FACULTY OF HUMANITIES
UNIVERSITY OF COPENHAGEN

Cover Page for Take-Home Assignments / Synopses

For Assignments at the Faculty of Humanities

Personal information	
Name	Timothy Powell
Student email *(e.g. abc123@alumni.ku.dk)*	tfv338@alumni.ku.dk
Phone number	45 91 86 96 01

Department

- ❍ ENGEROM
- ❍ IKK
- ◓ INSS
- ❍ IVA
- ❍ MEF
- ❍ SAXO
- ❍ ToRS

Study information	
Study programme	IT and cognition
Level	❍ BA ❍ BA elective ❍ KA ◓ MA elective ❍ Open University

Information about the exam					
Title of course	Project				
Title of exam	Project				
Subject-element code					
Curriculum					
Examination period	Winter				
Assignment title	Comparing Document Similarity Measures for Examining Media Summa				
Standard pages		No. of characters	16543	May the assignment be incl. in the assignment library?	◉ Yes ❍ No
Examiner					

Information in case of group examinations				
The author of each section in the paper is clearly stated (tick off)	◓ Yes	❍ No		
Co-writer 1	Name	Timothy Powell	KU username	tfv338
Co-writer 2	Name		KU username	
Co-writer 3	Name		KU username	
Co-writer 4	Name		KU username	

Declaration of Academic Honesty

I hereby confirm that I have completed this assignment on my own. In case of group examination, is the author of each section in the paper clearly stated in accordance with rules on group examinations. All quotations in the text have been marked as such and the assignment or fundamental parts of it have not been presented before in other contexts of assessment. The maximum number of standard pages has not been exceeded. *Handing in the assignment electronically by logging in at Absalon and uploading the material will replace a handwritten signature.*

You can find instructions on how to merge two PDFs on http://absalon.hum.ku.dk/english/students/onlinesubmission/

Rev. 23-April-2014

Comparing Document Similarity Measures for Examining Media Summaries of Large Documents

Timothy Powell

Abstract

Document similarity measures are used in various Information Retrieval tasks or as evaluation measures for automatic text summarization. In this paper I examine document similarity measures involving latent topic modelling through LDA for use with media articles that provide summaries of the report of the Truth and Reconciliation Commission of Canada. I compare topic vector representations made up of documents' topic probability distributions, with topic vectors in which the probabilities are weighted by each topic's coherence score. Evaluation of either approach is done by calculating the correlation between human ratings of document similarity over news data and cosine distances provided by either measure. My findings indicated a poor correlation for both methods, with weighted vectors displaying a lower correlation.

1 Introduction

Document similarity measures provide information for tasks such as document clustering, and the retrieval and ranking of documents resulting from a query in a search engine. Automatic text summarization involves the reduction of longer texts into shorter documents that convey important information from the source text. Evaluation of automatic summarization is often done with techniques that rely on ideal human-made text summaries, such as ROUGE, while approaches that rely solely on document similarity through various NLP methods have been attempted [2]. Automatic summarization is commonly applied to news articles in order to produce abstract summaries of one or several news stories. Rather than evaluate the quality of automatically generated news summaries, I'm interested in measuring the similarity between larger documents and news stories written with the partial intention of providing a summary account of a larger source text. Advancing methods that evaluate the quality of this class of news stories is of special importance when considering a source document of public interest whose size betrays its democratic value. Whether information leaks of classified documents, or

government commissioned reports, a journalist's duty is to relay this information to the public in a concise and representative manner. My intention is to explore NLP document similarity techniques for evaluating and visualizing how closely media articles characterize such documents. I have focused on document comparisons based on probability distributions generated by Latent Dirichlet Allocation (LDA). Comparing topic vectors produced by a topic model is a common approach to document similarity tasks in NLP and Information Retrieval, in addition to other methods such as measuring the cosine similarity of tf-idf vectors.

I am comparing 60 media articles related to the summary report of The Truth and Reconciliation Commission of Canada. The TRC was organized in order to acknowledge and record the experiences of Canada's First Nations people under the Residential School System and provide recommendations for reconciliation. The final report of the commission totals over 2 million words in 6 volumes while the summary report counts over 400 pages. The purpose of this project is to explore different approaches to measuring the similarity between the commission's report and online writing following its release. The current analysis uses LDA to generate a topic model over the the data, where the number of topics has been optimized by comparing the average topic coherence score across models. The topic coherence scores are used again as weights for the cosine distance metric between the topic distribution vector of the report and vectors of articles.

The second half of this paper describes an experiment conducted to evaluate the performance of either topic vector representations (weighted vs unweighted). The cosine distance between topic vectors of pairs of evaluation documents is compared to participant-annotated ratings of these document pairs' similarities.

2 Background

2.1 Document Similarity

Measuring the similarity between two documents involves transforming the plain text document into a representation corresponding to the input the document as a vector of features. A high-dimensional, sparse vector corresponding to the vocabulary in the document set can represent a document by containing counts of the words observed in the document, or with a boolean model (1 if the word is observed, 0 if not). The values of a count vector can be weighted with a tf-idf measure (term frequency-inverse document frequency) to capture the importance of the word within the document, relative to the corpus. Hazen[3] has shown that these

approaches, whereby a document is directly modelled with its observed features, are outperformed by document similarity approaches that make use of the latent modelling of documents, in which latent topics are inferred from the observed data.

A common way to measure the similarity between documents in vector space (represented by latent models such as LDA or direct models such as tf-idf) is to get the cosine of the angle between the vectors x and y, and the dissimilarity or distance between the vectors can be calculated by inverting the cosine similarity function:

$$cosdist(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||}$$

2.2 Latent Dirichlet Allocation

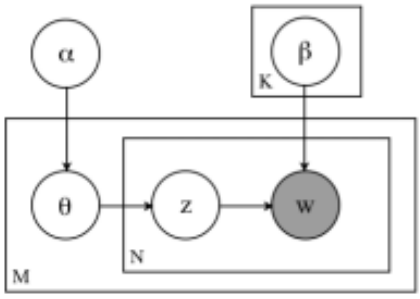

Figure 1: Plate notation of LDA

LDA is a generative probabilistic model that can be illustrated by the plate notation in figure 1. For every word w in a document M_i of size (1...N), the word and topic z to which it belongs is generated by drawing on distributions ϕ and θ, corresponding to Dirichlet distributions of words in a topic (out of K topics) and the mixture of topics in a document, respectively. α and β are concentration parameters of the per-document topic distribution and the per-topic word distribution. But these distributions have to come from somewhere. LDA involves an iterative algorithm, whereby parameters are initialized and topics and word distributions are chosen randomly. The following step iterates over all the words in all the documents, reassigning a new topic to each word based on the assumption that the topic assignment on all other words is correct. At one point, after enough iterations, the changes in reassignment of topics will converge.

A few things are worth noting about this process. Any document can be seen as a mixture of multiple topics, albeit with varying probabilities. Also, LDA does not take into consideration any structure or sequence of words. Each document is treated as a bag-of-words. It is unsupervised; no training documents are given, and all data is unlabeled, so we are effectively finding 'hidden' topics or themes in our data. The only observed random variables are the words. The topics, or rather topic-document and word-topic distributions, are latent variables whose statistical inference is the central task in LDA.

LDA is seen as an improvement on other probabilistic text modelling methods, such as pLSI and the mixture of unigram model, as the latter allows only one topic per document and the former requires a set of training documents [4]. Latent Semantic Indexing (LSI) is the non-probabilistic precursor to pLSI and was limited to performing a Singular Value Decomposition on a term-document matrix.

2.3 Topic coherence

Although topics are generated based on frequency of word co-occurrence, the coherence of sets of words on a semantic level may be imperfect. Two unrelated sets of words may combine to form one topic, or pairs of related words might all combine in a chain to form a topic of several unrelated pairs. A topic's coherence would thus be a measure to evaluate the quality of a topic by how related all the words in the set are. Mimno et al. [7] define a topic's coherence by calculating the probability that the words in the set co-occur in the documents of the corpus. Where $D(w_i)$ is defined as the count of documents containing word w_i, coherence is calculated by summing over all the scores of pairs in the topic where w_i is more common than w_j.

$$\text{Coherence} = \sum_{i<j} \text{score}(w_i, w_j)$$

This score is defined as:

$$\text{score}_{\text{UMass}}(w_i, w_j) = \log \frac{D(w_i, w_j) + 1}{D(w_i)}$$

This method is considered an improvement over pre-existing evaluation methods for topic models, where the quality of a model is measured by the likelihood of

unseen documents (perplexity). It also outperforms existing word-intrusion tests for detecting lower quality topics[7]. For these reasons, I am relying on this coherence score to compare consecutive runs of LDA with increasing number of topics. I am also relying on this measure to produce weighted vectors corresponding to the distribution of topics in my data, emphasizing the quality of the topics.

3 Approach

I am using the Mallet implementation of LDA[6], which uses Gibbs sampling to infer the posterior distribution of latent variables θ and ϕ, and performs hyperparameter optimization to determine α and β. I am running LDA on the entirety of my data to get my model, then inferring the topic distribution on either subset of the data required for document comparisons. Mallet provides a diagnostics file, which includes coherence scores on each topic.

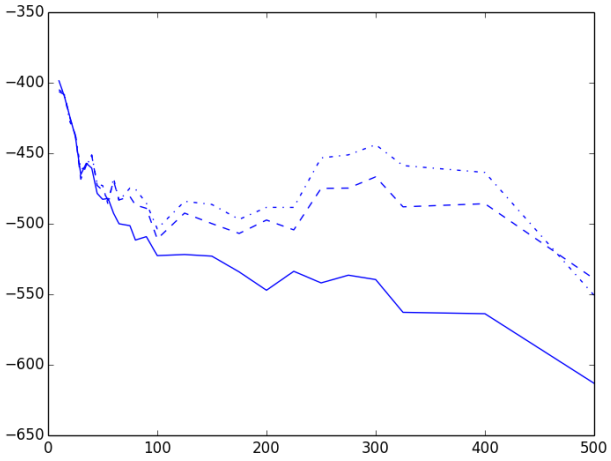

Figure 2: Mean, median and trimmed mean coherence scores for K=10...500

Figure 2 shows me average coherence over different topic numbers from 10 to 500. The solid line represents the average of coherence scores returned over all topics of each K=(10...500). The size and variety of the text in my corpus suggests a thematic distribution above 10 topics, yet this K returns the best average coherence score. I then plotted the median coherence score rather than only the average, represented by the line of dashes, in order to detect whether the higher granularity

Table 1: Topic terms for top 3 topics in TRC report, article with closest cosine distance, and article with furthest cosine distance

Document:	Report	Most similar article	Most dissimilar article
Topic rank 1	”thought” ”educate” ”continuing”	”indigenous” ”truth” ”reconciliation”	”world” ”chinese” ”war”
Topic rank 2	”projects” ”art” ”healing” ”community”	”culture” ”churches” ”make”	”peoples” ”people” ”european”
Topic rank 3	”treaties” ”land” ”indian”	”policy” ”cultural” ”government”	”report” ”recommendations” ”trc”

provided by an increase in topics also produces a handful of topics with very low coherence scores. This seems to be the case given the tendency for the distance between median and average score to grow as K increases. As a result, I plotted the average coherence score with outliers filtered out, based on absolute distance to the median. This allows me to choose 300 as the optimal K, as it is the highest score following a local maxima (but it's lower that K=10).

A cursory examination of the topics inferred on the TRC report and media articles reveal the top 3 topics of the report and its most similar and dissimilar articles (by cosine distance of unweighted topic vectors) containing terms displayed in Table 1.

The intuition behind my approach is that a vector representation of a document as a mixture of topics is more representative of its semantic content if the quality of each topic is factored in. Consequently, the similarity of these documents is captured better by the comparison of these 'informed' topic vectors. The following sections involve the evaluation of this approach. The similarity of the TRC's report to 60 related media articles is illustrated using 'buddy plots' as introduced in [1], where position encodings demonstrate changes in document similarity when the model or metric used is changed.

Figure 3 gives the distances of different media articles to the full report. Cosine distances of topic vectors (article vs full report) are represented on the top row, while the weighted topic vectors (with topic coherence acting as weights) are com-

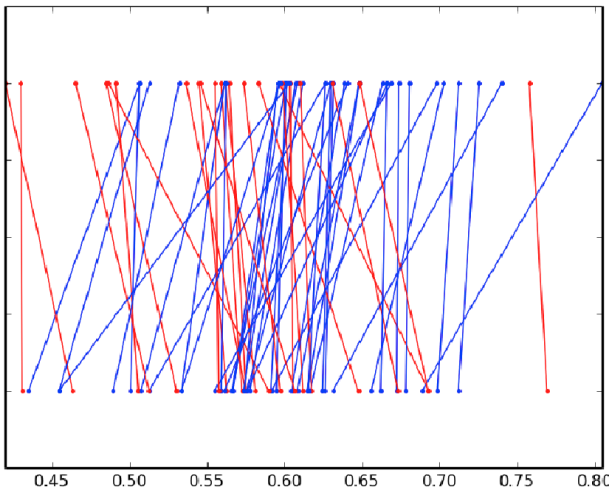

Figure 3: Buddy plot showing cosine vs weighted cosine distances to full report, K=300

pared on the lower row. Documents are connected by a blue line if the weighted vectors brings the document closer to the report, and red lines if the document is considered farther.

4 Experiment

One way to evaluate the performance of these two document similarity approaches is to see how they compare to human judgments of document similarity. I have employed data used by Lee et al.[5] in an experiment evaluating models of text document similarity, where 83 university students assessed the similarity of pairs of 50 documents. The participants were made to judge the similarity of Australian newswire texts (between 51 and 126 words) on a scale of one (”highly unrelated”) to five (”highly related”). I have calculated the average similarity rating of all 8-12 judgments of every potential pair.

Topic distribution vectors were created for each document, along with coherence-weighted topic distribution vectors, with the number of topics set to 100. The cosine distance between every potential pair was calculated (1224 pairs as documents are not compared to themselves) and I am using Pearson’s correlation coefficient to measure the correlation between participants’ similarity judgments of every pair and the cosine distance between topic vectors.

5 Results

	Correlation
LDA-human	-0.2337402
LDA+coherence-human	-0.1296247

Table 2: Correlation between topic vectors and human similarity judgments (LDA-human) and weighted topic vectors and judgments (LDA+coherence-human)

Table 2 shows Pearson's correlation coefficient scores between the average human rating of document pair similarity on a scale of 1-5 and either topic vector. I am calculating the distance between vectors rather than returning a similarity measure, so a negative correlation is to be expected, as similar documents should return a smaller distance but be rated with a higher score. Figure 4 and Figure 5 show scatter plots of these correlations, with non-weighted and weighted topic vectors respectively.

6 Discussion

The evaluation of either approach reveals a stronger negative correlation between human similarity judgments and cosine distances using topic vectors that have not been weighted. This can be accounted for by several factors. It is possible that measuring the distance between vectors using the cosine of the angle fails to capture the topic-quality information provided by the topic coherence weights. As I'm multiplying every topic value by the coherence score of that topic, a distance metric which measures the magnitude of the vectors as well, and not only orientation, might be more suitable.

Another reason for the low performance of the weighted-vector approach might be caused by the size of the evaluation data. 50 short texts might not be enough for LDA to produce meaningful topics. This possibility is supported by the unpredictable fluctuations in average coherence scores observed during the selection of the number of topics. The approach taken by [5], whereby 314 documents of "background data" was provided to the LSA topic model in order to make up for

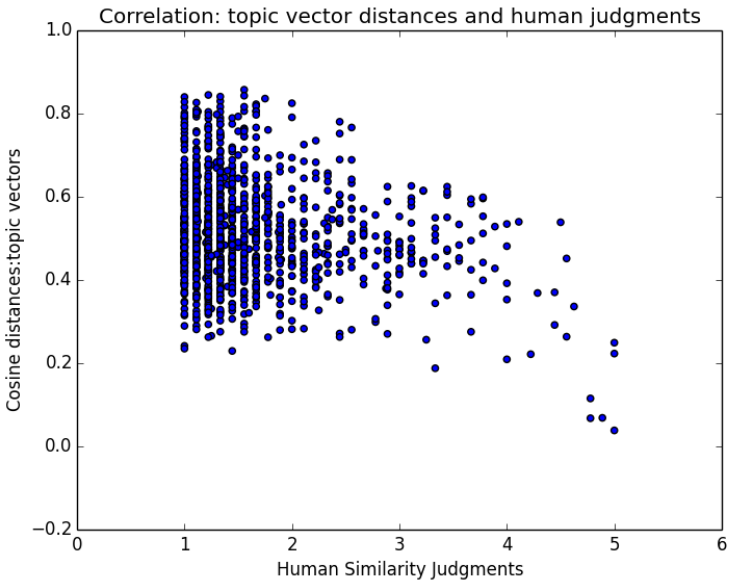

Figure 4: Human judgments plotted over non-weighted topic vector distances

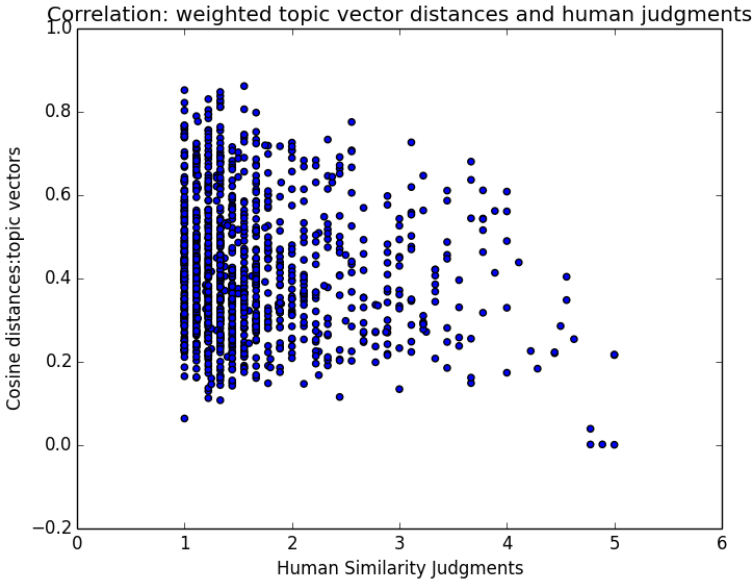

Figure 5: Human judgments plotted over weighted topic vector distances

the insufficiency of the evaluation data, also points to the size of the data being a consideration in experiments involving topic models. Additionally, as noted in [5], participants provided similarity judgments of pairs that leaned heavily towards the lower end of 'related' scale, which may impact both correlation scores (as both of my approaches resulted in poor correlations). Poor correlations between LDA based document features and human assessments of document similarity were also reported in [2].

7 Further Research

Future experiments could include a wider range of document modelling approaches and distance metrics. Results presented by [2] suggest a better performance by Word2vec's word embeddings and its extension for paragraphs, Doc2vec, when used to assess document similarity. In an experiment using the Fisher corpus of topically focused telephone conversations, KL divergence, cosine similarity and the dot product of vectors were compared as measures for identifying topically linked pairs of documents represented by different modelling techniques (such as LDA) [3]. The dot product measure's lower equal error rate indicates that further research should make use of more than cosine distances. Evaluation data for further research could consist of varyingly scored summaries from previous automatic document summarization tasks. A collection of documents that summarize one larger source document would provide more data for a topic model, reflects the intended purpose of my project, and might provide better results for document comparisons involving weighted topic vectors.

8 Conclusion

In this paper, I have explored the use of document similarity measures involving topic vectors generated by LDA topic modelling of documents. I have looked at ways of comparing media articles discussing the report of Canada's Truth and Reconciliation Commission with the report itself, employing LDA topic vectors, with and without weights extracted from each topic's coherence score. On an experiment conducted using human ratings of document similarities, I have found that the weighted topic vectors do not produce better cosine similarity scores, but the correlation between either vector and human judgments is low. Though I have focused on LDA topic vectors and used the summary of the TRC's full report, future work can make use of a variety of document modelling techniques, the full report, and different datasets for evaluation.

References

[1] ALEXANDER, E., AND GLEICHER, M. Task-driven comparison of topic models. *Visualization and Computer Graphics, IEEE Transactions on 22*, 1 (2016), 320–329.

[2] CAMPR, M., AND JEŽEK, K. Comparing semantic models for evaluating automatic document summarization. In *Text, Speech, and Dialogue* (2015), Springer, pp. 252–260.

[3] HAZEN, T. J. Direct and latent modeling techniques for computing spoken document similarity. In *Spoken Language Technology Workshop (SLT), 2010 IEEE* (2010), IEEE, pp. 366–371.

[4] HU, D. J. Latent dirichlet allocation for text, images, and music. *University of California, San Diego. Retrieved April 26* (2009), 2013.

[5] LEE, M., PINCOMBE, B., AND WELSH, M. An empirical evaluation of models of text document similarity. *Cognitive Science* (2005).

[6] MCCALLUM, A. K. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[7] MIMNO, D., WALLACH, H. M., TALLEY, E., LEENDERS, M., AND MCCALLUM, A. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2011), Association for Computational Linguistics, pp. 262–272.