External Evaluation of Topic Models

David Newman Sarvnaz Karimi Lawrence Cavedon

NICTA and The University of Melbourne
Parkville, Victoria 3010, Australia

{*david.newman, sarvnaz.karimi, lawrence.cavedon*}*@nicta.com.au*

Abstract *Topic models can learn topics that are highly interpretable, semantically-coherent and can be used similarly to subject headings. But sometimes learned topics are lists of words that do not convey much useful information. We propose models that score the usefulness of topics, including a model that computes a score based on pointwise mutual information (PMI) of pairs of words in a topic. Our PMI score, computed using word-pair co-occurrence statistics from external data sources, has relatively good agreement with human scoring. We also show that the ability to identify less useful topics can improve the results of a topic-based document similarity metric.*

Keywords Topic Modeling, Evaluation, Document Similarity, Natural Language Processing, Information Retrieval

1 Introduction

Topic models are unsupervised probabilistic models for document collections, and are generally regarded as the state-of-the-art for extracting course-grained semantic information from collections of text documents. The extracted semantic content is useful for a variety of applications including automatic categorization and faceted browsing. The topic model technique learns a set of thematic topics from words that tend to co-occur in documents. The technique assigns a small number of topics to each document, and those topics can then be used to explain and retrieve documents. However this explanation of a document is only useful if we can understand what is meant by a given topic.

Since the introduction of the original topic model approach [Blei et al., 2003, Griffiths and Steyvers, 2004], many researchers have modified and extended topic modeling in a variety of ways. However, there has been less effort on understanding the semantic nature of topics learned by topic models. While the list of the most likely (i.e. important) words in a topic provides good transparency to defining a topic, how can humans best interpret and understand the gist of a topic? Some researchers have started to address this problem, including Mei et al. [2007] who looked at the problem of automatic assignment of a short label for a topic, and Griffiths and Steyvers [2006] who applied topic models to word sense distinction tasks. Wallach et al. [2009] proposed methods for evaluating topic models, but they focused on the statistics of the model, not the meaning of individual topics.

The challenge of helping a user understand a discovered topic is exacerbated by the variable semantic quality of topics produced by a topic model. Certain types of document collections, for example collections of abstracts of research papers, produce mostly high-quality interpretable topics which have clear semantic meaning. However, the broader class of document collections — for example emails, blogs, news articles and books — tend to produce a wider mix of topics. The novelty of our work is targetting this challenge by focusing on evaluation of topics using their degree of usefulness to humans.

In this work we first ask humans to decide whether individual learned topics are useful or not (we define what is meant by useful). We then propose models that use external text data sources, such as Wikipedia or Google hits, to predict human judgements. Finally, we show how an assessment of useful and useless topics can improve the outcome of a document similarity task.

2 Topic Modeling

The topic model — also known as *latent Dirichlet allocation* or *discrete principal component analysis (PCA)* — is a Bayesian graphical model for text document collections represented by bags-of-words (see Blei et al. [2003], Griffiths and Steyvers [2004], Buntine and Jakulin [2004]). In a topic model, each document in the collection of D documents is modeled as a multinomial distribution over T topics, where each topic is a multinomial distribution over W words. Typically, only a small number of words are important (have high likelihood) in each topic, and only a small number of topics are present in each document.

The collapsed Gibbs [Geman and Geman, 1984] sampled topic model simultaneously learns the topics and the mixture of topics in documents by iteratively sampling the topic assignment z to every word in every document, using the Gibbs sampling update

Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia, 4 December 2009.

$$p(z_{id} = t | x_{id} = w, \mathbf{z}^{\neg id}) \propto$$

$$\frac{N_{wt}^{\neg id} + \beta}{\sum_w N_{wt}^{\neg id} + W\beta} \; \frac{N_{td}^{\neg id} + \alpha}{\sum_t N_{td}^{\neg id} + T\alpha},$$

where $z_{id} = t$ is the assignment of the i^{th} word in document d to topic t, $x_{id} = w$ indicates that the current observed word is w, and $\mathbf{z}^{\neg id}$ is the vector of all topic assignments not including the current word. N_{wt} represents integer count arrays (with the subscripts denoting what is counted), and α and β are Dirichlet priors.

The maximum a posterior (MAP) estimates of the topics $p(w|t)$, $t = 1 \ldots T$ and the mixture of topics in documents $p(t|d)$, $d = 1 \ldots D$ are given by

$$p(w|t) = \frac{N_{wt} + \beta}{\sum_w N_{wt} + W\beta},$$

$$p(t|d) = \frac{N_{td} + \alpha}{\sum_t N_{td} + T\alpha}.$$

Pathology of Learned Topics

Despite referring to the distributions $p(w|t)$ as topics, suggesting that they have sensible semantic meaning, they are in fact just statistics that explain count data according to the underlying generative model. To be more explicit, while many learned topics convey information similar to what is conveyed by a subject heading, topics themselves are not subject headings, and they sometimes are not at all related to a subject heading.

Since our focus in this paper is studying and evaluating the wide range of topics learned by topic models, we present examples of less useful topics learned by topic models. Note that these topics are not simply artifacts from one particular model started from some particular random initialization – they are stable features present in the data that can be repeatedly learned from different models, hyperparameter settings and random initializations. The following list shows an illustrative selection of less useful topics:

- north south carolina korea korean southern kim daewoo government country million flag thoreau economic war ... *This topic has associated Carolina with Korea via the words north and south.*
- friend thought wanted went knew wasn't love asked guy took remember kid doing couldn't kind ... *This is a typical "prose" style topic often learned from collections of emails, stories or news articles.*
- google domain search public copyright helping querying user automated file accessible publisher commercial legal ... *This is a topic of boilerplate copyright text that occurred in a large subset of a corpus.*
- effect significant increase decrease significantly change resulted measured changes caused ... *This is a topic of comparisons that was learned from a large collection of MEDLINE abstracts.*
- weekend december monday scott wood going camp richard bring miles think tent bike dec pretty ... *This topic includes a combination of several commonly occurring pathologies including lists of names, days of week, and months of year.*

Collections Modeled

We used two document collections: a collection of news articles, and a collection of books. These collections were chosen to produce sets of topics that have more variable quality than one typically observes when topic modeling collections of scientific literature. A collection of $D = 55,000$ news articles was selected from Linguistic Data Corporation's gigaword corpus, and a collection of $D = 12,000$ books was downloaded from the Internet Archive. We refer to these collections as "News Articles" and "Books" throughout the remainder of this paper.

Standard procedures were used to create the bags-of-words for the two collections. After tokenization, and removing stopwords and words that occurred fewer than ten times, we learned topic models of News Articles using $T = 50$ ($T50$) and $T = 200$ ($T200$) topics, and a topic model of Books using $T = 400$ ($T400$) topics. For each topic model, we printed the set of T topics. We define a topic as the list of ten most probable words in the topic. This cutoff at ten words is arbitrary, but it balances between having enough words to convey the meaning of a topic, but not too many words to complicate human judgements or our scoring models.

3 Human Scoring of Topics

We selected 117 topics from News Articles, including all 50 topics from the $T50$ topic model, and 67 selected topics from the $T200$ topic model. We selected 120 topics from the $T400$ topic model of Books. To increase the expected number of useful and useless topics, we pre-scored topics using our scoring models (described later) to select a mix of useful, useless, and in-between topics to make up the sample. We asked nine human subjects to score each of the 237 topics on a 3-point scale where 3="useful" and 1="useless".

We provided a rubric and some guidelines on how to judge whether a topic was useful or useless. In addition to showing several examples of useful and useless topics, we gave the following instructions to people performing the evaluation:

The topics learned by a topic model are usually sensible, meaningful, interpretable and coherent. But some topics learned (while statistically reasonable) are not particularly useful for human use. To evaluate our methods, we would like your judgment on how "useful" some learned topics are. Here, we are purposefully vague about what is "useful" ... it is some combination of coherent, meaningful, interpretable, words are related, subject-heading like, something you could easily label, etc.

Figure 1 shows selected useful and useless topics from News Articles, as scored by nine people. For our purposes, the usefulness of a topic can be thought of as whether one could imagine using the topic in a search interface to retrieve documents about a particular

Selected useful topics (unanimous score=3):
space earth moon science scientist light nasa mission planet mars ...
health disease aids virus vaccine infection hiv cases infected asthma ...
bush campaign party candidate republican mccain political presidential ...
stock market investor fund trading investment firm exchange companies ...
health care insurance patient hospital medical cost medicare coverage ...
car ford vehicle model auto truck engine sport wheel motor ...
cell human animal scientist research gene researcher brain university ...
health drug patient medical doctor hospital care cancer treatment disease ...

Selected useless topics (unanimous score=1):
king bond berry bill ray rate james treas byrd key ...
dog moment hand face love self eye turn young character ...
art budget bos code exp attn review add client sent ...
max crowd hand flag sam white young looked black stood ...
constitution color review coxnet page art photos available budget book ...
category houston filed thompson hearst following bonfire mean tag appear ...
johnson jones miller scott robinson george lawrence murphy mason ...
brook stone steven hewlett packard edge borge nov buck given ...

Figure 1: Selected useful and useless topics from collection of News Articles. Each line represents one topic.

Selected useful topics (unanimous score=3):
steam engine valve cylinder pressure piston boiler air pump pipe ...
furniture chair table cabinet wood leg mahogany piece oak louis ...
building architecture plan churches design architect century erected ...
cathedral church tower choir chapel window built gothic nave transept ...
god worship religion sacred ancient image temple sun earth symbol ...
loom cloth thread warp weaving machine wool cotton yarn mill ...
window nave aisle transept chapel tower arch pointed arches roof ...
cases bladder disease aneurism tumour sac hernia artery ligature pain ...

Selected useless topics (unanimous score=1):
entire finally condition position considered result follow highest greatest ...
aud lie bad pro hut pre able nature led want ...
soon short longer carried rest turned raised filled turn allowed ...
act sense adv person ppr plant sax genus applied dis ...
httle hke hfe hght able turn power lost bring eye ...
soon gave returned replied told appeared arrived received return saw ...
person occasion purpose respect answer short act sort receive rest ...
want look going deal try bad tell sure feel remember ...

Figure 2: Selected useful and useless topics from collection of Books.

subject. An indicator of usefulness is the ease by which one could think of a short label to describe a topic (for example “space exploration” could be a label for the first topic). The useless News Articles topics display little coherence and relatedness, and one would not expect them to be useful as categories or facets in a search interface.

We see similar results in Figure 2, which shows selected useful and useless topics from the Books collection. Again, the useful topics could directly relate to subject headings, and be used in a user interface for browse-by-subject. Note that the useless topics from both collections are not chance artifacts produced by the models, but are in fact stable and robust statistical features in the data sets.

Our human scoring of the 237 topics has high inter-rater reliability, as shown in Figure 3. Each human score has high agreement with the mean of the remaining scores (Pearson correlation coefficient $\rho = 0.78 \dots 0.81$). In the following sections we present models to predict these human judgements.

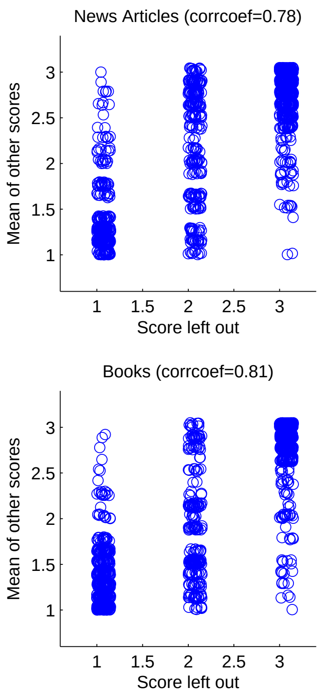

Figure 3: Inter-rater reliability, computed by leave-one-out, showing high agreement between the nine humans.

This inter-rater correlation is an upper bound on how well we can expect our scoring models to perform.

4 Scoring Model I: Pointwise Mutual Information

The intuition behind our first scoring model, pointwise mutual information (PMI) using external data, comes from the observation that occasionally a topic has some odd-words-out in the list of ten words. This leads to the idea of a scoring model based on word association between pairs of words, for all word pairs in a topic. But instead of using the collection itself to measure word association (which could reinforce noise or unusual word statistics), we use a large external text data source to provide *regularization*.

Specifically, we measured co-occurrence of word pairs from two huge external text datasets: all articles from English Wikipedia, and the Google n-grams data set. For Wikipedia we counted a co-occurrence as words w_i and w_j co-occurring in a 10-word window in any article, and for Google n-grams, we counted a co-occurrence as w_i and w_j co-occurring in any of the 5-grams. These co-occurrences are counted over corpora of 1B and 1T words respectively, so they produce reasonably reliable statistics.

We choose pointwise mutual information as the measure of word association, and define the following scoring formula for a topic $\mathbf{w}$:

$$\text{PMI-Score}(\mathbf{w}) = \text{median}\{\text{PMI}(w_i, w_j), ij \in 1 \dots 10\},$$

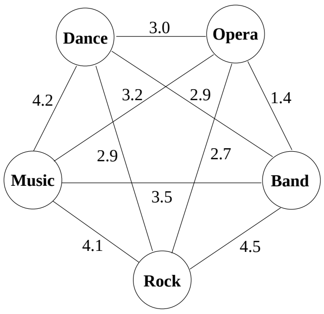

Figure 4: Illustration of pointwise mutual information between word pairs.

$$\mathrm{PMI}(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)},$$

where the top-ten list of words in a topic is denoted by $\mathbf{w} = (w_1, \ldots, w_{10})$, and we exclude the self PMI case of $i = j$. The PMI-Score for each topic is the median PMI for all pairs of words in a topic (so for a topic defined by the top-10 words, the PMI-Score is the median of 55 PMIs). Note that if two words are statistically independent, then their PMI is zero.

Our PMI-Score is illustrated in Figure 4 for a topic of five words: “music band rock dance opera”.[1] Using co-occurrence frequencies from Wikipedia, we see unsurprising high-scoring word pairs, such as PMI(rock,band)=4.5, and PMI(dance,music)=4.2. Some pairs exhibit greater independence, such as PMI(opera,band)=1.4. The PMI-Wiki-Score[2] for this topic is the median of all the PMIs, or PMI-Wiki-Score=3.1.

We see broad agreement between the PMI-Wiki-Score and the human scoring in Figure 5, which shows a scatterplot for all 237 topics. The correlation between the PMI-Wiki-Score and the mean human score is $\rho = 0.72$ for News Articles and $\rho = 0.73$ for Books (we define correlation ρ as the Pearson correlation coefficient). This correlation is relatively high given that the inter-rater-correlation is only slightly higher at $\rho = 0.78 \ldots 0.81$.

Using the Google 5-grams data instead of English Wikipedia for the external data source produces similar results, shown in Figure 6. In this case, the pointwise mutual information values are computed using word statistics from the 1 billion Google 5-grams instead of 2 million Wikipedia articles. The correlations are in a similar range ($\rho = 0.70 \ldots 0.78$) with a slightly higher correlation of $\rho = 0.78$ for News Articles.

Why does our PMI-Score model agree so well with human scoring of topics? Our intuition is that humans consider associations of pairs of words (or the association between one word and all the other words) to determine the relatedness and usefulness of a topic. This human process is somewhat approximated by the calculation of the PMI-Score.

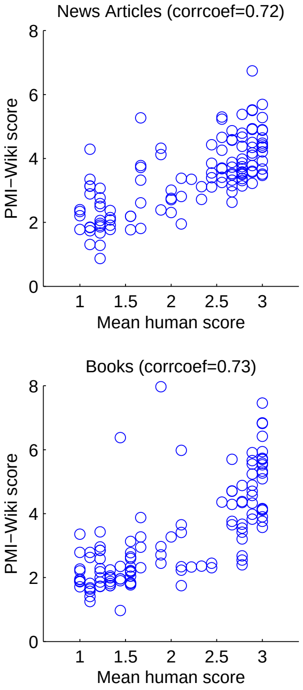

Figure 5: Scatterplot of PMI-Wiki-Score vs. mean human score.

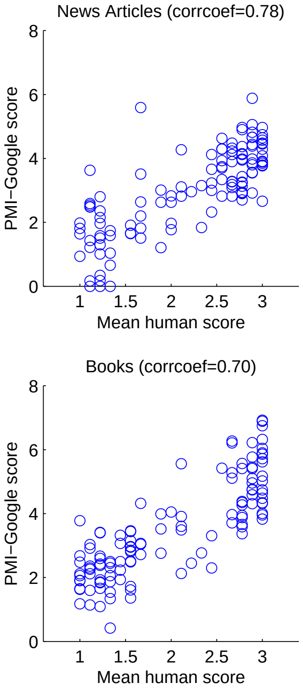

Figure 6: Scatterplot of PMI-Google-Score vs. mean human score.

5 Scoring Model II: Google

In this section we present a second scoring scheme, again based on a large external data source: this time

[1] We illustrate using 5 words instead of 10 for simplicity.

[2] This is the PMI-Score computed using frequency counts from Wikipedia.

the entire World Wide Web crawled by Google. We present two scoring formulas that use the Google search engine:

$$\text{Google-titles-match}(\mathbf{w}) = \mathbf{1}\left[w_i = v_j\right],$$

where $i = 1, \ldots, 10$ and $j = 1, \ldots, |V|$, and v_j are all the unique terms mentioned in the titles from the top-100 search results, and **1** is the indicator function to count matches; and

$$\text{Google-log-hits}(\mathbf{w}) = \log(\text{\# results from search for } \mathbf{w}),$$

where $\mathbf{w}$ is the search string "+w$_1$ +w$_2$ +w$_3$ . . . +w$_{10}$". We use the Google advanced search option '+' to search exactly as is and prevent Google from using synonyms. Our intuition is that the mention of topic words in URL titles — or the prevalence of documents that mention all ten words in the topic — may better correlate with a human notion of the usefulness of a topic.

For example, issuing the query to Google: "+space +earth +moon +science +scientist +light +nasa +mission +planet +mars" returns 171,000 results (so Google-log-hits($\mathbf{w}$)=5.2), and the following list shows the titles and URLs of the first 6 results:

1. NASA - STEREO Hunts for Remains of an Ancient Planet near Earth (science.nasa.gov/headlines/y2009/...)
2. NASA - Like Mars, Like Earth (www.nasa.gov/audience/foreducators/k-4/features/...)
3. NASA - Like Mars, Like Earth (www.nasa.gov/audience/forstudents/5-8/features/...)
4. ASP: The Silicon Valley Astronomy Lectures Podcasts (www.astrosociety.org/education/podcast/index.html)
5. NASA calls for ambitious outer solar system mission - space ... (www.newscientist.com/article/...)
6. NASA International Space Station Mission Shuttle Earth Science ... (spacestation-shuttle.blogspot.com/2009/08/...)

The underlined words show mentions of topic words in the URL titles, with the first six titles giving a total of 17 mentions. The top-100 URL titles include a total of 194 matches, so for this topic Google-titles-match($\mathbf{w}$)=194.

We see surprisingly good agreement between the Google-titles-match score and the human scoring in Figure 7 for the News Articles ($\rho = 0.78$), and a lower level of agreement for Books ($\rho = 0.52$). In the PMI-Scores there was no clear pattern of outliers in the scatterplots against the mean human score. However, we see a definite constraint of the Google-titles-match score, where there are many topics that received a high human score, but a low Google-titles-match score. Table 1 shows selected topics having a high human score (useful), but a low Google-titles-match score. The first three topics listed (from News Articles) show different types of problems. The first topic is clearly about cooking, but does not mention the word cooking. Furthermore, it is unlikely that URL titles would include words such as "teaspoon" or "pepper", so we are not surprised that Google-titles-match fails to give this topic a high score. The second topic is mostly about NASA and space exploration, but is polluted by the words "firefighter" and "worcester", which will severely limit the number of results returned. By using the median, the PMI-Score of this topic is less sensitive to these words that don't fit the topic, but the Google-titles-match has less hope of producing a useful list of search results when all ten words are included in the search query. Topics from Books follow, and we see a similar problem to the cooking topic from News Articles, where the words in the topic clearly convey something semantically coherent, but fail to evoke URL titles that mention those general terms.

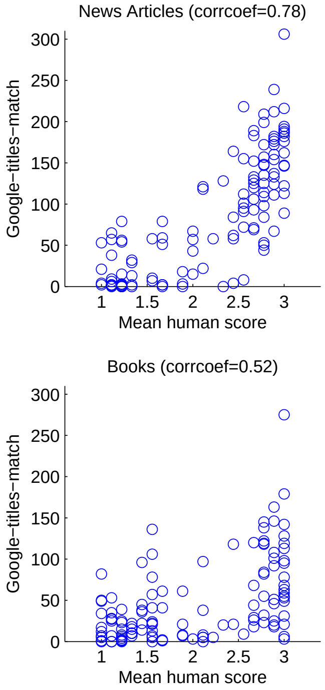

Figure 7: Scatterplot of Google-titles-match score vs. mean human score.

We see less promising results from our Google-log-hits score, which has relatively low correlation with the mean human scoring ($\rho = -0.09 \ldots 0.49$), as shown in the scatterplots in Figure 8. For this scoring formula we observed the reverse of the problem of Google-titles-match, namely we saw overly favorable scoring of many topics that received a low human score. Table 2 shows selected topics having a low human score (not useful), but a high Google-log-hits score. The topics in this table all exhibit the similar characteristic of all ten words being relatively common words. Consequently there exist many web pages that contain these words (issuing these topics as queries returned between 250,000 and 10,000,000 results). This behavior of Google-log-hits and failure to agree with human scoring (in this case) is relatively easy to understand.

Human	Titles-match	Topic
2.6	8	cup add tablespoon salt pepper teaspoon oil heat sugar pan ...
2.4	4	space nasa moon mission shuttle firefighter astronaut launch worcester rocket ...
2.3	0	oct series braves game yankees league bba met championship red ...
2.9	25	church altar churches stone chapel cathedral vestment service pulpit chancel ...
3.0	6	cases bladder disease aneurism tumour sac hernia artery ligature pain ...
2.8	23	art ancient statues statue marble phidias artist winckelmann pliny image ...
3.0	3	window nave aisle transept chapel tower arch pointed arches roof ...
2.9	18	crop land wheat corn cattle acre grain farmer manure plough ...
2.8	32	account cost item profit balance statement sale credit shown loss ...
2.9	20	pompeii herculaneum room naples painting inscription excavation marble bronze bath ...
3.0	21	window nave choir arch tower churches aisle chapel transept capital ...
3.0	31	drawing draw pencil pen drawn model cast sketches ink outline ...

Table 1: Disagreement between high human scores and low Google-titles-match scores.

Human	log hits	Topic
1.0	5.4	dog moment hand face love self eye turn young character ...
1.2	7.0	change mean different better result number example likely problem possible ...
1.2	6.4	fact change important different example sense mean matter reason women ...
1.1	5.9	friend thought wanted went knew wasn't love asked guy took ...
1.1	5.6	thought feel doesn't guy asked wanted tell friend doing went ...
1.1	6.1	bad doesn't maybe tell let guy mean isn't better ask ...
1.0	6.7	entire finally condition position considered result follow highest greatest fact ...
1.0	6.3	soon short longer carried rest turned raised filled turn allowed ...
1.1	6.1	modern view study turned face detail standing born return spring ...
1.2	6.3	sort deal simple fashion easy exactly call reason shape simply ...
1.1	6.4	proper require care properly required prevent laid making taking allowed ...
1.0	6.7	person occasion purpose respect answer short act sort receive rest ...
1.0	6.1	want look going deal try bad tell sure feel remember ...
1.2	6.3	saw cried looked heard stood asked sat answered began knew ...

Table 2: Disagreement between low human scores and high Google-log-hits scores.

6 Document Similarity

Discovering semantically similar documents in a collection of unstructured text has practical applications, such as search by example. Many studies have been proposed to calculate inter-document similarity since 1950s. For example, Grangier and Bengio [2005] use hyperlinks to score linked documents on the Web higher than unlinked for information retrieval tasks. Kaiser et al. [2009] use Wikipedia to find similar documents for a focused crawler (they also provide a good literature review on recent approaches that use support vector machines, latent semantic analysis (LSA), or explicit semantic analysis). Lee et al. [2005] empirically compare between three categories of binary, count, and LSA similarity models over a small corpus of human judged texts and concluded that evaluation of such models should occur in the context of their applications.

Humans judge two texts to be similar if they share the same concepts or topics [Kaiser et al., 2009]. We use our learned topics from News Articles to find similar documents and compare them against count-based models implemented in a search engine. Our preliminary findings show that if documents contain useless text — words that are not related to the main topic of the text or bear no content, such as advertisements — then they are likely to be mistakenly considered similar using document similarity metrics that rely on term frequencies. Below, we explain our experimental setup and results.

Count-Based Similarity

We used the Okapi BM25 [Walker et al., 1997] ranking function implemented in the Zettair[3] search engine. Similarity scores are based on term frequency and inverse document frequencies in a document collection.

Topic-Based Similarity

A document similarity measure using topics was computed using Hellinger distance. For every pair of documents d_i and d_j in a collection, and a set T of learned topics, Hellinger distance is computed as below:

$$\text{dist}(d_i, d_j) = \frac{1}{2} \sum_{t=1}^{T} \left(\sqrt{p(t|d_i)} - \sqrt{p(t|d_j)} \right)^2 ,$$

$$\text{dist}^*(d_i, d_j) = \frac{1}{2} \sum_{t \in \text{useful}} \left(\sqrt{p(t|d_i)} - \sqrt{p(t|d_j)} \right)^2 ,$$

[3] http://www.seg.rmit.edu.au/zettair/

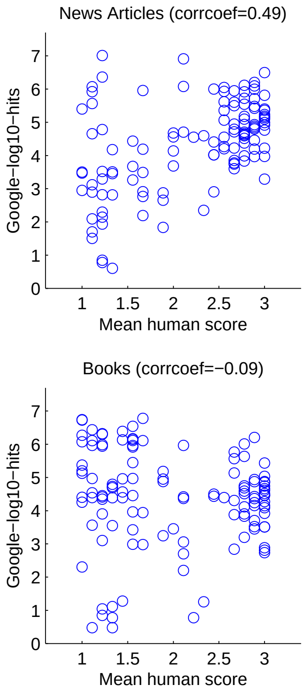

Figure 8: Scatterplot of Google-log-hits score vs. mean human score.

where $p(t|d_i)$ and $p(t|d_j)$ are probabilities of topics in documents i and j. We provide two formulas for Hellinger distance, one based on all topics, and dist* that uses just the “useful” topics.

Experimental Setup

Fifty documents were randomly selected from News Articles based on their proportion of useful and useless topics. An overview of the documents in the collection based on their percentages of useless text is shown in Figure 9. Our aim is to improve document similarity calculations on the right tail of this graph where the documents contain a larger proportion of useless text which could mislead document similarity methods that rely on the frequency of terms. We therefore first extracted those documents that contained at least 30% useful content (based on PMI-Wiki-Score) and at least 40% non-content text. We then calculated the similarity scores of 50 randomly selected documents from this subset with other documents in the collection. For count-based methods, we used each of these 50 full documents as queries to retrieve a ranked list of similar documents using the Zettair search engine. For the topic-based method, two approaches were used: using all the topics generated for the collection ($T200$), and using useful topics as based on the topics’ PMI-Wiki-Score.

In a preliminary experiment, a human judge was presented with original documents and the top most similar document (Top-1) extracted by each method. The human judge was not aware of the order of methods which the documents were retrieved. A simple binary scoring of *similar* or *not-similar* was used. The criteria for similarity was the overall subject of the documents, for example, both being about a specific sport. For 32 of 50 cases (64%), all methods successfully resulted in documents judged to be similar by the human judge. In only one case did Okapi outperform both topic-based methods. Using the useful-topics metric (dist*) led to 94% accuracy against similarity judgements; all topics (dist) was 88% accurate; Okapi was 70% accurate. Also, the overlap between the ranked outputs of the two systems, Okapi and useful topics, was very low: 30% in Top-1 overlapped (the documents were the same for the both systems).

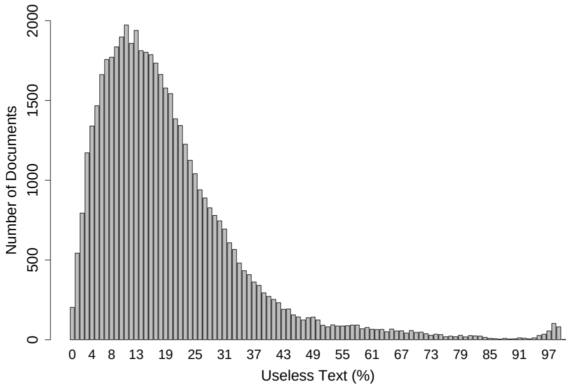

Figure 9: Number of documents versus proportion of usefuless content. 4.3% of documents have more than 50% useless text and 16.4% have more than 30% useless text.

Figure 10 shows an illustrative example where using topic modeling, in particular using good topics (i.e. dist*), outperforms Okapi when the original document contains a large proportion of non-content text.

While he experiments described in this section are limited in scope, they constitute an initial investigation into the task-level effectiveness of topic-based metrics that ignore “useless” topics. We believe that the results indicate that, for texts that contain “noise”, identifying the “useful” topics in a topic model has promising applications.

7 Conclusion

Evaluation of topic modeling — the analysis of large sets of unstructured documents and assignment of series of representative words as topics to clusters of documents — has hardly been investigated. In particular, meaning of the topics and human perception of their usefulness had not been studied before. Here, we investigated topic modeling evaluation using external data (Wikipedia documents, Google n-grams, and Google hits), and compared our proposed methods with human judgments on usefulness of the topics. According to our experiments on collections of news articles and books, a scoring method using pointwise mutual information

Original Document
At last! A biography that skips the saint-or-sinner debate. As Dusko Doder and Louise Branson abundantly document, Slobodan Milosevic, almost from the start, epitomized the Balkan-variety bad seed. The child of parents who both committed suicide, Milosevic aligned himself with a woman who hungered for power to avenge the ignominious death of her mother. Milosevic betrayed a college classmate, a mentor of two decades, and his next-door neighbor in lunging to the top of Yugoslavia's diseased post-Tito political leadership. And "Milosevic: Portrait of a Tyrant"...
...
(gm)
FOR WEDNESDAY AMs
Here are today's top news stories from The New York Times News Service for ally at LaSalle University for of Wednesday, Dec. 22: INTERNATIONAL ("i" code) CHINA-INTERNET (Beijing) - With the ambivalent blessing of the Chinese government, locally produced web sites and chat rooms have spread rapidly here in the last two years,... RUSSIA-U.S.-AID (Washington) - The State Department, invoking a seldom-used law, may block a $500 million loan package for Russia's oil sector. By David E. Sanger...

Okapi BM25 (Zettair)
More New Yorkers would vote against Hillary Rodham Clinton as a U.S. Senate candidate than vote for her, a new poll indicates. The survey by the Zogby International polling organization shows the probable Democratic nominee carrying an "unfavorable rating" of 48.4 percent among likely voters, as opposed to her "favorable rating" of 46.3 percent. It marks the first time the potential candidate's statistical negatives have eclipsed her positives in her still-undeclared campaign, pollster John eet, a city councilman in a tight race. "But I hope you of Utica said Tuesday.
...
(gm)
FOR WEDNESDAY AMs
Here are today's top news stories from The New York Times News Service for AMs of Wednesday, Dec. 22: INTERNATIONAL ("i" code) CHINA-INTERNET (Beijing) - With the ambivalent blessing of the Chinese government, locally produced web sites and chat rooms have spread rapidly here in the last two years,... RUSSIA-U.S.-AID (Washington) - The State Department, invoking a seldom-used law, may block a $500 million loan package for Russia's oil sector. By David E. Sanger.
...

All Topics
We may be living in a high-tech era but it still takes a low-tech truck to deliver something you've ordered over the Internet, which is why Forbes magazine picked Atlanta-based United Parcel Service as its "company of the year." "With 157,000 ground vehicles, 610 aircraft and $11 billion invested in technology, UPS moves both atoms and bits," says Forbes in announcing its "platinum list" of "America's best big companies." According to Forbes, UPS's role as a shipper of 6 percent of the nation's gross domestic product makes it "the missing link in the burgeoning world of E-commerce."
...
Story Filed By Cox Newspapers (gm) Here are the stories New York Times editors are planning for Tuesday, Dec. 28 Page 1. The NYT frontpage advisory, with layout description, will move by 7:30 p.m. ET. The NYT News Service Night Supervisor is Pat Ryan (888-346-9867). ISRAEL-POLITICS (Jerusalem) - The Shas political party, which represents Sephardic Jews of Middle Eastern and North African descent, announced Monday that it had decided to quit the coalition government of Israeli Prime Minister Ehud Barak.
...

Useful Topics
The Clinton administration, in a move intended to bolster opponents of President Slobodan Milosevic, has agreed to lift economic sanctions on Serbia as soon as there is a free election there, senior administration officials said on Tuesday. The administration had previously vowed that it would not lift the sanctions until Milosevic had been removed from power. But officials calculate that the new strategy should allow the Serbian opposition to increase popular pressure on Milosevic, to call early elections, since holding a free election would mean an end to an oil embargo, an air-travel ban and other sanctions that have weakened an already devastated Serbian economy. Secretary of State Madeleine Albright is expected to make the announcement Wednesday, but it carries a risk: that bickering opposition parties would so fragment the election results that Milosevic might be able to cling to power or, far less likely, that he would win outright in the balloting.
...
Although the constitution of the Yugoslav federation of Serbia and neighboring Montenegro does not grant Milosevic direct power to call new elections, the reality is that his powers are dictatorial
...

Figure 10: An example of top ranked similar documents returned by three methods: Okapi scores generated by Zettair, topic-based similarity using all topics (dist), and topic-based similarity only using useful topics. Using only useful topics (dist*) produces the best result.

on Wikipedia documents and Google n-grams has great potential to distinguish useful (or meaningful) topics from useless ones. This finding is supported by high correlation between our scoring approaches and human judgements on the same topics. We also showed a possible application for distinguished useful topics in extraction of similar documents in a collection.

Acknowledgements NICTA is funded by the Australian government as represented by Department of Broadband, Communication and Digital Economy, and the Australian Research Council through the ICT centre of Excellence programme. DN has also been supported by a grant from the Institute of Museum and Library Services, and a Google Research Award.
Authors are thankful to Timothy Baldwin for valuable discussions.

References

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

W. L. Buntine and A. Jakulin. Applying discrete PCA in data analysis. In *Proceedings of the 20th Uncertainty in Artificial Intelligence Conference*, pages 59–66, Banff, Canada, 2004.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images. volume 6, pages 721–741, November 1984.

D. Grangier and S. Bengio. Inferring document similarity from hyperlinks. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 359–360, Bremen, Germany, 2005.

T. Griffiths and M. Steyvers. Finding scientific topics. In *Proceedings of the National Academy of Sciences*, volume 101, pages 5228–5235, 2004.

T. Griffiths and M. Steyvers. Probabilistic topic models. In *Latent Semantic Analysis: A Road to Meaning*, 2006.

F. Kaiser, H. Schwarz, and M. Jakob. Using Wikipedia-based conceptual contexts to calculate document similarity. In *Proceedings of the 2009 Third International Conference on Digital Society*, pages 322–327, Cancun, Mexico, 2009.

M. D. Lee, B. Pincombe, and M. Welsh. An empirical evaluation of models of text document similarity. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 1254–1259, Mahwah, NJ, 2005.

Q. Mei, X. Shen, and C. Zhai. Automatic labeling of multinomial topic models. In *Proceedings of The 30th International Conference on Knowledge Discovery and Data Mining*, pages 490–499, 2007.

S. Walker, S. Robertson, M. Boughanem, G. Jones, and K. Sparck Jones. Okapi at TREC-6 automatic ad hoc, VLC, routing, filtering and QSDR. In *Proceedings of the 6th Text REtrieval Conference*, pages 125–136, 1997.

H. M. Wallach, I. Murray, R. Salakhutdinov, and D. M. Mimno. Evaluation methods for topic models. In *Proceedings of The 26th International Conference On Machine Learning*, pages 1105–1112, Quebec, Canada, 2009.