Matching Results of Latent Dirichlet Allocation for Text

Andreas Niekler (aniekler@fbm.htwk-leipzig.de)
Leipzig University of Applied Sciences (HTWK), Faculty of Media
Karl-Liebknecht-Str. 145, 04277 Leipzig, Germany

Patrick Jähnichen (jaehnichen@informatik.uni-leipzig.de)
Natural Language Processing Group, Department of Computer Science
University of Leipzig
Johannisgasse 26, 04103 Leipzig, Germany

Abstract

Many approaches have been introduced to enable Latent Dirichlet Allocation (LDA) models to be updated in an on-line manner. This includes inferring new documents into the model, passing parameter priors to the inference algorithm or a mixture of both, leading to more complicated and computationally expensive models. We present a method to match and compare the resulting LDA topics of different models with light weight easy to use similarity measures. We address the on-line problem by keeping the model inference simple and matching topics solely by their high probability word lists.

Keywords: Latent Dirichlet Allocation, topic distance measures, on-line topic tracking

Introduction

As massive amounts of information become available on-line, text mining applications have become an integral part of both industry and academia. One field of text mining is the identification and extraction of semantic concepts in text documents. Over the last decade, Latent Dirichlet Allocation (D. M. Blei, Ng, & Jordan, 2003) (LDA) has become one of the most popular methods to approach this task. LDA is a Bayesian model that makes use of latent variables[1], which represent the semantic concepts (associated with LDA and models building on LDA, these concepts are known as *topics*), to compute the posterior probability over the latent variables and model parameters to allow the extraction of latent semantic structures in texts (i.e. the topics). Examination of the posterior allows an approximation of probability distributions for both documents and topics[2].

Having a technique at hand to identify different topics, the wish to study their evolution over time evolves naturally. This includes both the analysis of static corpora as well as data that comes in constantly via a stream, the latter of which mostly relies on the segmentation of the data into different time slices of predefined size (e.g. one hour, one day, one year etc.), treating newly arrived data as a new time slice after its size is reached. Another way of handling the data and tracking topic trends without segmentation into time slices is that introduced by (Wang & McCallum, 2006), where the authors use the time stamps of documents as an additional (continuous) observed variable in the model. However, in our approach we resort to the notion of time slice separated data.

[1]For an introduction to latent variable models see (Bishop, 1999)

[2]For documents, a probability distribution over the set of latent topics and analogous to that, for topics, a probability distribution over a fixed vocabulary is inferred

The main problem of tracking topics' evolutions over time, either statically or in an on-line manner is the identification of identical topics in consecutive time slices or data windows[3]. To overcome this, previous approaches such as (D. Blei & Lafferty, 2006; AlSumait, Barbará, & Domeniconi, 2008) use the model outcome of time $t-1$ as a prior for the model at time t or analogously the outcome of a data window as a prior for another sub-set of documents. As this is rather an elegant way to align topics between time slices (from a mathematical point of view), these methods suffer from two serious drawbacks concerning the analysis of diachronic document collections. First, those models are restricted to use the same number and effectively the same topics in each time slice and are bound to measure the amount of change a specific topic undergoes from time $t-1$ to t instead of just aligning possibly identical topics. This prevents from finding newly arising and also from releasing "died", i.e. now unused topics or could even lead to the connection of unrelated topics. Second, the approach of using the outcome of the model at time $t-1$ (or a data window) as an input for the model at time t (or another set of documents) forces the analysis to be processed in a one-after-another fashion, preventing parallel processing of the data.

For the sake of completeness, it shall be stated that another approach to this field is known as *Topic Detection and Tracking*, for which (Allan, 2002) gives a detailed introduction.

In this paper we propose the matching of topics from subsequently trained LDA models via lightweight statistical similarity measures. Our approach is motivated by the finding that a major probability mass in topics' distributions over a vocabulary is represented only by a small number of highly probable words in the distribution. We therefore restrict ourselves to using only a subset of words, together with their probabilities to match different topics. This enables us to independently train the LDA models on each time slice, including both parameter and number of topics optimization per time slice. Further enhancements, such as using hierarchical Bayesian models (e.g. the hierarchical Dirichlet process model introduced by (Teh, Jordan, Beal, & Blei, 2006)) instead of optimizing the number of topics per time slice, are possible without altering the approach.

[3]A sub-set of documents from a bigger corpus.

The paper is organized as follows. In section 2, we review the underlying LDA model and describe our approach for matching topics of different time slices in section 3. Section 4 relates our approach to previous ones and subsumes them. We give an overview over the different similarity measures we took into consideration for solving the task in section 5 and present experiments and results in sections 6 and 7 using hand selected topics generated from a document corpus from the UK-based newspaper *The Guardian*, collected through an API on consecutive days from March, 10th through March, 15th 2011. Finally, we conclude giving an outlook to possible applications and future work.

LDA Model

Before defining our approach for matching topics, we first give a review of the statistical model of LDA and a Gibbs sampling algorithm introduced by (Griffiths & Steyvers, 2004), as a method for inference in the model. LDA is a hierarchical Bayesian model that encodes the relation between words and documents via the latent topics in a document corpus. Herein, documents are not directly linked to words but through latent variables z that govern the responsibility of a certain topic for the words in a document. As words are the observable variables in this model, conditional independence holds true for the document and topic distributions θ and ϕ. Placing prior distributions with hyperparameters α and β over θ and ϕ respectively completes the probabilistic model. A generative process for document generation is given by

1. draw K multinomials $\phi_k \propto Dir(\beta_k)$, one for each topic k
2. for each document d, $d = 1,\ldots,D$
   (a) draw multinomial $\theta_d \propto Dir(\alpha_d)$
   (b) for each word w_{dn} in document d, $n = 1,\ldots,N_d$
      i. draw a topic $z_{dn} \propto Multinomial(\theta_d)$
      ii. draw a word w_{dn} from $p(w_{dn}|\phi_{z_{dn}})$, the multinomial probability conditioned on topic z_{dn}

Exact inference is not tractable in this model, thus we utilize Gibbs sampling as described by (Griffiths & Steyvers, 2004). This includes computing the posterior distribution over all variables and model parameters instead of inferring θ and ϕ directly. Examination of the posterior then yields both distributions. The posterior distribution over topic assignments to words, conditioned on the words and all other topic assignments is given by

$$p(z_i = j|\mathbf{z}_{\setminus i},\mathbf{w}) \propto \frac{C^{VK}_{w_{\setminus i},j}+\beta_{w_i}}{\sum_{v=1}^{V}(C^{VK}_{v_{\setminus i},j}+\beta_w)} \frac{C^{DK}_{d_{\setminus i},j}+\alpha_j}{\sum_{k=1}^{K}(C^{DK}_{d_{\setminus i},k}+\alpha_k)} \quad (1)$$

where C^{VK} and C^{DK} are count matrices with dimensions $V \times K$ and $D \times K$, representing the number of times, a word has been assigned to a topic and the number of times, a topic has been assigned to a document, respectively. Subscript $\setminus i$ excludes the current assignment. Both matrices can be stored efficiently, using a sparse matrix representation, allowing a large vocabulary and thus large document corpora to be processed. Examination of the posterior leads to approximations of both ϕ and θ, which are given as the first and second fraction of equation (1). Consequently, ϕ can be interpreted as a matrix of size $V \times K$, containing the conditional probability $p(w_i|z_k)$ at position $\phi_{i,k}$. Hence, every column vector of ϕ, $\phi_{\cdot,k}$ can be construed as a probability distribution over the whole vocabulary of size V for topic k. The row vectors $\theta_{k,\cdot}$ of matrix θ with $\theta_{k,m} = p(z_k|d_m)$ can then be seen as probability distributions over all latent topics for every document m accordingly. A representation of the individual topics is usually given by a list of n words having highest probability in a topic. This is done by sorting the individual $\phi_{\cdot,k}$ in descending order and retrieving the first n entries afterwards as shown in Table 1.

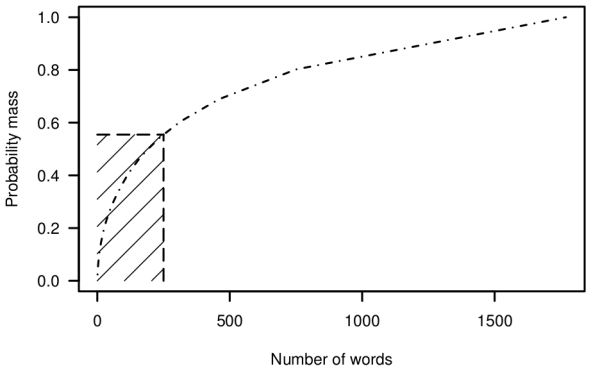

Figure 1: CDF plot for sorted $p(\mathbf{w}|z_k)$ probability distribution example

Matching LDA model posterior distributions

The target is to define a function $sim(p(\mathbf{w}|z_k), p(\mathbf{w}|z^*_k))$ that allows a satisfying separation of topics, so that we are able to define a threshold of similarity that adequately matches identical topics across different models. The outcome of the similarity function $sim(\cdot,\cdot)$ should span a wide range of values, i.e. the function's outcome for similar topics and dissimilar topics has to differ significantly. Otherwise, setting a general optimal threshold obviously becomes practically impossible. The posterior distributions over words given the topics $\phi_{\cdot,k} = p(\mathbf{w}|z_k), (k = 1,\ldots,K)$ can be interpreted as the semantic context or latent structure of the analyzed text corpus. These distributions are used to summarize the corpus contents as short lists of words, giving the intuition that there is only a small number of terms that form the main context of a topic within an LDA model. In Figure 1 we show that indeed only a minor count of terms represent a major probability portion within a topic $p(\mathbf{w}|z_k)$. To demonstrate this property, we built the cumulative distribution function (CDF) for an example

topic after sorting the distribution's probability values in descending order. Although the distribution over words for a topic depends on the β prior of the model, we observed this behavior in models where the inferred topics allow an intuitive interpretation (Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009; Newman, Lau, Grieser, & Baldwin, 2010). Another reading of this finding is the fact that only high probability words are of importance for the topic since all words (belonging to a long tail) of low probability in a topic have about the same mass within all other topics. Thus, the probability mass of the words with highest probability is also constant across topics and independent of the actual words. Considering this and the topic representations in Table 1, the intuition arises, that a similarity function based on simple word matching in sublists of high probability terms from the posterior distributions $p(\mathbf{w}|z_k)$ and $p(\mathbf{w}|z_k^*)$ of different models can help considerably to track topics across different models.

Related Work

To distinguish our work we will briefly discuss related approaches in more detail. The ability of topic models to analyze changes in semantic contexts of continuous document streams was introduced by (D. Blei & Lafferty, 2006; AlSumait et al., 2008). As already described in section 1, these approaches use the outcome of a model from a previous chunk of data, e.g. a time slice $t-1$, and utilize it as the prior for a new succeeding time slice t. In both setups the authors use a fixed number of topics to be inferred from the data[4]. In detail, they use the posterior distribution $p(\mathbf{w}|z_k)$ of topic k to formulate a prior β_k for model inference in succeeding data chunks. In a setup dealing with continuous streams or consecutive corpora, the main idea is, that contents in a data stream are stable over a ceratin time frame. Although, the method of generating priors from posteriors differs in both approaches, the idea of keeping the context of the corpus over time is the same. To incorporate knowledge about changes or stability of topics, measures like the KL divergence are used (see (AlSumait et al., 2008)). Finally, the change of a topic's context is anticipated when the topic's distribution in previous models differs from the current one.

Based on these ideas, analysis of the topics' evolution in a corpus is feasible by fixing the number of topics and dividing the data into chunks or time slices. Unfortunately, this approach is limited to using the same number of topics in each chunk, which is not optimal when the number of concepts in a text stream e.g. in news streams changes. In that case, having a fixed number of topics is inapt. Consequently, optimization of the topic models for each time slice/chunk of data separately seemed desirable to us, especially in the setting of highly dynamic news data streams. Thus independent topic models for each time slice have been used in our approach. Optimization includes inference of hyperparameters[5] and determining an optimal number of topics for the data (as in (Griffiths & Steyvers, 2004)). We produce the relationship between models afterwards via the proposed approach. The benefit of this idea is that we can detect newly arising as well as vanishing topics with exact quantities and can distributively process the models on different CPU's or machines.

Similarity measures

Different measures exist for comparing probability distributions (or real valued vectors in $\mathbb{R}^V$ as a generalization thereof). Since we are working with different corpora or text chunks of unequal size we cannot use absolute word counts to deduce the probability distributions $p(\mathbf{w}|\mathbf{z})$ for each model as has been done by (AlSumait et al., 2008). Instead, we use normalized probability distributions over the vocabulary as a representation of topics that are given by $\phi_{\cdot,k}$ for each topic k. Naturally, elements of $\phi_{\cdot,k}$ are probabilities in the range $]0,1[$. Thus using metrics based on point distances in euclidean space will result in very low values in general that tend to be useless to correctly distinguish between a match or a mismatch.

In our experiments we will create similarity matrices, hence we defined the proposed measures as similarities. The following measures have been evaluated in our experiments:
Jensen-Shannon divergence (JSD): Since we are dealing with probability distributions we chose this measure as a smoothed and symmetric alternative to the Kullback-Leibler (KL) divergence, which is a standard measure for comparing distributions. Note that the outcomes of JSD need to be normalized. The normalized values can than be transformed into a similarity measure by subtracting them from 1. In the following equation we set the distributions $p(\mathbf{w}|z_k)$ and $p(\mathbf{w}|z_k^*)$ to be compared as P and Q and use:

$$JSD(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M) \tag{2}$$

where

$$M = \frac{1}{2}(P+Q). \tag{3}$$

Cosine similarity: Interpreting the posterior distributions $p(\mathbf{w}|\mathbf{z})$ for a topic model as weighted word vectors, the cosine similarity is an unorthodox but nevertheless valid measure. Since it describes the angle between two vectors, the similarity is independent of the norm of the vectors and gives equal results as for unnormalized word counts. Note that the cosine similarity almost identical to the normalized correlation coefficient (Manning & Schütze, 1999) in our case: Since, due to the low probability of most words, the word distribution's mean is close to 0, the calculation of the correlation between

[4]Both approaches rely on LDA as the topic-generating statistical model, and thus are bound to define the LDA model parameter K, i.e. the number of topics the model produces

[5]Hyperparameters strongly influence the model outcome and thus must be optimized according to the intended task. One might analyze newspapers based on editorial departments, whereas others might search for very atomic topics. The latter, however, will not be possible using the mentioned prior based approaches due to a high variance in the topic counts.

two probability vectors will result in a value very close to the normalized correlation coefficient and won't take any negative values. For that reason computation of the correlation between two vectors has been skipped for its redundancy. We set the distributions $p(\mathbf{w}|z_k)$ and $p(\mathbf{w}|z_k^*)$ to be compared as A and B and use:

$$s(A,B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (4)$$

Dice's coefficient: Consider the summary of topics by a list of the top n words per topic as in Table 1. Looking at the lists, e.g. the japan topics, we can identify the similarity or the overlapping of the contents by just inspecting the words without using their actual probability. Following this idea, we also consider another similarity measure based on word sets, Dice's coefficient, that might seem unusual to compare different probability distributions. We set the words from the sorted distributions $p(\mathbf{w}|z_k)$ and $p(\mathbf{w}|z_k^*)$ to be compared as X and Y.

$$s = \frac{2|X \cap Y|}{|X| + |Y|} \quad (5)$$

Experiments

Our dataset consists of 2,133 news articles from five consecutive days (March 10th through 15th 2011) containing 64,674 unique word types, obtained through the API of the British newspaper *The Guardian*. Within this period there are two dominating news topics that we use as a basis for our experiments. Those are the riots in Libya and the consequences of the earthquake and tsunami catastrophe in Japan. Furthermore we will use one topic consisting of only stop words as a negative example with respect to the Japan and Libya topics, to test the performance of the similarity measures. In order to evaluate the different similarity measures we fit different topic models (one for each day) with comparable results. Since we also created a single corpus for each of the consecutive days we handpicked topics from the models. These topics are illustrated in Table 1 where we sorted the words by their probability and chose the 20 most probable words to summarize the contexts.

Based on these hand selected topics we built similarity matrices comparing the similarities of all topics using the similarity measures described in section 5. Additionally we tested the similarity measures on different word subsets of the topics. This means that we set all the probabilities within a topic distribution $p(\mathbf{w}|z_k)$ to 0 except those for the most probable n words. In our setup we chose $n \in \{2,5,10,20,40,80,160,320\}$. From the intuition that the most probable words sufficiently define a topic's context, we expect a more unambiguous and robust similarity matrix for comparisons based on small n. To decide how robust the similarities are, we measure the absolute deviation between the true and the desired similarity for each entry in a similarity matrix for a specific word sub-set. We average this value over all similarities for each topic. The mean absolute deviation for this setting is defined as

$$MD = \frac{1}{N_{topics}^2} \sum_{i=1}^{N_{topics}} \sum_{j=1}^{N_{topics}} \|s_{ij} - s_{ij}^*\| \quad (6)$$

where N_{topics} is the number of topics included in the similarity matrix, s_{ij} is the measured similarity and s_{ij}^* is the desired similarity. If two topics match, the desired similarity s_{ij}^* is equal to 1 whereas in contrast to that, the desired similarity for non-matching topics will be set to 0. If the intuition that the n most probable words sufficiently define the topic context/meaning is correct, incorporating only semantically relevant words into the comparison results in a decrease of the mean absolute deviation. To measure this behavior we calculate the mean absolute deviation of all elements within a similarity matrix for all defined values of n. To select the optimal similarity measure in combination with the optimal sub-set of words, we will determine the combination for which the mean absolute deviation has a minimum.

Note that the selection of the optimal sub-set of words needs to be rechecked for new tasks in new text sources since the probability distributions, and thus the number of meaningful words of the topics, strongly depend on those preferences.

Results

Performing the experiments with the procedure described above gives 27 similarity matrices.[6] For each matrix we calculated the mean absolute deviation of its entries. Figure 2 shows the performance of the different similarity measures. The x-axis represents the number of the most probable words used whereas the y-axis corresponds to the mean absolute deviation. Cosine similarity quite surprisingly yields the best results, i.e. the lowest mean absolute deviation for a sub-set of 10-40 words. A minimum of the mean absolute deviation of similarity values means that we have a higher tolerance to set a threshold. Similarities are close to their desired values and similarity values of positive and negative matches are spread over a wider range. Also, the intuition is verified that the spreading between the similarity values, and hence the distinguishability, rises when we exclude words from the comparison that scatter their probability mass over a large number of other topics: Incorporating all words of a topic's word distribution into the comparison always results in a certain amount of similarity among topics in a corpus. This is caused by the fact, that many words (belonging to the long tail of low probability words in a topic's word distribution) spread their probability mass across all topics in the corpus, i.e. they belong to the long tail of all other topics as well. Obviously, this provokes similarity to some degree, even if topics are not related at all. Thus, taking away low probability words results in higher similarity of topics that effectively mean the same.

[6] We compare three similarity measures. For each measure we built nine different similarity matrices based on the comparison of the topics with only the top n words left.

Table 1: Selected Topics from consecutive days 10th -15th March 2011.

Date	Shortname	Top 20 Words
12-03-2011	japan1	Japan nuclear plant tsunami earthquake reactor power Japanese disaster radiation water damage quake plants country Tokyo explosion reactors Fukushima reports
13-03-2011	japan2	nuclear Japan tsunami power earthquake reactor Japanese water disaster plant radiation crisis plants magnitude fuel reactors aftershocks rescue Friday prefecture explosion
14-03-2011	japan3	nuclear Japan reactor power plant Japanese earthquake tsunami explosion disaster Tokyo rescue reactors energy plants crisis radiation JST safety water
15-03-2011	japan4	nuclear Japan plant power radiation Japanese reactor reactors fuel earthquake levels Tokyo water disaster tsunami fire level crisis agency safety
10-03-2011	libya1	Libya Gaddafi forces military zone no-fly Nato Libyan Libyan oil foreign rebels rebel council Ras_Lanuf France fighting regime defence country
12-03-2011	libya2	Gaddafi Benghazi MP country regime revolution revolutionary Libya forces GG international council countries intervention foreign eurozone Libyan no-fly city army
13-03-2011	libya3	Gaddafi Libya oil foreign Arab Europe intervention no-fly Iraq zone support military forces regime rebels security western uprising Egypt Tunisia
14-03-2011	libya4	Cameron Labour Libya zone Gaddafi no-fly Miliband Balls Britain vote tax campaign action plan party Clegg ministers Labour rebels referendum
15-03-2011	libya5	no-fly zone Bahrain forces Gaddafi military Libya troops security rebels foreign torture regime Benghazi told Saudi_Arabia Britain France G8 town
15-03-2011	stopwords1	years public make work pay world made good UK back part long ve don day Germany week big report

As we stated before, these properties can vary for different text sources and tasks. Since other models need to fulfill different requirements for other content analysis tasks, they are often run with different sets of parameters or other preconditions. Hence, the proposed procedure needs to be reproduced for other text sources and/or models in order to select the optimal size of word sub-sets. However, cosine similarity definitely yields best results in the context of our matching process.

Applications and Future Work

In this paper, we presented a method to match the outcome of different topic models on the basis of the word distributions $p(\mathbf{w}|\mathbf{z})$. With this setting it is possible to train topic models on little chunks of text data and match the outcomes afterwards. An application for this is the generation of a topic models per hour, day, month or year where we can match the outcomes easily. With this on hand we can track and detect topics within diachronic news, patent or social media text sources. Furthermore we can handle very large datasets by dividing the text sources into document sub-sets and distributing the model training to many machines. Afterwards we can

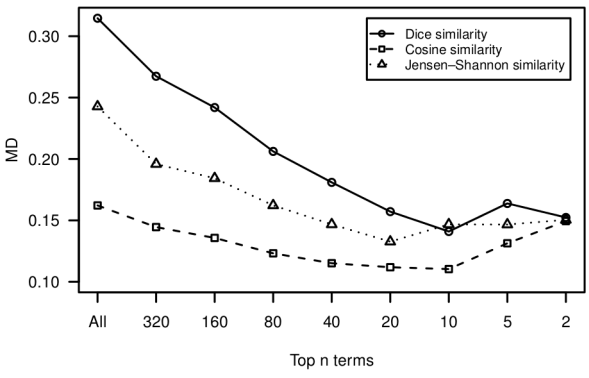

Figure 2: Mean absolute deviation for sub-sets of words

match the outcomes and give an accumulated view onto the whole corpus.

Future work will be focused on the selection of a threshold for different text sources and the definition of word sub-sets to use. Because of the diverse properties of certain text sources, specifying a general threshold for matching the topics proved to be inappropriate. For every text source, precision and recall of topic matching have to be optimized separately. To address this we will establish a procedure to test specific text sources for an optimal threshold. In (Silva, Stasiu, Orengo, & Heuser, 2007) a promising approach is shown, which can be adopted to this problem. Using this work it is possible to address the topic tracking problem with a mixture of lightweight similarity measures and simple fast processable topic models. With the connection of similar topics, time series data of consecutive chunks of text data e.g. consecutive days can be built, which can then be further analyzed to detect trends, unusual behavior or seasonal effects.

References

Allan, J. (2002). Introduction to topic detection and tracking. In J. Allan & W. B. Croft (Eds.), *Topic detection and tracking* (Vol. 12, p. 1-16). Springer US.

AlSumait, L., Barbará, D., & Domeniconi, C. (2008). Online lda: adaptive topic models for mining text streams with applications to topic detection and tracking. *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, 3–12.

Bishop, C. M. (1999). Latent variable models. In M. I. Jordan (Ed.), *Learning in graphical models* (pp. 371–403). MIT Press.

Blei, D., & Lafferty, J. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning* (pp. 113–120).

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

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22* (pp. 288–296).

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(Suppl 1), 5228–5235.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge Mass.: MIT Press.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010, June). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the north american chapter of the acl* (pp. 100–108). Los Angeles, California: Association for Computational Linguistics.

Silva, R. da, Stasiu, R. K., Orengo, V. M., & Heuser, C. A. (2007). Measuring quality of similarity functions in approximate data matching. *J. Informetrics*, *1*(1), 35–46.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*.

Wang, X., & McCallum, A. (2006). Topics over time: a non-markov continuous-time model of topical trends. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 424–433.

Appendix: Example similarity matrices

Figures 3 and 4 show the difference between an unassertive and a confident similarity matrix. A similarity of one corresponds to white, zero similarity is drawn in black. Note that we have a small amount of similarity between all topic pairings if we include all words for a match.

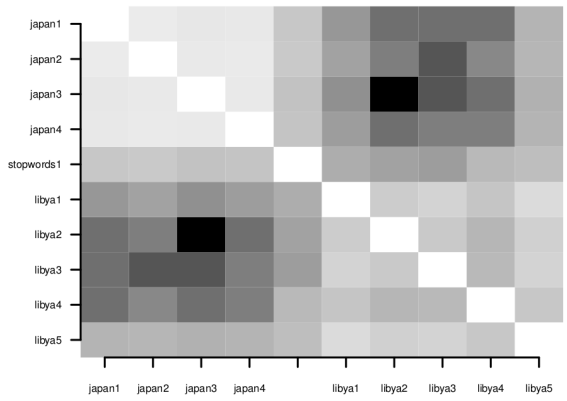

Figure 3: Similarity matrix with all words based on Jenson-Shannon divergence. Not only matching topics exhibit similarity.

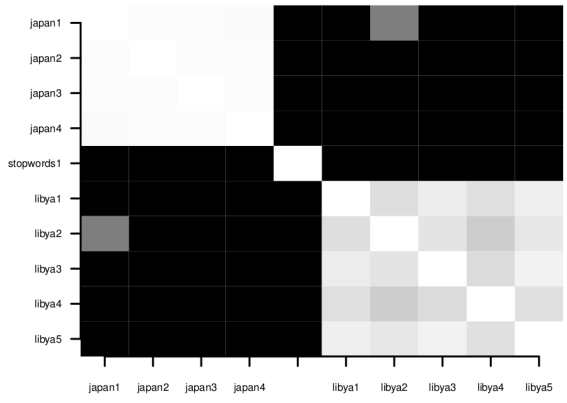

Figure 4: Similarity matrix with matching of 20 most probable words based on cosine similarity. High similarity is given to matching topics only.