A Systematic Comparison of Semantic Models on Human Similarity Rating Data: The Effectiveness of Subspacing

Benjamin P. Stone (bpstone@psychology.adelaide.edu.au)
School of Psychology, Level 4, Hughes Building, The University of Adelaide
Adelaide, SA 5005 Australia

Simon J. Dennis (dennis.210@osu.edu)
Department of Psychology, Ohio State University, 1827 Neil Ave
Columbus, OH 43210 USA

Peter J. Kwantes (peter.kwantes@drdc-rddc.gc.ca)
DRDC Toronto, 1133 Sheppard Ave W
Toronto, ON, Canada M3M 3B9

Abstract

A critical issue in the development of statistical models of lexical semantics is the nature of the background corpus that is used to derive term representations. Often good performance can be achieved if background corpora are hand selected, but performance can drop precipitously when this is not the case. In this study, we investigated querying a larger textbase (Wikipedia) to create subcorpora for analysis with lexical semantics models (Zelikovitz & Kogan, 2006). Similarities generated from six models of lexical semantics were compared against human ratings of document similarity on two document sets - the Internet Movie Database set and the newswire set from Lee, Pincomb and Welsh (2005). These methods included the vector space model (Salton, Wong & Yang, 1975), Latent Semantic Analysis (LSA, Kintsch, McNamara, Dennis, & Landauer, 2006), sparse Independent Components Analysis (Bronstein, Bronstein, Zibulevsky, & Zeevi, 2005), the topics model (Griffiths & Steyvers, 2002, Blei, Ng, & Jordan, 2003), nonnegative matrix factorization (Xu, Liu, & Gong, 2003) and the constructed semantics model (Kwantes, 2005). We found that on these datasets the creation of subcorpora can be very effective even improving upon the performance of corpora that had been hand selected for the domain. Surprisingly, the overall best performance was observed for the vectorspace model.

Keywords: Semantic models; human similarity judgment; corpora sub-spacing

Introduction

This paper describes the outcome of a systematic comparison of document similarities generated by a set of six statistical semantics models to document similarities generated by human participants. The purpose of the work was to determine whether subcorpora constructed by querying a larger corpus (Wikipedia) could produce sufficient performance to be generally useful (Zelikovitz & Kogan, 2006). In many applications the hand construction of corpora for a particular domain is not feasible, and so the ability to show a good match between human similarity evaluations and machine evaluations of document similarity using automated methods of corpus construction is critical. In the following sections, we will briefly describe each of the models and the data sets that we will be examining. Three studies are then described, the first two focus on corpora which were picked from specific knowledge domains, in the second this domain-specific corpus is also augmented by a more general purpose corpus (TASA).

In the final section, domain specific corpora are drawn from the Wikipedia database in research that explores the viability of using this large set of documents as a basis for automated corpus construction.

Models

The models examined were the vectorspace model (Salton, Wong & Yang, 1975), Latent Semantic Analysis (LSA, Kintsch, McNamara, Dennis, & Landauer, 2006), Sparse Independent Components Analysis (Bronstein, Bronstein, Zibulevsky, & Zeevi, 2005), the topics model (Griffiths & Steyvers, 2002, Blei, Ng, & Jordan, 2003), Sparse Nonnegative Matrix Factorization (Xu, Liu, & Gong, 2003) and the Constructed Semantics Model (Kwantes, 2005).

The vectorspace model (Salton, Wong & Yang, 1975): The vectorspace model assumes that terms can be represented by the set of documents in which they appear. Two terms will be similar to the extent that their document sets overlap. To construct a representation of a document, the vectors corresponding to the unique terms are multiplied by the log of the frequency within the document and divided by their entropy across documents and then added. Similarities are measured as the cosines between the resultant vectors for different documents.

Latent Semantic Analysis (LSA, Kintsch, McNamara, Dennis, & Landauer, 2006): LSA started with the same representation as the vectorspace model a term by document matrix with log entropy weighting. In order to reduce the contribution of noise to similarity ratings, however, the raw matrix is subjected to singular value decomposition (SVD). SVD decomposes the original matrix into a term by factor matrix, a diagonal matrix of singular values and a factor by document matrix. Typically, only a small number of factors (e.g., 300) are retained. To derive a vector representation of a novel document, term vectors are weighted, multiplied by the square root of the singular value vector and then added. As with the vector space model, the cosine is used to determine similarity.

Sparse Independent Components Analysis (Bronstein, Bronstein, Zibulevsky, & Zeevi, 2005): Sparse ICA is a

derivative of Independent Components Analysis that takes advantage of the fact that in sparse representations it can be assumed that the appearance of a particular value occurs as a consequence of one independent component and one component only. Deriving the components then reduces to normalizing the term vectors from the raw term by document matrix and then clustering in high dimensions. While this technique has primarily been applied to image processing problems, we were interested in investigating whether it would transfer to document processing applications. Note a number of clustering mechanisms can be employed. We investigated K-Means, Fuzzy C Means and Instant Runoff clustering. In pilot testing, only the K-Means version proved to be computationally feasible.

The topics model (Griffiths & Steyvers, 2002, Blei, Ng, & Jordan, 2003): The topics model is a Bayesian approach to document similarity that assumes a generative model in which a document is represented as a multinomial distribution of topics and topics are represented as multinomial distributions of words. The parameters of these models can be inferred from a corpus using either Markov Chain Monte Carlo techniques or variational Expectation Maximization. We implemented the former. Theoretically, document representations should then be calculated by running the sampler over a new corpus augmented with information from the new document. Even if one assumes that the background corpus assignments may be fixed such a procedure is prohibitively computationally expensive. Consequently, we choose to average the word representations to calculate topic distributions and then employed the dot product and Jensen Shannon measures of similarity.

Sparse Nonnegative Matrix Factorization (SpNMF, Xu, Liu, & Gong, 2003): Nonnegative Matrix Factorization is a technique similar to LSA, which in this context creates a matrix factorization of the weighted term by document matrix. This factorization involves just two matrices a term by factor matrix and a factor by term matrix - and is constrained to contain only nonnegative values. While nonnegative matrix factorization has been shown to create meaningful word representations using small document sets, in order to make it possible to apply it to large collections we implemented the sparse tensor method proposed by Shashua and Hazan (2005). As in LSA, log entropy weight term vectors were added to generate novel document vectors and the cosine was used as a measure of similarity.

The Constructed Semantics Model (Kwantes, 2005): The final model considered was the constructed semantics model (Kwantes, 2005). The CSM is unique in that it was created primarily as a cognitive model to explain the emergence of semantics from experience. It operates by taking the term by document matrix (using just log weighting) and multiplying it by its transpose. Consequently, terms do not have to appear together in order to be similar as is the case in the vectorspace model. Again terms are added to create novel document vectors and the cosine is used as a measure of similarity.

The Data

We used two datasets in this study – the IMDB set and the Lee set. The IMDB corpus is a large collection of "celebrity gossip" articles collected from the Internet Movie Database website between April 2000 and January 2006. The DRDC supplied human similarity rating of 25 of these documents for analysis in this study. The second dataset used was collected by Lee, Pincomb and Welsh (2004). It consists of ten independent ratings of the similarity of every pair of 50 short documents selected from the Australian Broadcasting Corporations news mail service, which provides text e-mails of headline stories. Wikipedia was utilized as a generic corpus from which smaller targeted sub-spaces could be sampled and compiled. To this end, the entire collection of Wikipedia entries were collected, and is current to March 2007. In total there were 2.8 million Wikipedia entries collected, however the document number was reduced to 1.57 million after the removal of incomplete articles contained in the original corpus. The incomplete articles removed were selected if they contained the word "Wikipedia" or "incomplete stub".

Methods and Results

The following section is delineated into three subsections. The first two subsections describe results where domain specific corpora have been collected. That is, each corpora (Toronto Star & IMDB-database) were thought to adequately represent the knowledge domains accessed by humans when comparing documents in the Lee and IMDB dataset, respectively. Furthermore, in Section Two, the IMDB corpora has also been augmented with the well known TASA general knowledge corpus. The third subsection examines the effectiveness of generating sub-corpora from Wikipedia for analysis with lexical semantics models on the Lee and IMDB datasets.

Initially, all corpora were preprocessed using standard methods; characters converted to lower case, numbers were zeroed (i.e., 31 Jan 2007 became 00 jan 0000), punctuation and words from a standard stoplist were removed, and words that appear only once in the corpus or in only one document were also removed. However, in the analysis described in Section Three (Part Two) below, further "stringent cleaning" of the Wikipedia corpus was performed to entirely remove numbers and single letters (such that 'J' and 'K' would be removed from 'J K Rowling' to create 'Rowling').

Section One: LEE-dataset & Toronto Star Corpora

For each of the semantics models, one or more spaces were compiled and similarity measures for each pair of documents from the Lee et al. (2005) collection were calculated. For the vectorspace and CSM models, no dimension reduction is attempted and so we compiled one space for each method. For the remaining models, one must select a number of dimensions in which to calculate similarities. Performance is likely to be influenced by this choice and so in each case we compared 50, 100 and 300 dimensional models. For the vec-

Table 1: Toronto Star, IMDB-TASA, TASA, IMDB, LEE-10000, IMDB-10000, LEE-1000 and IMDB-1000 corpus details.

	Number of Documents	Total words	Ave. Word/Doc	Unique Terms	Ave. Unique/Doc
Toronto Star	55021	15449673	280.80	97511	1.77
IMDB-TASA	48258	6195160	128.38	59054	1.22
TASA	35471	5169668	145.74	54531	1.54
IMDB	12787	1025492	80.20	22983	1.80
LEE-10000*	10000	2646818	264.68	69415	6.94
IMDB-10000*	10000	2715588	271.60	73511	7.35
LEE-1000*	1000	273720	273.72	13677	13.70
IMDB-1000*	1000	262054	262.21	18154	18.15

* Corpora "Stringently Cleaned" - see Section 3 (Part Two)

torspace, LSA, sparse ICA and sparse NMF models log entropy weighting was employed. For CSM, log weighting was used. A three day space compilation limit was also imposed. Spaces that had not compiled in this time were abandoned as they are likely to be impractical for use in applied domains.

The first result of interest identifies which of these models completed within the three day time limit that was imposed. Regardless of the advantageous properties of an algorithm if it does not compile in a reasonable period of time it will not be appropriate for most applications scaling is often a determining factor. Sparse NMF (300 dimensions), Sparse ICA (50, 100, and 300 dimensions), and CSM all failed to compile the Toronto Star corpus.

The correlations between similarity ratings made by humans and the models in this study were very low (Table 2). The topics model (300 topics) with the Jensen Shannon metric and Sparse NMF were the best performing models, correlating 0.13 with human judgment. However, Sparse NMF out-performed Topics-JS at lower dimensionality, so if it had not timed-out in the 300 dimension space compilation, it may have been the best performing model. Vectorspace had higher correlations than both LSA and non-Jensen Shannon Topics model. In nearly all cases, as model dimensionality became more complex, correlations with human ratings increased.

Lee's (2005) results indicate that LSA was the best performing model in their study, correlating about 0.6 with human judgments. While this was not the case in this study, poor performance may have been driven by a less than optimal match between the 'background' corpus and the documents rated. The likelihood of this scenario is supported by the generally low correlations with human results obtained by all of the models.

Section Two: IMDB-dataset & the IMDB and TASA corpora

As can be seen in Table 1, by augmenting the IMDB corpus with the TASA corpus, the resulting IMDB-TASA has more documents, longer documents and more unique terms. However, the average number of unique terms per document in the IMDB-TASA is less than both the IMDB and the TASA corpora.

Table 2: Correlations (r) between LEE-dataset similarity ratings made by human raters and the those made by LSA, Topics, Topics-JS (with Jensen- Shannon), Sparse ICA, Sparse NMF, Vectorspace, and CSM.

	Dimensions			
Model	None	50	100	300
LSA	-	0.04	0.05	0.06
Topics	-	0.02	0.01	0.06
Topics-JS	-	0.10	0.11	0.13
SpNMF	-	0.09	0.13	*
SICA	-	*	*	*
Vectorspace	0.10	-	-	-
CSM	*	-	-	-

*Space compilation exceeded 3 days
Note: Correlations did not include Same–Same documents.

Sparse ICA (with KMeans) failed to compile the IMDB-TASA corpus within three days, therefore it has not been included in this analysis.

Overall, LSA compiled at 300 dimension from the IMDB-TASA corpus was the model with the best match (0.23) to the human judgments in this study (Table 3). Furthermore, augmenting the IMDB with the TASA corpus increased the performance of LSA in all cases (50, 100, 300 dimensions). SpNMF and vectorspace models were also made more effective predictors of human judgments by augmented IMDB-TASA corpus. Conversely, Topics-JS (with Jensen Shannon equation) the second best performing model (0.20) using the IMDB-TASA, performed worse with only the IMDB (0.10). This trend was also displayed by Topics and CSM (Table 3).

Similar to the findings outlined in Section One, in this study increases to space dimensionality corresponded to increased correlations between model similarity assessments and judgments made by humans. It appears that the IMDB corpus may be a better match to the documents rated by humans in this study than the Toronto Star corpus used in the previous section. However, this would be expected given that the 25 documents rated by humans in this study were a subset of the IMDB corpus. In future research, it may prove inter-

esting to remove the documents rated by humans from the background corpus.

The results make it hard to discern if the augmentation of the IMDB corpus using TASA improved the performance of the models. Extending our background corpus produced performance improvements in latent semantic analysis, vectorspace model and sparse nonnegative matrix factorization, but decrements in both topics models and constructed semantics model.

Table 3: Similarity correlations (r) are between each semantic model and human judgments recorded in Kwantes study. Corpora augmentation: IMDB (I) compared to IMDB augmented with TASA corpus (IT).

	Dimensions			
Model	None	50	100	300
LSA (I)	-	0.21	0.20	0.19
LSA (IT)	-	0.22	0.22	0.23
Topics (I)	-	0.03	0.10	0.12
Topics (IT)	-	0.03	0.07	0.10
Topics-JS (I)	-	0.20	0.12	0.12
Topics-JS (IT)	-	0.10	0.11	0.12
SpNMF (I)	-	0.10	0.10	0.09
SpNMF (IT)	-	0.14	0.15	0.14
Vectorspace (I)	0.17	-	-	-
Vectorspace (IT)	0.19	-	-	-
CSM (I)	0.17	-	-	-
CSM (IT)	0.10	-	-	-

(I) - IMDB corpus
(IT) - IMDB corpus augmented with TASA corpus
Note: Correlations did not include Same–Same documents.

Section Three: Wikipedia Sub-corpora

To enable the creation of sub-space corpora, Lucene (a high performance text search engine) was used to index each document in the "cleaned" Wikipedia corpora. Lucene allows the user to retrieve documents based on customized queries. Like the more well known search engine Google, the documents returned by Lucene are ordered by relevance to the query.

Targeted queries were formulated for each group of documents that were rated by humans in the IMDB and Lee, Pincombe, and Welsh (2005) studies. The IMDB-based query was constructed by removing the stop-words and punctuation from each article's heading that humans had rated, and then joining the remaining words with "OR" statements. In contrast, the query devised for LEE-based data was more complex, in that it combined keywords chosen by the researcher with "AND" and 'OR" operators. Moreover, the LEE-based query used Lucene's the 'star' wildcard operator to return multiple results from word stems, for example "research*" would return documents containing the words "research", "researcher", and "researchers".

The following analysis has been delineated into two main parts. Part One deals with the initial investigation into subspacing using the LSA model and "cleaned" Wikipedia. For this initial exploratory analysis, LSA's was chosen for its quick compilation speeds and because of the generally good match that has been reported between LSA and human performance (Lee, Pincombe, & Welsh, 2005). This analysis indicated that both document size and the information position within a document (Lee & Corlett, 2003) affected the models ability to match to human similarity judgments. Part Two describes research using the Wikipedia corpora that had been more "stringently cleaned" (as described above). Removing numbers and single letters from these corpora substantially improve the performance of all the models (LSA, Topics, SpNMF, CSM, and Vectorspace) when comparing automated assessments of similarity with the human judgments recorded for the both IMDB and LEE documents.

Table 4: Correlations (r) between human and LSA (300 dimensions) on documents in the IMDB and LEE datasets. The ALL-300 corpus refers to the first 300 documents return by querying Wikipedia sorted in order of Lucene assess relevance. ALL- 500 is the first 500 documents, ALL-40000 the first 40000 documents.

Subspacing Type	LEE	IMDB
ALL-300	0.13	0.03
ALL-500	0.11	0.15
ALL-1000	0.09	0.15
ALL-10000	0.05	0.16
ALL-40000	0.04	0.17
FIRST-300	0.14	0.11
FIRST-500	0.14	0.09
FIRST-1000	0.15	0.11
FIRST-10000	0.16	0.10
FIRST-40000	0.14	0.15
SEG-300	0.13	0.11
SEG-500	0.10	0.10
SEG-1000	0.12	0.13
SEG-10000	0.09	0.17
SEG-40000	0.13	0.17
RANDOM-300	0.06	0.13
RANDOM-500	0.08	0.15
RANDOM-1000	0.06	0.15
RANDOM-10000	0.05	0.15
RANDOM-40000	0.04	0.15

Note: Correlations did not include Same–Same documents.

Part One: Sub-spacing using standard "cleansing" procedures. Four types of corpora were extracted from the Wikipedia document set. Three of the four types of corpora were based on the results returned by the Lucene queries described above. The documents return from these queries where either: accepted in full (ALL); truncated after the first 300 words (FIRST); or, generated from a Wikipedia corpus in which every document had been segmented in block of 300

words (SEG). The fourth type of corpora was generated from a set of randomly sampled documents from Wikipedia (RANDOM). Copora size was also manipulated for all of these types of documents, ranging from the most relevant (as assessed by the Lucene algorithm) 300 to 40000 documents. LSA (300 dimension) sub-spaces where compiled for all corpora, which where then assessed for their ability to match human similarity ratings from the IMDB and LEE tasks. The results of these comparison between similarity assessments of humans and the LSA model are displayed in Table 4.

Table 5: Average correlations (r) between human and LSA (300 dimensions) on documents in the IMDB and LEE datasets.

	LEE	IMDB	Total Average
ALL	0.08	0.13	0.11
FIRST	0.14	0.11	0.13
SEG	0.11	0.14	0.12
RANDOM	0.06	0.15	0.10

Note: Correlations did not include Same–Same documents.

Averaging over the figures in Table 4 across both the LEE and IMDB datasets indicates the best results were obtained using the FIRST 300 words (0.13) from each of the documents return by Lucene from the Wikipedia document set (Table 5). This finding is also supported by the work of Lee and Corlett (2003). In their study of Reuters documents, words with greater 'mean evidence' for a document were found at the beginning of documents (see Figure 1). It may be the case, that documents (webpages) in a web-based medium such as Wikipedia, may also follow this generalization. Intuitively, it seems likely that important descriptive information on a webpage will be positioned nearer the top of a page (within the first 300 words), so as not to be over-looked by the reader as the webpage scrolls or extends beneath screen view.

None of the possible sub-spacing methods discussed so far matched the human data significantly better that a set of documents drawn at random from the Wikipedia corpus (RANDOM). However, based on the trends displayed, the FIRST 300 words method was used when generating sub-spacing corpora for the rest of the analysis discussed in this report.

Part Two: A more "stringently cleaned" Wikipedia. Further examination of the Wikipedia corpora revealed many occurrences of both numbers (which had been zeroed from the initial cleaning process) and single letters. Removing the single letters reduced the Wikipedia corpus file size from 5.8Gb to 5.7Gb. The subsequent removal of the zeroed numbers reduced the Wikipedia corpus' size further to 5.5Gb. It is possible that the presence of these types of information creates noise for the corpora derived from Wikipedia. For example, the American declaration of independence in 1776 has little to do with Elvis Presley's birthday in 1935. Similarly, the 'Js' in 'J F K' and 'J K Rowling' do not indicate semantic similar-
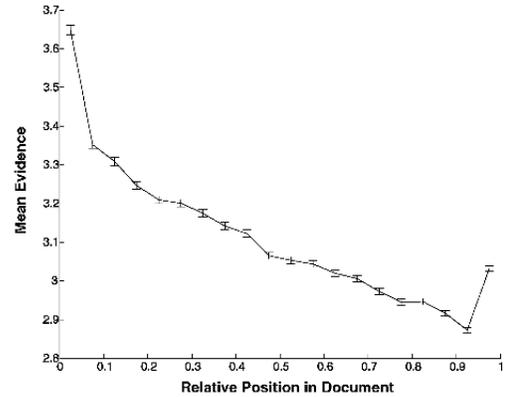

Figure 1: The mean absolute evidence provided by words in the Reuters-21578 corpus, as a function of there relative position in the document.

ity. Compared to the spaces created in Sections One and Two, spaces created using the "stringently cleaned" Wikipedia set of documents have a very high number of unique words per document. It is also clear that some documents are shorter and did not reach their 300 word cutoff point.

Model spaces generated from documents drawn using the "stringently cleaned" Wikipedia corpora generally outperformed those examined in Sections One and Two for every model (Table 6). Particularly encouraging was the performance of these models on the LEE dataset. In Section One the best performing models (Topics-JS & SpNMF) only correlated 0.13 with the human similarity ratings. In contrast, using 10000 "stringently cleaned" Wikipedia documents return from the LEE- based query, the vectorspace model correlates 0.51 with the human judgment. However, to put this figure into perspective, this correlation is less than the human inter-rater reliability (0.605) in Lee, Pincombe and Welsh's (2005) study.

Overall, the vectorspace model was the best performing model in comparison to the similarity judgments made by humans on both the IMDB (second best) and LEE datasets (best). SpNMF and Topics-JS (with Jensen Shannon equation) also perform comparatively well on both datasets. This is encouraging, given that Topics model has the benefit of constructing interpretable dimensions. That said, Topics-JS did not perform as well as LSA in Section Two when compared on either the IMDB or IMDB-TASA corpora. However, Topics-JS's poor performance on this task may have resulted from the generally small document sizes (average words per document 80.20) in the IMDB and (average words per document 128.38) IMDB-TASA corpora (see Table 1). In future research, it would be possible to explore this effect of document length by truncating the Wikipedia document lengths to 100 words on a IMDB-based query.

CSM correlations were higher on the IMDB-based Wikipedia subspace than on the LEE-based dataset. It is also worth noting that the spaces created by CSM are substan-

tially larger than those created by any other model. For the 10000 document IMDB-based Wikipedia subspace, the CSM space was 1.2Gb, much larger than the size of the vectorspace model space which was 215Mb. It was unfortunate given the relative solid perform of Sparse NMF in the 10000 document conditions on both IMDB and LEE datasets, that it failed to compile in the three days allocated for this purpose. Finally, increasing the number of documents from 1000 to 10000 generally increased the models correlations with the human data in this study, however this pattern was not observed for the topics model.

Table 6: Correlations (*r*) between similarity assessments made by semantic models and humans on both the IMDB and LEE datasets using the first 300 words from 1000 and 10000 stringently cleaned Wikipedia documents. All model spaces except Vectorspace and CSM were created with 300 dimensions.

	LEE		IMDB	
Model	1000	10000	1000	10000
Vectorspace	0.40	0.51	0.14	0.29
LSA	0.33	0.36	0.13	0.20
Topics	0.43	0.38	0.05	0.24
Topics-JS	0.33	0.38	0.20	0.26
CSM	0.17	0.19	0.18	0.27
SpNMF	0.33	0.36	0.26	0.37

Note: Correlations did not include Same–Same documents.

Conclusion

In this study, we investigated querying a larger textbase (Wikipedia) to create subcorpora for analysis with lexical semantics models (Zelikovitz & Kogan, 2006). Similarities generated from six models of lexical semantics were compared against human ratings of document similarity on two document sets - the Internet Movie Database set and the newswire set from Lee, Pincomb and Welsh (2005). These methods included the vector space model (Salton, Wong & Yang, 1975), Latent Semant Analysis (LSA, Kintsch, McNamara, Dennis, & Landauer, 2006), sparse Independent Components Analysis (Bronstein, Bronstein, Zibulevsky, & Zeevi, 2005), the topics model (Griffiths & Steyvers, 2002, Blei, Ng, & Jordan, 2003), nonnegative matrix factorization (Xu, Liu, & Gong, 2003) and the constructed semantics model (Kwantes, 2005).

We found that on these datasets the creation of Wikipedia-based subcorpora can be effective even improving upon the performance of corpora that had been hand selected for the domain. This performance increase may be caused by the greater average number of unique terms found in each document of the Wikipedia subcorpora.

The performance of all models improved when using subcorpora drawn from Wikipedia. However, this improvement was in part driven by a more "stringent" preprocessing of cor-

pus documents. Future research will explore the removal of both numbers and single letters from the Toronto Star corpora and IMDB to investigate whether this improves the performance of semantic models using these more "stringently cleaned" corpora.

Overall, the best performance was observed for the vectorspace model. This was surprising given that, compared to the other models described in this research, vectorspace is both computationally less intensive and easy to implement.

References

Blei, D., Ng, A. Y., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bronstein, A. M., Bronstein, M. M., Zibulevsky, M., & Zeevi, Y. Y. (2005). Sparse ica for blind separation of transmitted and reflected images. *International Journal of Imaging Science and Technology*, *15*, 84–91.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In *Proceedings of the 24th annual conference of the cognitive society* (pp. 381–386). Mahwah, NJ: Lawrence Erlbaum Associates.

Kintsch, W., McNamara, D., Dennis, S., & Landauer, T. (2006). *Handbook of latent semantic analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kwantes, P. J. (2005). Using context to build semantics. *Psychonomic Bulletin and Review*, *12*, 703–710.

Lee, M. D., & Corlett, E. Y. (2003). Sequential sampling models of human text classification. *Cognitive Science*, *27*, 159–193.

Lee, M. D., Pincombe, B. M., & Welsh, M. B. (2005). An empirical evaluation on models of text document similarity. In *Proceedings of the 27th annual conference of the cognitive society* (pp. 1254–1259). Mahwah, NJ: Lawrence Erlbaum Associates.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*, 613–620.

Shashua, A., & Hazan, T. (2005). Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22nd international conference on machine learning* (pp. 792–799). New York, NY: ACM Press.

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international acm sigir conference on research and development in informaion retrieval* (pp. 267–273). New York, NY: ACM Press.

Zelikovitz, S., & Kogan, M. (2006). Using web searches on important words to create background sets for lsi classification. In *Proceedings of the 19th international flairs conference* (pp. 598–603). Menlo Park, CA: AAAI Press.