south	0		16 record	0
	17 resulting	0.375		17 number	0.125		17 world	0.125
	18 called	0		18 domestic	1		18 state	0.125
	19 including	0		19 record	0		19 animal	1
	20 quantum	0		20 united	0		20 series	0
		0.2625			0.29375			0.3125

Weapon	**Term**	**Relatedness**	**Weapon**	**Term**	**Relatedness**	**Weapon**	**Term**	**Relatedness**
	1 weapon	1		1 weapon	1		1 weapon	1
	2 weapons	1		2 weapons	1		2 weapons	1
	3 device	0.875		3 design	0.375		3 design	0.375
	4 force	1		4 military	1		4 system	0
	5 projectile	1		5 device	0.875		5 develop	0.75
	6 modern	0.125		6 include	0		6 state	0.625
	7 include	0		7 designed	0.375		7 large	0.25
	8 advantage	0.25		8 power	1		8 military	1
	9 combat	0.875		9 purpose	0.375		9 states	0.625
	10 produce	0.25		10 similar	0.125		10 united	0.125
	11 implements	0		11 refer	0.125		11 power	1
	12 warfare	1		12 common	0.125		12 release	0.25
	13 damage	1		13 modern	0.125		13 include	0
	14 purpose	0.375		14 america	0.875		14 force	1
	15 power	1		15 combat	0.875		15 designed	0.375
	16 great	0		16 force	1		16 device	0.875
	17 increase	0.25		17 state	0.625		17 based	0
	18 effect	0.5		18 explosive	0.875		18 refer	0.125
	19 missile	1		19 intended	0.375		19 released	0.125
	20 projectiles	0		20 general	0		20 purpose	0.375
		0.575			0.55625			0.49375

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

School	Term	Relatedness
	1 schools	1
	2 school	1
	3 based	0.25
	4 education	1
	5 economic	0.5
	6 public	1
	7 place	0.375
	8 government	0.375
	9 national	0.5
	10 setting	0.125
	11 century	0.25
	12 world	0.375
	13 students	1
	14 building	0.875
	15 taught	1
	16 credit	0.875
	17 direction	0
	18 mental	0
	19 individuals	0.625
	20 state	0.875
		0.6

School	Term	Relatedness
	1 education	1
	2 school	1
	3 schools	1
	4 educational	1
	5 students	1
	6 student	1
	7 founded	0.5
	8 college	1
	9 state	0.875
	10 located	0.25
	11 secondary	0.875
	12 private	1
	13 independent	0.5
	14 institution	0.875
	15 campus	1
	16 grade	1
	17 board	0.75
	18 member	0.25
	19 include	0.625
	20 england	0.25
		0.7875

School	Term	Relatedness
	1 education	1
	2 school	1
	3 schools	1
	4 educational	1
	5 student	1
	6 students	1
	7 independent	0.5
	8 state	0.875
	9 located	0.25
	10 college	1
	11 founded	0.5
	12 board	0.75
	13 preparatory	0.625
	14 boarding	0.75
	15 england	0.25
	16 private	1
	17 grade	1
	18 include	0.625
	19 secondary	0.875
	20 states	0.875
		0.79375

Police	Term	Relatedness
	1 force	1
	2 police	1
	3 state	1
	4 officers	1
	5 office	0.875
	6 officer	1
	7 enforce	0.875
	8 enforcement	0.875
	9 australia	0.25
	10 traffic	0.875
	11 service	1
	12 units	1
	13 members	0.25
	14 station	1
	15 agency	0.625
	16 separate	0.25
	17 search	0.875
	18 special	0.375
	19 motor	0.25
	20 forces	1
		0.76875

Police	Term	Relatedness
	1 force	1
	2 police	1
	3 office	0.875
	4 officer	1
	5 state	1
	6 officers	1
	7 enforce	0.875
	8 crime	1
	9 service	1
	10 enforcement	0.875
	11 policing	0.875
	12 person	0.75
	13 forces	1
	14 united	0.75
	15 states	1
	16 public	0.875
	17 agency	0.625
	18 power	0.75
	19 common	0.375
	20 territorial	0.5
		0.85625

Police	Term	Relatedness
	1 force	1
	2 police	1
	3 office	0.875
	4 officer	1
	5 officers	1
	6 state	1
	7 crime	1
	8 policing	0.875
	9 enforcement	0.875
	10 service	1
	11 forces	1
	12 public	0.875
	13 territorial	0.5
	14 agency	0.625
	15 responsible	0.875
	16 united	0.75
	17 states	1
	18 special	0.375
	19 person	0.75
	20 include	0.5
		0.84375

Fruit	Term	Relatedness
	1 seeds	1
	2 fruit	1
	3 apple	1
	4 apples	1
	5 sweet	0.875
	6 include	0
	7 state	0
	8 fruits	1
	9 varies	0.125
	10 today	0.125
	11 strawberries	1
	12 orange	1
	13 flesh	0.5
	14 pears	1
	15 called	0
	16 produce	0.375
	17 specific	0.125
	18 syrup	0
	19 flowering	0
	20 animals	0.125
		0.5125

Fruit	Term	Relatedness
	1 fruits	1
	2 fruit	1
	3 apple	1
	4 common	0
	5 sweet	0.875
	6 produce	0.375
	7 range	0.125
	8 include	0
	9 called	0
	10 similar	0.25
	11 sugar	0.625
	12 orange	1
	13 released	0
	14 large	0.125
	15 product	0.375
	16 album	0
	17 flavor	0.875
	18 based	0
	19 america	0
	20 version	0
		0.38125

Fruit	Term	Relatedness
	1 species	0.125
	2 fruit	1
	3 fruits	1
	4 range	0.125
	5 large	0.125
	6 produce	0.375
	7 apple	1
	8 include	0
	9 united	0
	10 sweet	0.875
	11 small	0.125
	12 orange	1
	13 called	0
	14 product	0.375
	15 america	0
	16 natural	0.75
	17 eaten	0.75
	18 common	0
	19 plant	1
	20 south	0.125
		0.4375

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

Meat	Term	Relatedness
1	refer	0.125
2	process	0.5
3	america	0.375
4	american	0.375
5	species	0.625
6	served	0.875
7	world	0
8	include	0
9	album	0
10	protein	0.875
11	hunted	0.75
12	source	0.25
13	success	0
14	concerns	0.375
15	performance	0
16	reference	0
17	south	0.125
18	consume	0
19	animal	1
20	water	0.375
		0.33125

Meat	Term	Relatedness
1	animal	1
2	include	0
3	refer	0.125
4	process	0.5
5	animals	1
6	consumption	1
7	world	0
8	product	0.625
9	method	0.125
10	common	0
11	produce	0.5
12	meats	1
13	called	0
14	united	0
15	cooked	1
16	human	0.375
17	including	0.125
18	states	0
19	america	0.375
20	large	0.125
		0.39375

Meat	Term	Relatedness
1	animal	1
2	product	0.625
3	process	0.5
4	include	0
5	state	0
6	produce	0.5
7	release	0
8	united	0
9	released	0
10	states	0
11	world	0
12	album	0
13	including	0.125
14	large	0.125
15	america	0.375
16	ground	0.375
17	animals	1
18	called	0
19	products	0.625
20	consumption	1
		0.3125

Telephone	Term	Relatedness
1	telephone	1
2	phone	1
3	service	0.875
4	phones	1
5	system	0.5
6	connect	0.75
7	electric	0.75
8	telephones	1
9	internet	0.75
10	electrical	0.625
11	called	0.75
12	network	1
13	access	0.5
14	device	0.75
15	small	0.5
16	switch	0.25
17	number	0.875
18	invented	0
19	transmission	0.5
20	modern	0
		0.66875

Telephone	Term	Relatedness
1	telephone	1
2	phone	1
3	service	0.875
4	phones	1
5	connect	0.75
6	number	0.875
7	telephones	1
8	communicati	1
9	system	0.5
10	company	0.5
11	network	1
12	communicati	1
13	telecommun	1
14	services	0.75
15	common	0.125
16	exchange	0.25
17	switch	0.25
18	electric	0.75
19	place	0.375
20	state	0
		0.7

Telephone	Term	Relatedness
1	telephone	1
2	phone	1
3	number	0.875
4	service	0.875
5	system	0.5
6	communicati	1
7	phones	1
8	network	1
9	communicati	1
10	company	0.5
11	numbers	0.875
12	telecom	1
13	services	0.75
14	digit	0.875
15	telephones	1
16	telecommun	1
17	connect	0.75
18	public	0.375
19	national	0.125
20	state	0
		0.775

Machine	Term	Relatedness
1	machine	1
2	system	0.75
3	computer	0.875
4	machines	1
5	systems	0.625
6	number	0.375
7	large	0.375
8	design	0.5
9	general	0.25
10	language	0.125
11	limited	0.125
12	device	0.625
13	german	0.375
14	personal	0
15	hardware	0.875
16	change	0.125
17	translate	0.125
18	considered	0.125
19	produce	0.75
20	parts	0.75
		0.4875

Machine	Term	Relatedness
1	machines	1
2	machine	1
3	computer	0.875
4	process	0.625
5	system	0.75
6	produce	0.75
7	device	0.625
8	perform	0.5
9	design	0.5
10	called	0.125
11	state	0.375
12	general	0.25
13	model	0.625
14	application	0.375
15	number	0.375
16	complex	0.375
17	large	0.375
18	designed	0.25
19	mechanical	0.875
20	metal	0.5
		0.55625

Machine	Term	Relatedness
1	machines	1
2	machine	1
3	computer	0.875
4	device	0.625
5	produce	0.75
6	system	0.75
7	design	0.5
8	number	0.375
9	called	0.125
10	process	0.625
11	state	0.375
12	mechanical	0.875
13	designed	0.25
14	common	0
15	original	0.25
16	class	0.25
17	perform	0.5
18	model	0.625
19	release	0.25
20	large	0.375
		0.51875

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

Computer	Term	Relatedness	Computer	Term	Relatedness	Computer	Term	Relatedness
1	computer	1	1	computer	1	1	computer	1
2	computers	1	2	computers	1	2	computers	1
3	program	0.875	3	system	0.75	3	system	0.75
4	application	0.875	4	systems	0.875	4	systems	0.875
5	machine	0.875	5	program	0.875	5	design	0.25
6	solve	0.375	6	design	0.25	6	machine	0.875
7	world	0.125	7	include	0	7	process	0.625
8	electronic	0.75	8	process	0.625	8	include	0
9	applications	0.875	9	develop	0.5	9	program	0.875
10	process	0.625	10	based	0.125	10	based	0.125
11	system	0.75	11	electronic	0.75	11	format	0.75
12	state	0.375	12	computing	1	12	software	1
13	systems	0.875	13	device	0.75	13	hardware	1
14	develop	0.5	14	machine	0.875	14	computing	1
15	design	0.25	15	refer	0.125	15	develop	0.5
16	digital	0.75	16	software	1	16	digital	0.75
17	order	0.25	17	person	0	17	electronic	0.75
18	general	0	18	digital	0.75	18	perform	0.5
19	perform	0.5	19	require	0.125	19	early	0
20	problem	0.625	20	application	0.875	20	device	0.75
		0.6125			0.6125			0.66875

Space	Term	Relatedness	Space	Term	Relatedness	Space	Term	Relatedness
1	space	0.875	1	mathematics	0.875	1	mathematics	0.875
2	spaces	0.75	2	space	0.875	2	space	0.875
3	mathematics	0.875	3	spaces	0.75	3	spaces	0.75
4	point	0.625	4	logic	0.125	4	called	0
5	finite	0.5	5	called	0	5	point	0.625
6	concept	0.5	6	topological	0.5	6	topological	0.5
7	defined	0.375	7	system	0.875	7	general	0
8	euclidean	0.375	8	point	0.625	8	topology	0.25
9	topological	0.5	9	topology	0.25	9	system	0.875
10	number	0.375	10	structure	0.125	10	metric	0.5
11	dimension	0.625	11	earth	0.875	11	function	0.25
12	element	0.625	12	include	0.25	12	theory	0.625
13	euclid	0.375	13	general	0	13	related	0.25
14	define	0.375	14	element	0.625	14	structure	0.125
15	definition	0.25	15	orbit	0.625	15	element	0.625
16	system	0.875	16	related	0.25	16	algebra	0.375
17	metric	0.5	17	function	0.25	17	earth	0.875
18	mathematici	0	18	concept	0.5	18	vector	0.5
19	century	0	19	metric	0.5	19	points	0.375
20	short	0	20	theory	0.625	20	include	0.25
		0.46875			0.475			0.475

Game	Term	Relatedness	Game	Term	Relatedness	Game	Term	Relatedness
1	player	1	1	player	1	1	player	1
2	games	0.875	2	games	0.875	2	games	0.875
3	players	0.875	3	players	0.875	3	players	0.875
4	involve	0.5	4	video	1	4	video	1
5	playing	1	5	board	1	5	played	0.875
6	person	0.5	6	played	0.875	6	number	0.25
7	board	1	7	number	0.25	7	include	0.375
8	human	0.5	8	called	0.125	8	board	1
9	number	0.25	9	playing	1	9	general	0.125
10	played	0.875	10	general	0.125	10	playing	1
11	computer	1	11	require	0.125	11	called	0.125
12	early	0.125	12	common	0.25	12	computer	1
13	require	0.125	13	action	0.625	13	world	0.5
14	purpose	0.5	14	based	0.25	14	popular	0.375
15	rules	0.875	15	computer	1	15	action	0.625
16	range	0.25	16	include	0.375	16	theory	0.625
17	exist	0.375	17	involve	0.5	17	involve	0.5
18	present	0.375	18	chess	1	18	common	0.25
19	involves	0.5	19	popular	0.375	19	chess	1
20	distinction	0	20	theory	0.625	20	require	0.125
		0.575			0.6125			0.625

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

Power	Term	Relatedness	Power	Term	Relatedness	Power	Term	Relatedness
1	ratio	0.5	1	system	0.625	1	electric	0.75
2	power	0.875	2	power	0.875	2	power	0.875
3	define	0.125	3	electric	0.75	3	system	0.625
4	energy	1	4	ratio	0.5	4	energy	1
5	tribute	0.375	5	electrical	1	5	refer	0.125
6	electric	0.75	6	refer	0.125	6	large	0.25
7	number	0.375	7	state	0.25	7	electrical	1
8	generate	0.625	8	control	1	8	state	0.25
9	press	0.375	9	energy	1	9	source	1
10	system	0.625	10	large	0.25	10	america	0.125
11	defined	0.375	11	force	0.75	11	control	1
12	place	0	12	ability	0.25	12	include	0.25
13	additional	0	13	concept	0.125	13	general	0.125
14	instance	0	14	point	0.25	14	national	0.5
15	operation	0.5	15	general	0.125	15	transmission	0.625
16	equal	0.25	16	transmission	0.625	16	ability	0.25
17	developer	0.25	17	means	0.125	17	world	0.25
18	equivalent	0.25	18	require	0.375	18	american	0.125
19	developed	0.375	19	states	0.25	19	systems	0.625
20	works	0.375	20	include	0.25	20	called	0.125
		0.4			0.475			0.49375

Piano	Term	Relatedness	Piano	Term	Relatedness	Piano	Term	Relatedness
1	strings	0.875	1	music	1	1	music	1
2	piano	1	2	piano	1	2	piano	1
3	instrument	0.75	3	instrument	0.75	3	instrument	0.75
4	string	0.875	4	pianos	0.625	4	pianos	0.625
5	keyboard	0.625	5	played	0.75	5	pianist	1
6	system	0	6	perform	0.875	6	piece	0.5
7	italian	0.75	7	sound	1	7	sound	1
8	popular	0.125	8	instruments	0.875	8	instruments	0.875
9	board	0.125	9	pianist	1	9	written	0.125
10	music	1	10	keyboard	0.625	10	played	0.75
11	pianos	0.625	11	popular	0.125	11	released	0.25
12	sound	1	12	string	0.875	12	string	0.875
13	refer	0	13	produce	0.125	13	keyboard	0.625
14	acoustic	0	14	record	0.375	14	musician	0.625
15	classical	1	15	musical	0.625	15	album	0.375
16	press	0.25	16	feature	0.375	16	popular	0.125
17	instruments	0.875	17	musician	0.625	17	original	0.25
18	referred	0	18	piece	0.5	18	musical	0.625
19	hammers	0	19	performance	1	19	concert	1
20	range	0.625	20	range	0.625	20	refer	0
		0.525			0.6875			0.61875

Moon	Term	Relatedness	Moon	Term	Relatedness	Moon	Term	Relatedness
1	solar	0.875	1	american	0.125	1	released	0
2	occur	0.625	2	america	0.125	2	release	0
3	occurs	0.75	3	music	0.375	3	american	0.125
4	precisely	0.125	4	written	0.125	4	written	0.125
5	calendar	0.375	5	state	0.125	5	album	0.125
6	precise	0	6	fiction	0.125	6	record	0.375
7	phases	0.25	7	record	0.375	7	story	0.375
8	earth	0.375	8	character	0	8	fiction	0.125
9	calendars	0.25	9	lunar	1	9	state	0.125
10	eclipse	0.875	10	story	0.375	10	music	0.375
11	plane	0	11	space	0.75	11	published	0
12	phase	0.625	12	released	0	12	united	0
13	lunar	1	13	published	0	13	based	0
14	influence	0.375	14	occur	0.625	14	played	0.125
15	including	0	15	united	0	15	states	0.125
16	released	0	16	states	0.125	16	original	0.125
17	resulting	0.125	17	played	0.125	17	national	0
18	outer	0.25	18	based	0	18	member	0
19	closest	0	19	march	0.25	19	series	0
20	space	0.75	20	popular	0.125	20	character	0
		0.38125			0.2375			0.10625

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

Screen	Term	Relatedness	Screen	Term	Relatedness	Screen	Term	Relatedness
	1 screen	1		1 screen	1		1 screen	1
	2 screens	1		2 screens	1		2 screens	1
	3 original	0.125		3 include	0.25		3 vision	0.75
	4 present	0.625		4 design	0.75		4 include	0.25
	5 select	0.375		5 product	0.625		5 television	1
	6 content	0.75		6 system	0.375		6 product	0.625
	7 origin	0		7 cover	0.625		7 design	0.75
	8 sight	0.75		8 vision	0.75		8 display	1
	9 system	0.375		9 feature	0.5		9 produce	0.5
	10 technique	0.5		10 television	1		10 video	1
	11 support	0.125		11 movie	0.875		11 system	0.375
	12 design	0.75		12 video	1		12 production	0.5
	13 technology	0.875		13 material	0.125		13 movie	0.875
	14 interest	0.375		14 light	1		14 cinema	1
	15 visually	1		15 display	1		15 table	0.375
	16 function	0.25		16 function	0.25		16 media	1
	17 suitable	0.25		17 common	0.125		17 america	0
	18 number	0.125		18 image	1		18 based	0.125
	19 keyboard	0		19 table	0.375		19 national	0
	20 generally	0		20 motion	0.75		20 material	0.125
		0.4625			0.66875			0.6125

Car	Term	Relatedness	Car	Term	Relatedness	Car	Term	Relatedness
	1 automobile	1		1 vehicle	1		1 vehicles	1
	2 america	0.625		2 america	0.625		2 vehicle	1
	3 large	0.375		3 mobile	0.75		3 automobile	1
	4 vehicle	1		4 automobile	1		4 mobile	0.75
	5 american	0.125		5 vehicles	1		5 america	0.625
	6 europe	0.25		6 design	0.625		6 sport	0.5
	7 larger	0.375		7 large	0.375		7 include	0
	8 states	0.125		8 sport	0.5		8 large	0.375
	9 united	0		9 small	0.375		9 drive	1
	10 short	0.25		10 engine	1		10 design	0.625
	11 north	0		11 include	0		11 small	0.375
	12 light	0.25		12 general	0.375		12 general	0.375
	13 include	0		13 drive	1		13 built	0.5
	14 built	0.5		14 class	0.5		14 engine	1
	15 passenger	1		15 power	0.875		15 train	0.5
	16 carry	0.5		16 united	0		16 typical	0.125
	17 motor	1		17 designed	0.625		17 motor	1
	18 common	0.5		18 passenger	1		18 passenger	1
	19 years	0.375		19 american	0.125		19 power	0.875
	20 market	0.625		20 train	0.5		20 designed	0.625
		0.44375			0.6125			0.6625

Book	Term	Relatedness	Book	Term	Relatedness	Book	Term	Relatedness
	1 large	0.125		1 books	1		1 books	1
	2 books	1		2 format	0.875		2 print	0.875
	3 print	0.875		3 print	0.875		3 format	0.875
	4 store	0.875		4 general	0.125		4 general	0.125
	5 pages	1		5 large	0.125		5 pages	1
	6 printed	0.875		6 include	0.125		6 state	0.125
	7 publisher	1		7 collect	0.375		7 printed	0.875
	8 paper	0.875		8 refer	0.5		8 paper	0.875
	9 market	0.5		9 pages	1		9 cover	1
	10 bookseller	1		10 paper	0.875		10 large	0.125
	11 electronic	0.625		11 graph	0.125		11 include	0.125
	12 magazine	0.75		12 common	0.125		12 published	0.875
	13 longer	0.375		13 called	0		13 united	0
	14 written	1		14 printed	0.875		14 common	0.125
	15 series	0.875		15 generally	0		15 refer	0.5
	16 bookstore	1		16 cover	1		16 author	1
	17 stores	0.75		17 state	0.125		17 states	0
	18 organ	0		18 author	1		18 generally	0
	19 title	0		19 place	0.125		19 called	0
	20 purchase	0		20 small	0.125		20 public	0.5
		0.675			0.46875			0.5

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

Bike	Term	Relatedness
1	bicycle	1
2	cycle	1
3	bikes	1
4	purpose	0.125
5	cycling	1
6	typical	0.125
7	typically	0.125
8	large	0.125
9	drive	0.625
10	place	0.125
11	cyclist	1
12	depending	0
13	similar	0.125
14	train	0.375
15	handle	1
16	power	0.375
17	larger	0
18	cyclists	1
19	serve	0
20	states	0
		0.45625

Bike	Term	Relatedness
1	bicycle	1
2	cycle	1
3	bikes	1
4	cycling	1
5	event	0.125
6	motorcycle	0.875
7	design	0.625
8	rider	1
9	state	0.125
10	wheel	1
11	public	0.375
12	include	0.125
13	street	0.625
14	bicycles	1
15	vehicle	0.625
16	small	0.25
17	people	0.375
18	sport	0.5
19	united	0
20	riders	1
		0.63125

Bike	Term	Relatedness
1	bicycle	1
2	cycle	1
3	cycling	1
4	bikes	1
5	event	0.125
6	mountain	1
7	state	0.125
8	include	0.125
9	sport	0.5
10	cyclists	1
11	design	0.625
12	rider	1
13	trail	1
14	world	0.125
15	motorcycle	0.875
16	cross	0.375
17	public	0.375
18	country	0.125
19	south	0
20	riders	1
		0.61875

Soldier	Term	Relatedness
1	soldier	1
2	soldiers	1
3	album	0
4	produce	0
5	state	0.625
6	group	0.75
7	number	0.5
8	fight	1
9	states	0.5
10	force	0.875
11	originally	0
12	united	0.625
13	south	0
14	produced	0.125
15	project	0
16	figures	0.375
17	native	0
18	military	1
19	created	0
20	based	0.25
		0.43125
		6

Soldier	Term	Relatedness
1	soldier	1
2	release	0.125
3	world	0.375
4	soldiers	1
5	state	0.625
6	america	0.625
7	released	0.125
8	united	0.625
9	original	0
10	based	0.25
11	american	0.5
12	states	0.5
13	series	0.25
14	album	0
15	including	0
16	group	0.75
17	directed	0.5
18	national	0.75
19	south	0
20	music	0
		0.4
		7

Soldier	Term	Relatedness
1	soldier	1
2	release	0.125
3	released	0.125
4	america	0.625
5	soldiers	1
6	world	0.375
7	state	0.625
8	written	0
9	series	0.25
10	american	0.5
11	album	0
12	original	0
13	united	0.625
14	based	0.25
15	record	0
16	directed	0.5
17	states	0.5
18	including	0
19	title	0.5
20	north	0
		0.35
		7

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

LM

10 articles			75 articles			150 articles		
Cloud	**Term**	**Relatedness**	**Cloud**	**Term**	**Relatedness**	**Cloud**	**Term**	**Relatedness**
	1 served	0.25		1 clouds	1		1 clouds	1
	2 cloud	1		2 cloud	1		2 cloud	1
	3 depend	0.25		3 altitude	0.75		3 large	0.375
	4 limit	0		4 water	1		4 america	0
	5 article	0		5 large	0.375		5 water	1
	6 observations	0.75		6 cumulus	0.75		6 altitude	0.75
	7 clouds	1		7 precipitation	0.375		7 american	0
	8 temperature	0.875		8 generally	0.125		8 release	0
	9 occur	0.375		9 character	0		9 light	0.875
	10 water	1		10 surface	0.5		10 cumulus	0.75
	11 observation	0.75		11 vapor	0.875		11 common	0
	12 aerosol	0		12 common	0		12 service	0.5
	13 stellar	0		13 ground	0		13 based	0.125
	14 ionized	0		14 produce	0		14 system	0.25
	15 state	0		15 occur	0.375		15 released	0
	16 space	0		16 level	0.375		16 generally	0.125
	17 clusters	0		17 develop	0.125		17 level	0.375
	18 cluster	0		18 small	0.375		18 produce	0
	19 galaxies	0		19 system	0.25		19 formation	1
	20 confirmed	0		20 stratus	0.625		20 including	0
		0.3125			0.44375			0.40625
Cat	**Term**	**Relatedness**	**Cat**	**Term**	**Relatedness**	**Cat**	**Term**	**Relatedness**
	1 america	0		1 species	1		1 released	0.125
	2 vision	0.375		2 america	0		2 america	0
	3 species	1		3 origin	0.375		3 american	0
	4 common	0.125		4 small	0.375		4 small	0.375
	5 result	0.25		5 common	0.125		5 common	0.125
	6 early	0		6 character	0.625		6 origin	0.375
	7 popular	0.5		7 called	0		7 produce	0
	8 super	0.125		8 world	0.125		8 short	0.375
	9 specific	0.25		9 original	0.25		9 breed	0.875
	10 domestic	1		10 population	0.375		10 domestic	1
	11 sense	0.75		11 series	0		11 include	0
	12 similar	0.25		12 breed	0.875		12 called	0
	13 american	0		13 american	0		13 species	1
	14 resulting	0.375		14 state	0.125		14 character	0.625
	15 number	0.125		15 felis	0.5		15 number	0.125
	16 language	0		16 south	0		16 record	0
	17 artist	0.125		17 number	0.125		17 world	0.125
	18 experiment	0		18 domestic	1		18 state	0.125
	19 action	0		19 record	0		19 animal	1
	20 called	0		20 united	0		20 series	0
		0.2625			0.29375			0.3125
Weapon	**Term**	**Relatedness**	**Weapon**	**Term**	**Relatedness**	**Weapon**	**Term**	**Relatedness**
	1 weapon	1		1 weapon	1		1 weapon	1
	2 weapons	1		2 weapons	1		2 weapons	1
	3 projectile	1		3 design	0.375		3 design	0.375
	4 force	1		4 device	0.875		4 system	0
	5 device	0.875		5 military	1		5 develop	0.75
	6 modern	0.125		6 include	0		6 state	0.625
	7 advantage	0.25		7 designed	0.375		7 large	0.25
	8 produce	0.25		8 purpose	0.375		8 military	1
	9 warfare	1		9 power	1		9 states	0.625
	10 effect	0.5		10 similar	0.125		10 united	0.125
	11 missile	1		11 common	0.125		11 power	1
	12 great	0		12 refer	0.125		12 release	0.25
	13 purpose	0.375		13 state	0.625		13 force	1
	14 damage	1		14 explosive	0.875		14 include	0
	15 include	0		15 america	0.875		15 designed	0.375
	16 power	1		16 combat	0.875		16 device	0.875
	17 combat	0.875		17 intended	0.375		17 refer	0.125
	18 increase	0.25		18 modern	0.125		18 based	0
	19 implements	0		19 force	1		19 released	0.125
	20 explosive	0.875		20 person	0		20 explosive	0.875
		0.61875			0.55625			0.51875

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

School	Term	Relatedness	School	Term	Relatedness	School	Term	Relatedness
	1 school	1		1 education	1		1 education	1
	2 chool	0		2 school	1		2 school	1
	3 schools	1		3 schools	1		3 schools	1
	4 education	1		4 educational	1		4 educational	1
	5 schooling	1		5 student	1		5 student	1
	6 generally	0		6 students	1		6 students	1
	7 public	1		7 founded	0.5		7 independent	0.5
	8 institution	0.875		8 college	1		8 college	1
	9 private	1		9 state	0.875		9 state	0.875
	10 building	0.875		10 located	0.25		10 founded	0.5
	11 attend	1		11 independent	0.5		11 located	0.25
	12 economic	0.5		12 private	1		12 board	0.75
	13 include	0.625		13 secondary	0.875		13 england	0.25
	14 credit	0.875		14 grade	1		14 private	1
	15 century	0.25		15 england	0.25		15 preparatory	0.625
	16 common	0.625		16 campus	1		16 grade	1
	17 government	0.375		17 institution	0.875		17 boarding	0.75
	18 countries	0.375		18 member	0.25		18 include	0.625
	19 child	1		19 class	1		19 states	0.875
	20 children	1		20 board	0.75		20 secondary	0.875
		0.71875			0.80625			0.79375

Police	Term	Relatedness	Police	Term	Relatedness	Police	Term	Relatedness
	1 force	1		1 force	1		1 force	1
	2 police	1		2 police	1		2 police	1
	3 officers	1		3 office	0.875		3 office	0.875
	4 office	0.875		4 officer	1		4 officer	1
	5 state	1		5 state	1		5 officers	1
	6 officer	1		6 officers	1		6 state	1
	7 enforcement	0.875		7 enforce	0.875		7 crime	1
	8 enforce	0.875		8 crime	1		8 policing	0.875
	9 members	0.25		9 policing	0.875		9 enforcement	0.875
	10 units	1		10 service	1		10 service	1
	11 traffic	0.875		11 enforcement	0.875		11 forces	1
	12 service	1		12 person	0.75		12 territorial	0.5
	13 australia	0.25		13 forces	1		13 public	0.875
	14 search	0.875		14 united	0.75		14 agency	0.625
	15 agency	0.625		15 public	0.875		15 responsible	0.875
	16 forces	1		16 states	1		16 united	0.75
	17 special	0.375		17 common	0.375		17 special	0.375
	18 services	0.875		18 power	0.75		18 states	1
	19 civil	0.625		19 responsible	0.875		19 local	0.75
	20 motor	0.25		20 territorial	0.5		20 include	0.5
		0.78125			0.86875			0.84375

Fruit	Term	Relatedness	Fruit	Term	Relatedness	Fruit	Term	Relatedness
	1 apple	1		1 fruits	1		1 species	0.125
	2 fruit	1		2 fruit	1		2 fruit	1
	3 seeds	1		3 apple	1		3 fruits	1
	4 state	0		4 common	0		4 range	0.125
	5 sweet	0.875		5 sweet	0.875		5 large	0.125
	6 fruits	1		6 produce	0.375		6 produce	0.375
	7 apples	1		7 range	0.125		7 apple	1
	8 include	0		8 include	0		8 include	0
	9 strawberries	1		9 called	0		9 united	0
	10 pears	1		10 orange	1		10 sweet	0.875
	11 flesh	0.5		11 similar	0.25		11 product	0.375
	12 varies	0.125		12 sugar	0.625		12 small	0.125
	13 today	0.125		13 product	0.375		13 orange	1
	14 specific	0.125		14 released	0		14 called	0
	15 produce	0.375		15 large	0.125		15 plant	1
	16 orange	1		16 album	0		16 america	0
	17 called	0		17 flavor	0.875		17 south	0.125
	18 flower	0.875		18 based	0		18 eaten	0.75
	19 states	0		19 version	0		19 common	0
	20 taste	0		20 flavour	0.875		20 natural	0.75
		0.55			0.425			0.4375

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

Meat	Term	Relatedness
1	refer	0.125
2	process	0.5
3	america	0.375
4	species	0.625
5	american	0.375
6	served	0.875
7	protein	0.875
8	world	0
9	include	0
10	album	0
11	product	0.625
12	school	0
13	vocal	0
14	parts	0.375
15	consume	0
16	concerns	0.375
17	water	0.375
18	significant	0
19	animal	1
20	source	0.25
		0.3375

Meat	Term	Relatedness
1	animal	1
2	include	0
3	refer	0.125
4	process	0.5
5	animals	1
6	consumption	1
7	world	0
8	method	0.125
9	product	0.625
10	common	0
11	produce	0.5
12	meats	1
13	called	0
14	united	0
15	cooked	1
16	human	0.375
17	including	0.125
18	states	0
19	america	0.375
20	large	0.125
		0.39375

Meat	Term	Relatedness
1	animal	1
2	product	0.625
3	process	0.5
4	include	0
5	state	0
6	produce	0.5
7	release	0
8	united	0
9	released	0
10	states	0
11	world	0
12	album	0
13	including	0.125
14	america	0.375
15	large	0.125
16	ground	0.375
17	animals	1
18	called	0
19	products	0.625
20	consumption	1
		0.3125

Telephone	Term	Relatedness
1	telephone	1
2	phone	1
3	service	0.875
4	phones	1
5	telephones	1
6	electric	0.75
7	system	0.5
8	connect	0.75
9	internet	0.75
10	device	0.75
11	network	1
12	called	0.75
13	electrical	0.625
14	number	0.875
15	switch	0.25
16	access	0.5
17	small	0.5
18	services	0.75
19	converts	0.125
20	invented	0
		0.6875

Telephone	Term	Relatedness
1	telephone	1
2	phone	1
3	service	0.875
4	phones	1
5	connect	0.75
6	number	0.875
7	telephones	1
8	communicati	1
9	system	0.5
10	company	0.5
11	network	1
12	communicati	1
13	telecommun	1
14	services	0.75
15	common	0.125
16	exchange	0.25
17	electric	0.75
18	switch	0.25
19	place	0.375
20	state	0
		0.7

Telephone	Term	Relatedness
1	telephone	1
2	phone	1
3	number	0.875
4	service	0.875
5	system	0.5
6	communicati	1
7	phones	1
8	network	1
9	communicati	1
10	company	0.5
11	numbers	0.875
12	telecom	1
13	services	0.75
14	digit	0.875
15	telephones	1
16	telecommun	1
17	connect	0.75
18	public	0.375
19	national	0.125
20	state	0
		0.775

Machine	Term	Relatedness
1	machine	1
2	system	0.75
3	computer	0.875
4	machines	1
5	number	0.375
6	systems	0.625
7	general	0.25
8	design	0.5
9	large	0.375
10	language	0.125
11	considered	0.125
12	fully	0
13	power	0.5
14	require	0
15	effective	0
16	application	0.375
17	level	0.375
18	ratio	0.25
19	change	0.125
20	government	0.125
		0.3875

Machine	Term	Relatedness
1	machines	1
2	machine	1
3	computer	0.875
4	process	0.625
5	system	0.75
6	produce	0.75
7	device	0.625
8	perform	0.5
9	design	0.5
10	called	0.125
11	state	0.375
12	general	0.25
13	model	0.625
14	designed	0.25
15	application	0.375
16	large	0.375
17	number	0.375
18	complex	0.375
19	original	0.25
20	mechanical	0.875
		0.54375

Machine	Term	Relatedness
1	machines	1
2	machine	1
3	computer	0.875
4	device	0.625
5	produce	0.75
6	system	0.75
7	design	0.5
8	number	0.375
9	process	0.625
10	state	0.375
11	called	0.125
12	designed	0.25
13	mechanical	0.875
14	common	0
15	class	0.25
16	model	0.625
17	perform	0.5
18	original	0.25
19	large	0.375
20	release	0.25
		0.51875

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

Computer	Term	Relatedness
1	computer	1
2	computers	1
3	program	0.875
4	application	0.875
5	electronic	0.75
6	machine	0.875
7	applications	0.875
8	process	0.625
9	system	0.75
10	world	0.125
11	solve	0.375
12	design	0.25
13	digital	0.75
14	general	0
15	develop	0.5
16	problem	0.625
17	systems	0.875
18	state	0.375
19	perform	0.5
20	order	0.25
		0.6125

Computer	Term	Relatedness
1	computer	1
2	computers	1
3	system	0.75
4	systems	0.875
5	program	0.875
6	design	0.25
7	include	0
8	process	0.625
9	based	0.125
10	develop	0.5
11	electronic	0.75
12	computing	1
13	device	0.75
14	refer	0.125
15	software	1
16	machine	0.875
17	digital	0.75
18	person	0
19	application	0.875
20	signed	0.125
		0.6125

Computer	Term	Relatedness
1	computer	1
2	computers	1
3	system	0.75
4	systems	0.875
5	design	0.25
6	machine	0.875
7	include	0
8	process	0.625
9	program	0.875
10	based	0.125
11	computing	1
12	hardware	1
13	software	1
14	format	0.75
15	develop	0.5
16	electronic	0.75
17	digital	0.75
18	perform	0.5
19	device	0.75
20	early	0
		0.66875

Space	Term	Relatedness
1	mathematics	0.875
2	space	0.875
3	spaces	0.75
4	point	0.625
5	define	0.375
6	concept	0.5
7	dimension	0.625
8	element	0.625
9	system	0.875
10	definition	0.25
11	euclidean	0.375
12	finite	0.5
13	number	0.375
14	topological	0.5
15	euclid	0.375
16	defined	0.375
17	elements	0
18	typical	0
19	sequence	0.125
20	introduced	0
		0.45

Space	Term	Relatedness
1	mathematics	0.875
2	space	0.875
3	spaces	0.75
4	logic	0.125
5	called	0
6	topological	0.5
7	point	0.625
8	system	0.875
9	topology	0.25
10	structure	0.125
11	earth	0.875
12	general	0
13	include	0.25
14	element	0.625
15	related	0.25
16	function	0.25
17	orbit	0.625
18	exist	0.375
19	metric	0.5
20	theory	0.625
		0.46875

Space	Term	Relatedness
1	mathematics	0.875
2	space	0.875
3	spaces	0.75
4	called	0
5	point	0.625
6	topological	0.5
7	general	0
8	topology	0.25
9	system	0.875
10	metric	0.5
11	function	0.25
12	theory	0.625
13	related	0.25
14	algebra	0.375
15	element	0.625
16	structure	0.125
17	earth	0.875
18	vector	0.5
19	points	0.375
20	include	0.25
		0.475

Game	Term	Relatedness
1	player	1
2	games	0.875
3	involve	0.5
4	players	0.875
5	playing	1
6	person	0.5
7	board	1
8	played	0.875
9	early	0.125
10	number	0.25
11	human	0.5
12	require	0.125
13	computer	1
14	rules	0.875
15	range	0.25
16	exist	0.375
17	purpose	0.5
18	present	0.375
19	interaction	0
20	strategy	0
		0.55

Game	Term	Relatedness
1	player	1
2	games	0.875
3	players	0.875
4	video	1
5	board	1
6	number	0.25
7	played	0.875
8	called	0.125
9	playing	1
10	require	0.125
11	general	0.125
12	action	0.625
13	common	0.25
14	include	0.375
15	based	0.25
16	computer	1
17	involve	0.5
18	popular	0.375
19	chess	1
20	theory	0.625
		0.6125

Game	Term	Relatedness
1	player	1
2	games	0.875
3	players	0.875
4	video	1
5	played	0.875
6	number	0.25
7	include	0.375
8	board	1
9	general	0.125
10	playing	1
11	called	0.125
12	computer	1
13	world	0.5
14	popular	0.375
15	action	0.625
16	theory	0.625
17	involve	0.5
18	common	0.25
19	chess	1
20	require	0.125
		0.625

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

Power		Term	Relatedness
	1	ratio	0.5
	2	power	0.875
	3	system	0.625
	4	define	0.125
	5	generate	0.625
	6	electric	0.75
	7	electrical	1
	8	number	0.375
	9	energy	1
	10	output	0.5
	11	measure	0.5
	12	defined	0.375
	13	distribution	0.5
	14	total	0
	15	signal	1
	16	person	0
	17	years	0
	18	works	0.375
	19	generation	0.5
	20	measured	0.5
			0.50625

Power		Term	Relatedness
	1	system	0.625
	2	power	0.875
	3	electric	0.75
	4	refer	0.125
	5	electrical	1
	6	ratio	0.5
	7	control	1
	8	energy	1
	9	state	0.25
	10	force	0.75
	11	large	0.25
	12	ability	0.25
	13	concept	0.125
	14	point	0.25
	15	general	0.125
	16	transmission	0.625
	17	states	0.25
	18	require	0.375
	19	means	0.125
	20	developed	0.375
			0.48125

Power		Term	Relatedness
	1	electric	0.75
	2	power	0.875
	3	system	0.625
	4	energy	1
	5	large	0.25
	6	refer	0.125
	7	electrical	1
	8	state	0.25
	9	source	1
	10	america	0.125
	11	control	1
	12	include	0.25
	13	general	0.125
	14	national	0.5
	15	ability	0.25
	16	transmission	0.625
	17	world	0.25
	18	company	0.375
	19	called	0.125
	20	american	0.125
			0.48125

Piano		Term	Relatedness
	1	strings	0.875
	2	piano	1
	3	string	0.875
	4	instrument	0.75
	5	board	0.125
	6	pianos	0.625
	7	system	0
	8	sound	1
	9	popular	0.125
	10	keyboard	0.625
	11	italian	0.75
	12	music	1
	13	refer	0
	14	similar	0
	15	octaves	1
	16	instruments	0.875
	17	range	0.625
	18	requires	0
	19	padded	0.125
	20	store	0
			0.51875

Piano		Term	Relatedness
	1	music	1
	2	piano	1
	3	instrument	0.75
	4	pianos	0.625
	5	played	0.75
	6	perform	0.875
	7	instruments	0.875
	8	sound	1
	9	pianist	1
	10	keyboard	0.625
	11	popular	0.125
	12	string	0.875
	13	produce	0.125
	14	record	0.375
	15	feature	0.375
	16	musical	0.625
	17	musician	0.625
	18	piece	0.5
	19	performance	1
	20	range	0.625
			0.6875

Piano		Term	Relatedness
	1	music	1
	2	piano	1
	3	instrument	0.75
	4	pianos	0.625
	5	pianist	1
	6	piece	0.5
	7	sound	1
	8	instruments	0.875
	9	written	0.125
	10	played	0.75
	11	released	0.25
	12	string	0.875
	13	keyboard	0.625
	14	musician	0.625
	15	album	0.375
	16	original	0.25
	17	musical	0.625
	18	popular	0.125
	19	concert	1
	20	refer	0
			0.61875

Moon		Term	Relatedness
	1	solar	0.875
	2	occur	0.625
	3	occurs	0.75
	4	earth	0.375
	5	eclipse	0.875
	6	calendars	0.25
	7	calendar	0.375
	8	plane	0
	9	phases	0.25
	10	lunar	1
	11	precisely	0.125
	12	phase	0.625
	13	precise	0
	14	influence	0.375
	15	including	0
	16	released	0
	17	outer	0.25
	18	orbit	0.75
	19	planet	0.625
	20	ecliptic	0.125
			0.4125

Moon		Term	Relatedness
	1	american	0.125
	2	america	0.125
	3	music	0.375
	4	written	0.125
	5	state	0.125
	6	fiction	0.125
	7	character	0
	8	record	0.375
	9	lunar	1
	10	story	0.375
	11	released	0
	12	space	0.75
	13	published	0
	14	occur	0.625
	15	states	0.125
	16	united	0
	17	played	0.125
	18	based	0
	19	march	0.25
	20	popular	0.125
			0.2375

Moon		Term	Relatedness
	1	released	0
	2	release	0
	3	american	0.125
	4	written	0.125
	5	album	0.125
	6	record	0.375
	7	story	0.375
	8	fiction	0.125
	9	state	0.125
	10	music	0.375
	11	published	0
	12	united	0
	13	based	0
	14	played	0.125
	15	states	0.125
	16	original	0.125
	17	national	0
	18	member	0
	19	series	0
	20	character	0
			0.10625

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

Screen	Term	Relatedness	Screen	Term	Relatedness	Screen	Term	Relatedness
	1 screen	1		1 screen	1		1 screen	1
	2 screens	1		2 screens	1		2 screens	1
	3 sight	0.75		3 include	0.25		3 vision	0.75
	4 system	0.375		4 design	0.75		4 include	0.25
	5 technique	0.5		5 product	0.625		5 television	1
	6 original	0.125		6 system	0.375		6 product	0.625
	7 content	0.75		7 cover	0.625		7 design	0.75
	8 select	0.375		8 vision	0.75		8 display	1
	9 present	0.625		9 feature	0.5		9 produce	0.5
	10 origin	0		10 television	1		10 system	0.375
	11 front	0.125		11 material	0.125		11 video	1
	12 design	0.75		12 movie	0.875		12 media	1
	13 range	0.375		13 video	1		13 table	0.375
	14 common	0.125		14 function	0.25		14 cinema	1
	15 material	0.125		15 display	1		15 production	0.5
	16 support	0.125		16 common	0.125		16 america	0
	17 print	0.5		17 light	1		17 movie	0.875
	18 visual	1		18 image	1		18 material	0.125
	19 cinema	1		19 table	0.375		19 national	0
	20 surface	0.875		20 production	0.5		20 based	0.125
		0.525			0.65625			0.6125

Car	Term	Relatedness	Car	Term	Relatedness	Car	Term	Relatedness
	1 automobile	1		1 vehicle	1		1 vehicles	1
	2 america	0.625		2 america	0.625		2 vehicle	1
	3 large	0.375		3 automobile	1		3 automobile	1
	4 american	0.125		4 mobile	0.75		4 mobile	0.75
	5 vehicle	1		5 vehicles	1		5 america	0.625
	6 europe	0.25		6 design	0.625		6 sport	0.5
	7 larger	0.375		7 large	0.375		7 include	0
	8 united	0		8 sport	0.5		8 large	0.375
	9 states	0.125		9 small	0.375		9 drive	1
	10 short	0.25		10 engine	1		10 design	0.625
	11 north	0		11 include	0		11 small	0.375
	12 light	0.25		12 general	0.375		12 general	0.375
	13 include	0		13 drive	1		13 built	0.5
	14 built	0.5		14 class	0.5		14 engine	1
	15 carry	0.5		15 power	0.875		15 train	0.5
	16 passenger	1		16 united	0		16 motor	1
	17 motor	1		17 designed	0.625		17 typical	0.125
	18 common	0.5		18 passenger	1		18 passenger	1
	19 family	0		19 train	0.5		19 power	0.875
	20 customers	0		20 american	0.125		20 designed	0.625
		0.39375			0.6125			0.6625

Book	Term	Relatedness	Book	Term	Relatedness	Book	Term	Relatedness
	1 large	0.125		1 books	1		1 books	1
	2 books	1		2 format	0.875		2 print	0.875
	3 print	0.875		3 print	0.875		3 format	0.875
	4 store	0.875		4 general	0.125		4 general	0.125
	5 printed	0.875		5 large	0.125		5 pages	1
	6 publisher	1		6 include	0.125		6 state	0.125
	7 market	0.5		7 collect	0.375		7 paper	0.875
	8 pages	1		8 refer	0.5		8 printed	0.875
	9 paper	0.875		9 pages	1		9 large	0.125
	10 bookstore	1		10 paper	0.875		10 cover	1
	11 bookseller	1		11 graph	0.125		11 include	0.125
	12 electronic	0.625		12 common	0.125		12 published	0.875
	13 series	0.875		13 called	0		13 united	0
	14 written	1		14 printed	0.875		14 common	0.125
	15 longer	0.375		15 generally	0		15 refer	0.5
	16 magazine	0.75		16 cover	1		16 author	1
	17 stores	0.75		17 state	0.125		17 states	0
	18 organ	0		18 author	1		18 generally	0
	19 cover	1		19 place	0.125		19 called	0
	20 united	0		20 small	0.125		20 public	0.5
		0.725			0.46875			0.5

Appendix B: Chi-Square Mutual Information: Relatedness of all terms from all queries for all sizes of N

Bike	Term	Relatedness	Bike	Term	Relatedness	Bike	Term	Relatedness
	1 bicycle	1		1 bicycle	1		1 bicycle	1
	2 cycle	1		2 cycle	1		2 cycle	1
	3 bikes	1		3 bikes	1		3 cycling	1
	4 purpose	0.125		4 cycling	1		4 bikes	1
	5 place	0.125		5 event	0.125		5 event	0.125
	6 drive	0.625		6 motorcycle	0.875		6 mountain	1
	7 large	0.125		7 design	0.625		7 state	0.125
	8 typical	0.125		8 rider	1		8 include	0.125
	9 cycling	1		9 state	0.125		9 sport	0.5
	10 typically	0.125		10 wheel	1		10 cyclists	1
	11 cyclist	1		11 public	0.375		11 rider	1
	12 created	0		12 include	0.125		12 design	0.625
	13 refer	0		13 street	0.625		13 trail	1
	14 train	0.375		14 bicycles	1		14 world	0.125
	15 power	0.375		15 vehicle	0.625		15 motorcycle	0.875
	16 vehicle	0.625		16 small	0.25		16 cross	0.375
	17 variety	0.125		17 people	0.375		17 public	0.375
	18 purposes	0.125		18 sport	0.5		18 country	0.125
	19 country	0.125		19 united	0		19 south	0
	20 pedal	1		20 riders	1		20 riders	1
		0.45			0.63125			0.61875

Soldier	Term	Relatedness	Soldier	Term	Relatedness	Soldier	Term	Relatedness
	1 soldier	1		1 world	0.375		1 soldier	1
	2 soldiers	1		2 soldier	1		2 release	0.125
	3 produce	0		3 release	0.125		3 released	0.125
	4 state	0.625		4 soldiers	1		4 america	0.625
	5 album	0		5 america	0.625		5 soldiers	1
	6 fight	1		6 state	0.625		6 world	0.375
	7 states	0.5		7 released	0.125		7 state	0.625
	8 force	0.875		8 united	0.625		8 written	0
	9 united	0.625		9 original	0		9 series	0.25
	10 group	0.75		10 based	0.25		10 american	0.5
	11 number	0.5		11 states	0.5		11 original	0
	12 originally	0		12 american	0.5		12 album	0
	13 produced	0.125		13 directed	0.5		13 united	0.625
	14 project	0		14 album	0		14 based	0.25
	15 early	0		15 including	0		15 record	0
	16 figures	0.375		16 group	0.75		16 directed	0.5
	17 created	0		17 series	0.25		17 including	0
	18 based	0.25		18 national	0.75		18 states	0.5
	19 record	0		19 include	0		19 north	0
	20 sergeant	1		20 force	0.875		20 title	0.5
		0.43125			0.44375			0.35
		5			**8**			**7**

10 Appendix C: P@n values for CSMI and NT extrapolated term relatedness

Table 9: CSMI values of P@n and AP with query = "Cloud"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.825	0.825	0.825
P@10	0.588	0.538	0.588
P@15	0.45	0.475	0.475
P@20	0.444	0.494	0.444
AP	0.632	0.647	0.65

Table 10: CSMI values of P@n and AP with query ="Cat"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.375	0.375	0.375
P@10	0.338	0.325	0.325
P@15	0.292	0.317	0.317
P@20	0.3	0.294	0.294
AP	0.379	0.385	0.38

Table 11: CSMI values of P@n and AP with query = "Weapon"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.85	0.85	0.85
P@10	0.613	0.613	0.613
P@15	0.6	0.55	0.583
P@20	0.581	0.556	0.556
AP	0.68	0.671	0.67

Table 12: CSMI values of P@n and AP with query = "School"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	1	1	1
P@10	0.863	0.863	0.863
P@15	0.8	0.858	0.817
P@20	0.769	0.788	0.806
AP	0.873	0.905	0.896

Table 13: CSMI values of P@n and AP with query = "Police"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.975	0.975	0.975
P@10	0.95	0.963	0.963
P@15	0.908	0.933	0.925
P@20	0.85	0.856	0.869
AP	0.934	0.946	0.945

Table 14: CSMI values of P@n and AP with query = "Fruit"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.969	0.75	0.75
P@10	0.538	0.463	0.538
P@15	0.45	0.45	0.45
P@20	0.425	0.383	0.425
AP	0.606	0.591	0.601

Table 15: CSMI values of P@n and AP with query ="Meat"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.525	0.525	0.525
P@10	0.438	0.438	0.438
P@15	0.483	0.458	0.458
P@20	0.406	0.394	0.394
AP	0.495	0.487	0.484

Table 16: CSMI values of P@n and AP with query = "Telephone"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.925	0.925	0.925
P@10	0.85	0.85	0.85
P@15	0.825	0.825	0.825
P@20	0.7	0.7	0.7
AP	0.872	0.872	0.873

Table 17: CSMI values of P@n and AP with query = "Machine"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.75	0.85	0.85
P@10	0.675	0.675	0.675
P@15	0.55	0.583	0.575
P@20	0.53125	0.55625	0.54375
AP	0.65	0.717	0.713

Table 18: CSMI values of P@n and AP with query = "Computer"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.75	0.9	0.9
P@10	0.6	0.6	0.6
P@15	0.633	0.633	0.642
P@20	0.606	0.613	0.613
AP	0.662	0.722	0.722

Table 19: CSMI values of P@n and AP with query = "Space"

P@n	Vector	BM25	LM
P@1	0.875	0.875	0.875
P@5	0.625	0.525	0.525
P@10	0.538	0.5	0.5
P@15	0.508	0.492	0.467
P@20	0.419	0.475	0.469
AP	0.556	0.566	0.56

Table 20: CSMI values of P@n and AP with query = "Game"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.95	0.95	0.95
P@10	0.713	0.713	0.713
P@15	0.633	0.625	0.583
P@20	0.613	0.613	0.613
AP	0.749	0.748	0.741

Table 21: CSMI values of P@n and AP with query = "Power"

P@n	Vector	BM25	LM
P@1	0.75	0.625	0.625
P@5	0.75	0.75	0.675
P@10	0.638	0.638	0.688
P@15	0.525	0.525	0.525
P@20	0.475	0.475	0.481
AP	0.645	0.609	0.614

Table 22: CSMI values of P@n and AP with query = "Piano"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.85	0.825	0.825
P@10	0.85	0.85	0.85
P@15	0.75	0.708	0.692
P@20	0.731	0.688	0.688
AP	0.818	0.801	0.799

Table 23: CSMI values of P@n and AP with query = "Moon"

P@n	Vector	BM25	LM
P@1	0.125	0.125	0.125
P@5	0.075	0.175	0.175
P@10	0.15	0.275	0.275
P@15	0.133	0.275	0.283
P@20	0.225	0.238	0.238
AP	0.146	0.23	0.224

Table 24: CSMI values of P@n and AP with query = "Screen"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.725	0.725	0.725
P@10	0.538	0.688	0.688
P@15	0.525	0.725	0.675
P@20	0.625	0.669	0.656
AP	0.634	0.724	0.709

Table 25: CSMI values of P@n and AP with query = "Car"

P@n	Vector	BM25	LM
P@1	0.75	1	1
P@5	0.875	0.875	0.875
P@10	0.7375	0.725	0.725
P@15	0.608	0.667	0.667
P@20	0.581	0.613	0.613
AP	0.713	0.724	0.729

Table 26: CSMI values of P@n and AP with query = "Book"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.6	0.6	0.6
P@10	0.563	0.588	0.588
P@15	0.467	0.467	0.467
P@20	0.488	0.469	0.469
AP	0.583	0.589	0.589

Table 27: CSMI values of P@n and AP with query = "Bike"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.975	0.825	0.825
P@10	0.725	0.775	0.775
P@15	0.667	0.7	0.7
P@20	0.656	0.631	0.631
AP	0.792	0.777	0.777

Table 28: CSMI values of P@n and AP with query = "Soldier"

P@n	Vector	BM25	LM
P@1	1	1	0.375
P@5	0.525	0.625	0.625
P@10	0.475	0.475	0.475
P@15	0.483	0.4	0.417
P@20	0.4	0.4	0.444
AP	0.525	0.52	0.5

Table 29: NT values of P@n and AP with query = "Cloud"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.675	0.675	0.675
P@10	0.475	0.475	0.475
P@15	0.425	0.425	0.425
P@20	0.444	0.444	0.444
AP	0.555	0.555	0.555

Table 30: NT values of P@n and AP with query ="Cat"

P@n	Vector	BM25	LM
P@1	0.125	0.125	0.125
P@5	0.25	0.25	0.25
P@10	0.1875	0.1875	0.1875
P@15	0.258	0.258	0.258
P@20	0.225	0.225	0.225
AP	0.211	0.211	0.211

Table 31: NT values of P@n and AP with query = "Weapon"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.8	0.8	0.8
P@10	0.7125	0.7125	0.7125
P@15	0.642	0.642	0.642
P@20	0.65	0.65	0.65
AP	0.754	0.752	0.752

Table 32: NT values of P@n and AP with query = "School"

P@n	Vector	BM25	LM
P@1	0.75	0.625	0.75
P@5	0.875	0.875	0.875
P@10	0.788	0.763	0.75
P@15	0.625	0.617	0.617
P@20	0.6	0.566	0.566
AP	0.746	0.716	0.731

Table 33: NT values of P@n and AP with query = "Police"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.9	0.9	0.9
P@10	0.863	0.838	0.838
P@15	0.708	0.642	0.642
P@20	0.656	0.656	0.656
AP	0.822	0.804	0.804

Table 34: NT values of P@n and AP with query = "Fruit"

P@n	Vector	BM25	LM
P@1	0	0	0
P@5	0.45	0.35	0.35
P@10	0.375	0.375	0.375
P@15	0.4	0.408	0.408
P@20	0.35	0.4	0.4
AP	0.324	0.364	0.364

Table 35: NT values of P@n and AP with query ="Meat"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.6	0.6	0.6
P@10	0.525	0.525	0.525
P@15	0.475	0.475	0.475
P@20	0.488	0.488	0.488
AP	0.55	0.55	0.55

Table 36: NT values of P@n and AP with query = "Telephone"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.475	0.475	0.475
P@10	0.563	0.563	0.563
P@15	0.6	0.6	0.6
P@20	0.544	0.581	0.581
AP	0.577	0.579	0.579

Table 37: NT values of P@n and AP with query = "Machine"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.55	0.55	0.55
P@10	0.425	0.425	0.425
P@15	0.425	0.425	0.425
P@20	0.413	0.413	0.413
AP	0.514	0.514	0.514

Table 38: NT values of P@n and AP with query = "Computer"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.825	0.825	0.825
P@10	0.863	0.863	0.863
P@15	0.833	0.758	0.758
P@20	0.744	0.706	0.706
AP	0.85	0.812	0.812

Table 39: NT values of P@n and AP with query = "Space"

P@n	Vector	BM25	LM
P@1	0.875	0.875	0.875
P@5	0.65	0.65	0.65
P@10	0.575	0.575	0.575
P@15	0.533	0.525	0.525
P@20	0.5	0.544	0.544
AP	0.58	0.596	0.596

Table 40: NT values of P@n and AP with query = "Game"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.825	0.825	0.825
P@10	0.6	0.663	0.663
P@15	0.625	0.658	0.658
P@20	0.66875	0.675	0.675
AP	0.732	0.74	0.74

Table 41: NT values of P@n and AP with query = "Power"

P@n	Vector	BM25	LM
P@1	0.75	0.75	0.75
P@5	0.85	0.85	0.85
P@10	0.75	0.75	0.75
P@15	0.642	0.683	0.683
P@20	0.681	0.713	0.713
AP	0.724	0.74	0.748

Table 42: NT values of P@n and AP with query = "Piano"

P@n	Vector	BM25	LM
P@1	1	0.75	0.75
P@5	0.7	0.65	0.65
P@10	0.688	0.663	0.663
P@15	0.675	0.642	0.642
P@20	0.688	0.669	0.669
AP	0.721	0.636	0.636

Table 43: NT values of P@n and AP with query = "Moon"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.85	0.825	0.825
P@10	0.65	0.65	0.65
P@15	0.555	0.57	0.57
P@20	0.525	0.481	0.481
AP	0.67	0.675	0.675

Table 44: NT values of P@n and AP with query = "Screen"

P@n	Vector	BM25	LM
P@1	0.25	0.25	0.25
P@5	0.475	0.475	0.475
P@10	0.588	0.588	0.588
P@15	0.517	0.567	0.567
P@20	0.544	0.588	0.588
AP	0.55	0.561	0.561

Table 45: NT values of P@n and AP with query = "Car"

P@n	Vector	BM25	LM
P@1	0.5	1	1
P@5	0.625	0.675	0.675
P@10	0.6	0.563	0.563
P@15	0.533	0.567	0.567
P@20	0.594	0.569	0.569
AP	0.593	0.601	0.601

Table 46: NT values of P@n and AP with query = "Book"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.45	0.275	0.275
P@10	0.263	0.263	0.263
P@15	0.367	0.3	0.3
P@20	0.356	0.369	0.369
AP	0.386	0.368	0.368

Table 47: NT values of P@n and AP with query = "Bike"

P@n	Vector	BM25	LM
P@1	1	1	1
P@5	0.625	0.625	0.625
P@10	0.713	0.713	0.713
P@15	0.583	0.583	0.583
P@20	0.556	0.556	0.556
AP	0.645	0.645	0.645

Table 48: NT values of P@n and AP with query = "Soldier"

P@n	Vector	BM25	LM
P@1	0.375	0.375	0.375
P@5	0.0.35	0.35	0.35
P@10	0.5	0.425	0.425
P@15	0.492	0.475	0.475
P@20	0.425	0.406	0.406
AP	0.458	0.434	0.434

11 Appendix D: Results of Image Retrieval using CSMI and NT

Table 49: Results of image retrieval using CSMI for query = "Cloud"

Query: cloud ET: clouds water altitude large cumulus		
Image #	Related?	Contains ET
1	Yes	cloud large cumulus
2	Yes	clouds cumulus large
3	Yes	cumulus cloud
4	Yes	cloud clouds cumulus
5	Yes	clouds cumulus cloud
6	Yes	water clouds cumulus large
7	Yes	cloud cumulus
8	Yes	clouds cumulus
9	Yes	altitude clouds
10	Yes	clouds cumulus

Table 50: Results of image retrieval using CSMI for query = "cat"

Query: cat ET: species america origin small		
Image #	Related image?	Contains ET
1	No	origin species
2	No	origin species
3	No	small species
4	Yes	cat small
5	Yes	cat small
6	Yes	america cat
7	No	origin
8	No	origin
9	Yes	cat small
10	Yes	cat small

Table 51: Results of image retrieval using CSMI for query = "weapon"

Query: weapon ET: weapons design military device include		
Image #	Related image?	Contains ET
1	Yes	military weapon
2	Yes	weapon weapons
3	Yes	weapon weapons
4	Yes	weapon weapons
5	Yes	weapon weapons
6	No	device
7	No	device
8	Yes	military weapon
9	No	design device
10	Yes	military weapons

Table 52: Results of image retrieval using CSMI for query = "school"

Query: school ET: education schools educational students student		
Image #	Related image?	Contains ET
1	Yes	school students education
2	Yes	school education students
3	Yes	education students
4	Yes	education student
5	Yes	education student school
6	Yes	school education student
7	Yes	school education student
8	Yes	education school students
9	Yes	students school
10	Yes	school students

Table 53: Results of image retrieval using CSMI for query = "police"

Query: police ET: force office officer officers state		
Image #	Related image?	Contains ET

1	Yes	police officers
2	Yes	police officer
3	Yes	police officer
4	Yes	officer police
5	Yes	police officer
6	Yes	police officers
7	No	officer officers
8	No	officer officers
9	Yes	police officer
10	Yes	police officer

Table 54: Results of image retrieval using CSMI for query = "fruit"

Query: fruit ET: fruits apple common sweet produce		
Image #	Related image?	Contains ET
1	Yes	apple fruits fruit
2	Yes	fruit produce sweet
3	Yes	fruit produce
4	Yes	fruit fruits
5	Yes	fruit produce
6	Yes	produce fruit
7	Yes	fruit produce
8	Yes	fruit produce
9	Yes	fruit produce
10	Yes	fruit produce

Table 55: Results of image retrieval using CSMI for query = "meat"

Query: meat ET: animal include refer process animals		
Image #	Related image?	Contains ET
1	Yes	animals meat
2	No	animal animals
3	No	animals animal
4	No	animals animal
5	Yes	include meat

6	Yes	meat animal
7	No	animal animals
8	No	animals animal
9	No	animal animals
10	No	animal animals

Table 56: Results of image retrieval using CSMI for query = "telephone"

Query: telephone ET: phone service phones connect number		
Image #	Related image?	Contains ET
1	Yes	phone phones telephone
2	Yes	phone telephone
3	Yes	number telephone phone
4	Yes	phone phones
5	Yes	telephone phone
6	Yes	telephone phone
7	Yes	telephone phone
8	Yes	phone telephone
9	Yes	phone number
10	Yes	phones

Table 57: Results of image retrieval using CSMI for query = "machine"

Query: machine ET: machines computer process system produce		
Image #	Related image?	Contains ET
1	Yes	machine machines
2	Yes	machine machines
3	Yes	machine machines
4	No	computer machines
5	Yes	system machine
6	Yes	computer system
7	No	system
8	No	system
9	Yes	computer machine
10	Yes	computer machine

Table 58: Results of image retrieval using CSMI for query = "computer"

Query: computer ET: computers system systems program design		
Image #	Related image?	Contains ET
1	Yes	computer program
2	Yes	computers computer
3	Yes	computers computer
4	Yes	computers computer
5	Yes	computers computer
6	Yes	computer computers
7	Yes	computers computer
8	No	computers computer
9	No	design computer
10	No	program

Table 59: Results of image retrieval using CSMI for query = "space"

Query: space ET: mathematics spaces logic called topological		
Image #	Related image?	Contains ET
1	No	spaces
2	No	space mathematics
3	No	space mathematics
4	No	called
5	No	spaces
6	No	space spaces
7	No	mathematics
8	No	mathematics
9	No	mathematics
10	No	called

Table 60: Results of image retrieval using CSMI for query = "game"

Query: game ET: player games players video board

Image #	Related image?	Contains ET
1	Yes	board games game
2	Yes	game board
3	Yes	game games player players
4	Yes	game games
5	Yes	video games
6	Yes	games players
7	Yes	games players
8	Yes	game board
9	Yes	game games
10	Yes	games game

Table 61: Results of image retrieval using CSMI for query = "power"

Query: power ET: system electric electrical refer ratio		
Image #	Related image?	Contains ET
1	Yes	power electrical
2	Yes	electrical power
3	Yes	electric electrical power
4	No	power electrical
5	Yes	electric electrical power
6	No	power electrical
7	Yes	power electrical
8	Yes	electric power
9	Yes	power electric
10	Yes	power electrical electric

Table 62: Results of image retrieval using CSMI for query = "piano"

Query: piano ET: music instrument pianos perform played		
Image #	Related image?	Contains ET
1	Yes	piano instrument music
2	Yes	piano music instrument
3	No	piano instrument music
4	Yes	piano music instrument

5	No	piano pianos music
6	Yes	piano instrument
7	Yes	piano instrument
8	Yes	perform piano
9	No	music instrument
10	No	music instrument

Table 63: Results of image retrieval using CSMI for query = "moon"

Query: moon ET: american america music written state		
Image #	Related image?	Contains ET
1	No	america american
2	Yes	moon state
3	No	written
4	No	written
5	No	written
6	No	written
7	No	america american
8	No	american state
9	No	american music
10	No	moon music

Table 64: Results of image retrieval using CSMI for query = "screen"

Query: screen ET: screens include product system design		
Image #	Related image?	Contains ET
1	Yes	screen screens
2	No	design product
3	No	design system
4	No	design product
5	Yes	product screen
6	No	screen product
7	No	design screen
8	No	design screen product
9	Yes	screens screen
10	No	system

Table 65: Results of image retrieval using CSMI for query = "car"

Query: car ET: vehicle america mobile automobile vehicles		
Image #	Related image?	Contains ET
1	Yes	vehicle vehicles automobile car
2	Yes	automobile vehicle car
3	Yes	automobile vehicle car
4	Yes	car automobile vehicle
5	Yes	car automobile vehicle
6	Yes	car vehicle vehicles
7	Yes	automobile car vehicle
8	Yes	automobile vehicle car
9	Yes	automobile vehicle car
10	Yes	automobile vehicle car

Table 66: Results of image retrieval using CSMI for query = "bike"

Query: bike ET: bicycle cycle bikes cycling event		
Image #	Related image?	Contains ET
1	Yes	bike bicycle bikes cycling cycle
2	Yes	bike bikes bicycle cycle cycling
3	Yes	bike bikes bicycle cycling cycle
4	Yes	bike bikes bicycle cycling cycle
5	Yes	bike bikes bicycle cycling cycle
6	Yes	bike cycle cycling bikes bicycle
7	Yes	bike bicycle cycle cycling
8	Yes	bike bicycle cycle cycling
9	Yes	bikes cycle bike bicycle
10	Yes	cycling bikes bicycle bike

Table 67: Results of image retrieval using CSMI for query = "soldier"

Query: soldier ET: world release soldiers state released		
Image #	Related image?	Contains ET

1	Yes	soldier soldiers
2	Yes	soldiers soldier
3	Yes	soldier soldiers
4	Yes	soldiers soldier
5	Yes	soldiers soldier
6	Yes	soldier soldiers
7	Yes	soldier soldiers
8	Yes	soldiers soldier
9	Yes	soldiers soldier
10	Yes	soldiers soldier

Table 68: Results of image retrieval using NT for query = "cloud"

Query: cloud ET: computing cirrus clouds point form		
Image #	Related image?	Contains ET
1	Yes	cloud cirrus
2	Yes	cloud clouds cirrus
3	Yes	cloud cirrus
4	Yes	cirrus clouds
5	Yes	clouds cloud cirrus
6	Yes	clouds cloud cirrus
7	Yes	cloud clouds cirrus
8	Yes	cirrus cloud
9	Yes	point clouds cloud
10	Yes	cirrus clouds

Table 69: Results of image retrieval using NT for query = "cat"

Query: cat ET: hanna joseph population size jerry		
Image #	Related image?	Contains ET
1	No	hanna
2	No	joseph
3	No	hanna
4	No	hanna
5	Yes	short cat
6	Yes	cat short
7	No	joseph short
8	Yes	cat jerry
9	No	hanna
10	No	hanna

Table 70: Results of image retrieval using NT for query = "weapon"

Query: weapon ET: lethal nuclear weapons bit explosive		
Image #	Related image?	Contains ET

1	Yes	lethal weapon
2	Yes	weapon weapons
3	Yes	weapon weapons
4	Yes	weapon weapons
5	Yes	weapon weapons
6	No	weapons nuclear
7	Yes	weapons weapon
8	Yes	weapon weapons
9	Yes	weapon weapons
10	Yes	weapon weapons

Table 71: Results of image retrieval using NT for query = "school"

Query: school ET: boarding college preparatory public schools		
Image #	Related image?	Contains ET
1	Yes	college public
2	No	boarding school
3	No	boarding public
4	Yes	college school public
5	Yes	college school schools
6	Yes	college public school
7	No	public schools school
8	Yes	boarding school
9	Yes	college school
10	No	college school

Table 72: Results of image retrieval using NT for query = "police"

Query: police ET: service force responsible chief constable		
Image #	Related image?	Contains ET
1	No	police chief service
2	No	police chief service
3	Yes	police constable
4	Yes	police constable
5	Yes	police constable

6	No	service force
7	No	service force
8	No	chief
9	Yes	police constable
10	Yes	constable police

Table 73: Results of image retrieval using NT for query = "fruit"

Query: fruit ET: dove ptilinopus bizarre tree iucn		
Image #	Related image?	Contains ET
1	Yes	fruit tree bizarre
2	Yes	fruit tree bizarre
3	No	ptilinopus dove
4	No	tree dove
5	Yes	tree red fruit
6	Yes	tree red fruit
7	Yes	fruit red bizarre
8	No	dove
9	No	tree dove
10	Yes	red iucn

Table 74: Results of image retrieval using NT for query = "meat"

Query: meat ET: puppets products pie pies packing		
Image #	Related image?	Contains ET
1	Yes	meat pie
2	No	pies pie
3	No	pie pies
4	No	pie pies
5	No	pie pies
6	No	meat products
7	No	pies
8	No	pies pie
9	No	pie pies
10	No	pies pie

Table 75: Results of image retrieval using NT for query = "telephone"

Query: telephone ET: numbering plan area codes local		
Image #	Related image?	Contains ET
1	No	local area
2	No	plan
3	No	plan
4	No	numbering
5	Yes	telephone
6	No	telephone
7	No	plan
8	No	plan
9	No	numbering
10	No	numbering

Table 76: Results of image retrieval using NT for query = "machine"

Query: machine ET: tools iron man pistols darwin		
Image #	Related image?	Contains ET
1	Yes	man machine iron
2	Yes	iron machine
3	No	machine man
4	No	darwin man
5	No	iron tools
6	No	tools iron
7	No	tools iron
8	No	iron machine
9	No	man darwin
10	No	darwin man

Table 77: Results of image retrieval using NT for query = "computer"

Query: computer ET: desktop software systems vision programs		
Image #	Related image?	Contains ET

1	No	computer software
2	Yes	desktop computer
3	Yes	desktop computer
4	Yes	desktop computer
5	No	computer software desktop
6	No	computer software desktop
7	Yes	desktop software
8	Yes	computer software
9	No	computer software
10	Yes	computer software

Table 78: Results of image retrieval using NT for query = "space"

Query: space ET: outer sample station personal phase		
Image #	Related image?	Contains ET
1	No	space personal
2	Yes	space outer
3	No	personal
4	No	sample
5	No	outer space
6	No	phase
7	No	personal space
8	No	personal space
9	Yes	space station
10	Yes	space station

Table 79: Results of image retrieval using NT for query = "game"

Query: game ET: players games played complete genre		
Image #	Related image?	Contains ET
1	Yes	game games players
2	Yes	game games
3	Yes	games played
4	Yes	games players
5	Yes	games players

6	Yes	game complete
7	Yes	game players
8	Yes	game players
9	Yes	game players
10	Yes	game players

Table 80: Results of image retrieval using NT for query = "power"

Query: power ET: electric market supply rating vacuum		
Image #	Related image?	Contains ET
1	Yes	power supply
2	Yes	power supply
3	Yes	electric power
4	Yes	power electric
5	Yes	electric vacuum
6	Yes	electric power
7	Yes	electric power
8	No	electric power
9	Yes	power electric
10	Yes	power electric

Table 81: Results of image retrieval using NT for query = "piano"

Query: piano ET: roll grand minor op novelty		
Image #	Related image?	Contains ET
1	Yes	piano grand
2	Yes	piano
3	Yes	piano grand
4	No	novelty
5	No	novelty
6	No	minor
7	Yes	piano grand
8	No	op
9	No	grand roll
10	No	grand roll

Table 82: Results of image retrieval using NT for query = "moon"

Query: moon ET: occurs natural satellite full short		
Image #	Related image?	Contains ET
1	Yes	moon natural satellite
2	Yes	moon full
3	Yes	full moon
4	Yes	moon full
5	Yes	moon full
6	Yes	moon full
7	No	full natural
8	Yes	moon full
9	Yes	moon full
10	Yes	moon full

Table 83: Results of image retrieval using NT for query = "screen"

Query: screen ET: guild theater clear view fire		
Image #	Related image?	Contains ET
1	No	guild
2	No	clear view
3	No	clear
4	No	guild
5	Yes	theater screen
6	No	theater
7	No	guild
8	No	fire clear
9	No	view clear
10	Yes	fire screen

Table 84: Results of image retrieval using NT for query = "car"

Query: car ET: body style cars include drag		
Image #	Related image?	Contains ET

1	Yes	drag racing cars
2	Yes	drag racing cars
3	Yes	drag racing cars
4	Yes	drag racing
5	Yes	drag racing
6	Yes	drag racing
7	No	car cars racing
8	Yes	car cars racing
9	Yes	car drag racing
10	Yes	drag racing car cars

Table 85: Results of image retrieval using NT for query = "book"

Query: book ET: music lungs tokens print run		
Image #	Related image?	Contains ET
1	Yes	book music
2	Yes	book print music
3	No	lungs
4	No	music print
5	Yes	print book
6	Yes	print book
7	Yes	music book
8	Yes	music book
9	No	tokens
10	No	print

Table 86: Results of image retrieval using NT for query = "bike"

Query: bike ET: park path week paths bar		
Image #	Related image?	Contains ET
1	Yes	bike week
2	No	bike week
3	Yes	bikes bike
4	No	path paths
5	No	park paths path

6	Yes	bike bikes
7	Yes	park bike bikes
8	Yes	bike week
9	Yes	bike week
10	Yes	bikes paths

Table 87: Results of image retrieval using NT for query = "soldier"

Query: soldier ET: lone soldiers game released song		
Image #	Related image?	Contains ET
1	Yes	soldier soldiers game
2	Yes	soldier soldiers
3	Yes	soldiers soldier
4	Yes	soldier soldiers
5	Yes	soldiers soldier
6	Yes	soldiers soldier
7	Yes	soldier soldiers
8	No	song written
9	Yes	soldiers soldier
10	Yes	soldiers soldier