The Influence of Input Matrix Representation on Topic Modelling Performance

Alta de Waal
Human Language Technology Competence Area
CSIR Meraka Institute
Pretoria, SOUTH AFRICA
Email: adewaal@csir.co.za

Etienne Barnard
Multilingual Speech Technologies Group
North West University
Vanderbijlpark 1900, SOUTH AFRICA
Email: etienne.barnard@gmail.com

***Abstract*—Topic models explain a collection of documents with a small set of distributions over terms. These distributions over terms define the topics. Topic models ignore the structure of documents and use a bag-of-words approach which relies solely on the frequency of words in the corpus.**

We challenge the bag-of-word assumption and propose a method to structure single words into concepts. In this way, the inherent meaning of the feature space is enriched by more descriptive concepts rather than single words. We turn to the field of natural language processing to find processes to structure words into concepts.

In order to compare the performance of structured features with the bag-of-words approach, we sketch an evaluation framework that accommodates different feature dimension sizes. This is in contrast with existing methods such as perplexity, which depend on the size of the vocabulary modelled and can therefore not be used to compare models which use different input feature sets. We use a stability-based validation index to measure a model's ability to replicate similar solutions of independent data sets generated from the same probabilistic source. Stability-based validation acts more consistently across feature dimensions than perplexity or information-theoretic measures.

I. Introduction

Vast amounts of electronic data are available, including news articles, scientific articles, newsgroup entries, emails and social network artifacts. The size of these data sets grow every day, making it increasingly difficult to make sense, and extract useful information from such information sources. The data sets are typically unstructured, unlabelled and dynamic in nature. This has stimulated the development of novel processing techniques in order to extract, summarise and understand the information contained therein. In many text mining applications, no or little prior knowledge is available about the content of the text data [1] which calls for unsupervised techniques with the goal of structuring and associating related text sources.

Topic modelling is a technique for the unsupervised analysis of large document collections. The fundamental assumption of topic models is that the semantic context of a document is a mixture of topics [2]. The topics are shared across the corpus by various documents and a topic is defined as a distribution over the vocabulary set of the document collection. Topic models infer document-topic associations, or clusters. These clusters are probabilistic in nature - each document exhibits a probability of being assigned to a topic.

The quality of the latent topic space is important for two reasons: Firstly, it associates unseen documents with existing documents and predicts latent similarities, thereby exhibiting its *predictive* abilities. Secondly, it summarises the corpus with a set of topics, thereby exhibiting its *exploratory* abilities. When measuring the predictive abilities of a topic model, perplexity is an appropriate measure. It provides an indication of the model's ability to generalise by measuring the exponent of the mean log-likelihood of words in a held-out test set of the corpus. The exploratory abilities of the latent topic space are generally measured by means of human interpretation. This is done by examining the top-*n* words in a topic and (rather subjectively) assigning a label to the topic.

In this paper, we discuss techniques to improve both the predictive and exploratory abilities of topic models. In particular, we discuss the properties of the input *document* $\times$ *word* matrix that contribute to the quality of inferred topics. We introduce an evaluation framework to measure this quality as the standard measure, perplexity, becomes inappropriate when the dimensions of the *document* $\times$ *word* matrix change.

II. Related Work

The vocabulary of a text corpus defines the parameter space of a topic model. The accurate representation of the corpus through topics (and therefore the value of topic models) is challenged by this high dimensional, data sparse parameter space [3]. Strategies to address this issue have been developed, such as vocabulary reduction. In fact, Rigoste *et al.* [3] have indicated a significant increase in performance of topic models when reducing the vocabulary size significantly (900 out of 40,000). Only frequent words were kept in the data set, discarding rare words.

In the field of text categorization, feature selection methods such as mutual information are often used to reduce the vocabulary size in order to increase model performance [4]. [5] used the information bottleneck method [6] to extract words capturing most information about a document. The full vocabulary is then replaced with the word clusters. The information bottleneck method was then applied again on this compact representation of document information in order to create document clusters. In this way, the original high dimensional vocabulary space is reduced significantly and thereby

increasing the performance of the algorithm. The information bottleneck method differs from probabilistic topic models in the sense that it makes no statistical assumption about the structure of the data (no hidden variables are defined).

The relaxation of the bag-of-words assumption most often used for topic models provides a wealth of opportunity for better interpretation of topic models. Word order is very important for lexical meaning [7].

Although *n*-gram approaches provide more contextual information, they come with a high price in computational complexity [8], [7]. To address this problem, [7] extend the bag-of-word assumption and introduce topical *N*-gram models: In the generative process, a topic is sampled for each word and then the words status as unigram or *n*-gram is determined based on context. The model then samples the word from a topic-specific unigram or *n*-gram distribution. The statistical simplicity of models based on the bag-of-words assumption is lost in this approach and although the *n*-gram output produces better interpretation of the topics, it is not clear whether or not it performs better than topic models based on the bag-of-word assumption.

III. Stability-based validation for Topic Models

A. Introduction

In this section, we introduce an evaluation framework for topic models. In order to understand if the structuring of features will contribute to the quality of inferred topics, alternative evaluation methods to perplexity need to be considered.

The valuation of topic models is a challenge because (a) topic models are often applied to unlabelled data, so that a ground truth does not exist and (b) "soft" (probabilistic) document clusters are created by state-of-the-art topic models, which complicates comparisons even when ground truth labels are available. In general, unsupervised techniques do not allow for comparison of predicted outcomes with ground truth outcomes; therefore, traditional classification performance metrics cannot be used. Hence, indirect measures of generalization, such as perplexity, are commonly employed as performance measures for topic models. Perplexity comes in handy for model selection purposes and can measure the relative performance between different topic models and the number of topics. It depends on the size of the vocabulary modelled – it can therefore not be used to compare models which use different input feature sets or across different languages.

We turn to cluster validity techniques in the data clustering field to search for alternative performance metrics for topic models. Clustering algorithms aim to extract the natural grouping structure in data [9]. Data clustering algorithms include *k*-means [10], *k*-nearest neighbour [11] and self-organising feature - or Kohonen maps [12], a type of artificial neural network. Cluster validity needs to consider various issues related to clustering algorithms. A cluster algorithm will cluster data, even when no natural clusters in the data exist. Different cluster algorithms may produce different clusters, which raises the question whether the resulting clusters are a true reflection of the data or imposed by the particular algorithm [12]. A perturbation measure of some sort is usually implemented in a cluster validation scheme in order to validate the clustering solutions [12] and should evaluate the output of a clustering algorithm quantitatively and objectively. Furthermore, the validation scheme should be applicable to all clustering algorithms - it should not rely on assumptions about specific group structures in the data that is not captured by the clustering algorithm itself [9].

B. Stability-based validation

The basic idea of stability-based validation is to compare clustering solutions for two different data sets generated from the same probabilistic source. This assesses the replicability of the clustering solution. Because the two data sets are mutually exclusive, the derived clustering solutions are not directly comparable and the clustering solution of the first data set needs to be transferred to the clustering solution of the second data set by means of a classifier. We use an SVM (support vector machine) as the classifier to transfer the solution.

The process can be explained as follows:

Let $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{X}' = (X'_1, \ldots, X'_m)$ be finite data sets. A clustering algorithm $\mathcal{A}$ constructs a solution $\mathbf{Y} := \mathcal{A}(\mathbf{X})$ where each sample X_i is associated with a label Y_i. [9] proposed the following mechanism to make a direct comparison between solutions possible: The data set $\mathbf{X}$ together with its clustering solution $\mathbf{Y} := \mathcal{A}(\mathbf{X})$ can be considered as a training set used to infer a classifier, ϕ. The classifier ϕ is now used to predict a label $\phi(X')$ for a new data point X' in a test set $\mathbf{X}'$ - the second data set. The predicted labels $\phi(\mathbf{X}')$ are subsequently compared with the labels generated from the clustering solution on the second data set: $\mathbf{Y} := \mathcal{A}(\mathbf{X}')$. In this way, the solution of the first data set is transferred to the solution of the second data set, using the classifier ϕ [9].

C. Stability Measure

The normalised Hamming distance was proposed in [9] to quantify the fraction of labelled entities $\phi(\mathbf{X}')$ that were misclassified (not matching the labels of $\mathcal{A}(\mathbf{X}')$). This gives a good indication of the match between the cluster solutions of the two data sets in an intuitive way. It is called the stability or dissimilarity measure and is the average distance between solutions for two data sets $\mathbf{X}$ and $\mathbf{X}'$. This approach to derive a stability index is optimised for hard clustering solutions, i.e. the clustering solution $\mathbf{Y} := \mathcal{A}(\mathbf{X})$ is used as labelling information for the data set $\mathbf{X}$ in order to create a training set. Furthermore, the use of the Hamming distance as dissimilarity measure calls for a one-to-one comparison of the predicted label and cluster solution.

In the case of topic modelling we use the average document correlation of aligned topics in the two clustering solutions for data set $\mathbf{X}'$ as the stability index,

$$S(\mathcal{A}) := E_{\mathbf{X},\mathbf{X}'}[\text{corr}(\theta, \theta')] \tag{1}$$

The stability index as defined above is used throughout the paper to describe the topic model performance. Figure 1

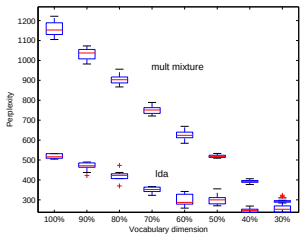

Fig. 1. Perplexity vs feature dimensionality (CRAN Corpus)

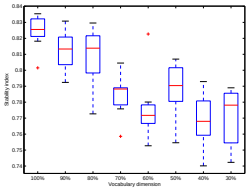

Fig. 2. Stability index vs feature dimensionality (CRAN Corpus)

illustrates the perplexity measures accross vocabulary dimensions for two topic models, namely LDA and Multinomial Mixture. Figure 2 illustrates the stability index measures accross vocabulary dimensions for the LDA topic model. The stability index measures act more consistently across feature (vocabulary) dimensions than perplexity. Both experiments were done using the CRAN corpus (explained in more detail in section V).

IV. Structured Features

A. Introduction

The input data to topic models are contained in a *document* $\times$ *word* matrix and the output data in *document* $\times$ *topic* and *topic* $\times$ *word* matrices.

The bag-of-words assumption is the core assumption of most topic models such as LDA and multinomial mixture. The bag-of-words approach turns natural text in multiple documents into a *word* $\times$ *document* matrix where $cell_{ij}$ represents the frequency of $word_i$ in $document_j$.. The advantage of the bag-of-word approach is that it simplifies the computational process of the topic model significantly because of the independence assumption. The limitation of the bag-of-words approach is that significant phrases get lost in the use of single terms, because critical word order and phrases are not captured.

B. Chunking

In this section we investigate the structuring of sets of words into concepts that will still maintain the bag-of-words assumption, but also introduce other benefits that cannot be obtained with single terms. The input data to the topic model remain a *document* $\times$ '*unit*' matrix where *unit* represents a concept of joint words. We turn to the field of natural language processing (NLP) with the aim of reducing feature space dimensionality. We study different syntactic strategies to group adjacent words into concepts. At the core of this approach is part-of-speech (POS) tagging of words in the corpus: A sequence of non-overlapping words are grouped based on their POS tags, forming a concept.

Two NLP tasks are the search for structure and meaning in streams of text. The most common methods to perform these tasks are segmentation and labelling. Segmentation comprises breaking up a stream of characters into 'linguistically meaningful segments' like words. These segments are then labelled with their respective part-of-speech categories. The search for structure and meaning can be construed as a combination of segmentation and labelling. The segmentation is based on a non-overlapping sequence of words that makes syntactically sense. These segments are named 'chunks'.

The first step in the chunking process is to label, or tag words. For the purpose of tagging, we assume words and punctuation markers to be the tokens in streams of text. Tagging is the assignment of part-of-speech labels to each token in the corpus.

We use a simple bigram tagger, trained on the Penn Treebank Corpus to classify the words into part-of-speech tags.We combine the bigram tagger with a unigram tagger as well as a default tagger as backoff algorithm to fall back on if the bigram tagger fails to tag the word. The default tagger tags a word as a noun by default. This tagger combination achieves an accuracy of 88.56% on the Penn Treebank corpus.

The next step in the chunking process is to segment the tagged words into meaningful, non-overlapping phrases. Many different phrases can be defined to be chunked, such as verb phrases, noun phrases and even more specifically, proper noun phrases. For the purpose of structuring features for topic models, we are interested in noun and verb phrases and different patterns thereof. These patterns are defined according to a chunk grammar that consists of rules on how sentences should be chunked [13].

C. Chunking processes for topic models

One of the objectives of chunking for topic modelling is to reduce the dimensionality of the feature space. Furthermore, it should improve the intelligibility of the topics. In a sense these two objectives cause ambivalence when designing a segmentation pattern. On the one hand, we want the chunk to be as exhaustive as possible (to improve intelligibility) but on the other hand the chunk should be as generic as possible in order to cover as many as possible occurrences over all documents and hence, reduce the dimensionality of the parameter space.

1) Noun phrases: Our first chunking process is to include only noun phrases in the feature set with regular expression <**NN.***>+. This will include any number of adjacent nouns of any kind.

2) Noun and verb phrases: This chunking process is made up with two patterns: The first pattern includes any number of adjacent nouns of any kind - <**NN.***>+. The second pattern includes any number of adjacent verbs of any kind - <**VB.***>+.

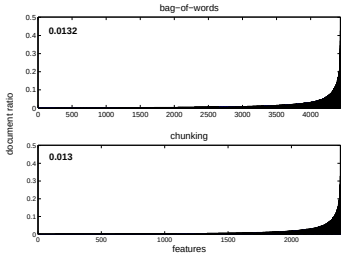

Fig. 3. Ratio of documents containing one or more occurrences of feature (ordered)

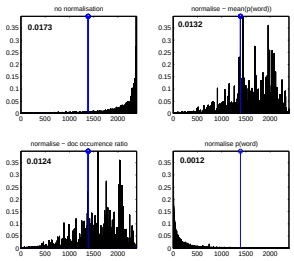

Fig. 4. Illustration of document × chunk matrix

3) Verb and noun with adjectives phrases: This chunking process is made up with two patterns: The first pattern includes any number of adjacent verbs of any kind - <**VB.***>+. The second pattern is made up with zero or more adjectives, followed by one or more nouns - <**JJ.***>*<**NN.***>+.

D. Including significant chunks in the data set

As mentioned before, the bag-of-words assumption contributes to the statistical simplicity of topic models such as LDA. The features, or words, provide information to the topic model about the way in which documents were generated. Furthermore, it discriminates between documents and allocates documents to topics. Attributes of words that will do this effectively are the following:

- A high variance in occurrence of the word across documents. This excludes words with a consistent high count, such as stop words, or a consistent low count across documents, such as foreign words.
- At least two occurrences of the word in the corpus, otherwise it has no useful statistical properties.

One measure of a high variance of words, or features across documents is the ratio of documents that contain one or more occurrences of the specific feature. Figure 3 illustrates the (sorted) ratio of documents (y axis) containing occurrences of the word, or features as represented on the x axis. A ratio of 0.5 means that 50% of documents in the corpus have at least one occurrence of the feature. The figure was generated for the CRAN corpus (this corpus will be discussed in 'Experimental Evaluation') where the top graph represents bag-of-words features and the lower graph represents the feature set generated by the '<NN.*>+' chunking strategy. The graphs indicate that both bag-of-words and chunking produce a small number of features with high representation in documents - most features have a small document occurrence ratio. A lower document occurrence ratio implies a higher variance in occurrence of the feature across documents. By inspection it is clear that the chunking feature set follows the same document occurrence pattern as the bag-of-words feature set. The mean document occurrence ratio is 0.0132 and 0.013 for the bag-of-words and chunking feature sets respectively.

Although the structuring of chunks lowers the dimensionality of the feature space, it does not guarantee an improvement on the above mentioned attributes as can be seen both from figure 3 and the mean document occurrence ratio also given in figure 3. In fact, the frequency of a chunk in a document is equal or lower than the lowest frequency word in the chunk. This implies that unfiltered use of the *document* × *chunk* matrix will not improve the performance of the topic model as the feature set resulting from the chunking strategy is not richer in variance across documents than the original bag-of-words feature set. In fact, it is easy to see that a larger proportion of chunks than words occur only once or twice in the data set. The chunk set needs to be filtered first in order to reflect the desired attributes by including chunks that have a high variance across documents.

We calculate a variance measure to order chunks as follows: For each document, we calculate the probability of each chunk. For each chunk, we then calculate the variance across documents and experiment with different ways to normalise the variance. A new feature set only includes chunks with a high normalized variance. The average document occurrence ratio of this feature set should be lower than that of the bag-of-words feature set.

We experiment with the following strategies to normalize the variance:

- No normalization.
- Normalize with the average chunk probability across documents.
  $\sigma^2(\text{chunk}) = \sigma^2(\text{chunk}) / mean(p(\text{chunk}))$
- Normalize with the document occurrence ratio.
  $\sigma^2(\text{chunk}) = \sigma^2(\text{chunk}) / ratio(\text{document occurrence})$
- Normalize the chunk probability across documents before calculating the variance.
  $\sigma^2(\text{chunk}) = \sigma^2(\text{norm(p(chunk))})$

We ordered the features in ascending order of variance across documents and plotted the corresponding document occurrence ratio for different normalization strategies in figure 4. Once the chunk set is ordered using the normalized variance measure, the top *n* chunks in terms of variance are used as a new feature set to form a *document* × *chunk* matrix. All chunks on the right of the blue line are included in the new feature set. These chunks have a high variance across documents. The graphs display the document occurrence ratio of the chunks. The first graph (upper left quadrant) indicates that no normalization of the variance will select chunks with a high document occurrence ratio. The next two graphs represent

a mixture of high and low document occurrence ratios included in the filtered chunks set and the last graph (lower right quadrant) represent inclusion of low document occurrence ratios. The average document occurrence scores for the four normalization strategies are indicated in the upper left corner of each graph. Although the graph in the lower right quadrant produced the lowest average document occurrence ratio, many documents are left empty when using this feature set, which does not make it a feasible feature set for topic modelling. In the next section, we calculate the stability index for the three normalisation strategies and compare it with the bag-of-words and complete chunk feature sets.

V. Experimental Evaluation

Our experiments are based on two corpora:

- The Cranfield collection (CRAN) of aerodynamic abstracts has 1397 documents and a vocabulary of size 4437.
- A subset of the Reuters-21578, distribution 1.0 newswire articles (Reuters) is used, containing 6600 documents with 15822 unique terms.

For all experiments, the Latent Dirichlet Allocation (LDA) [14] was used as topic model. For experimental evaluation, we calculate the stability index on the *document* $\times$ *chunk* matrix for each chunking strategy and each normalisation strategy, both for the CRAN and Reuters corpora. Each experiment was repeated ten times, using different initial conditions of the model parameters with each iteration. The data sets are split into 80% train and 20% test sets. The number of topics for each experiment is set to 25, both for the CRAN and Reuters corpus.

A. Results

The rows in tables I and II, display the results on the following matrices:

- Bag-of-words: *document* $\times$ *word* matrix
- All chunks: *document* $\times$ *chunk* matrix
- No normalization: Subset of chunks[1] with high variances, no normalization performed.
- $mean(p(\text{chunk}))$: Subset of chunks with high variances, normalized with $mean(p(\text{chunk}))$.
- $ratio(\text{documents containing chunk})$: Subset of chunks with high variances, normalized with $ratio(\text{document occurrence ratio})$.

Some interesting topics are displayed in tables III and IV for the CRAN and Reuters corpora, respectively. The topics are represented by the top 10 phrases in the selected topic distribution. These topics are inferred from the chunking process: verb and noun with adjectives phrases. The phrases are clearly more intelligible than only single word phrases in many cases, thus demonstrating the qualitative advantage of the proposed method.

[1]For the CRAN corpus, each subset of chunks includes the top 1000 chunks with the highest variability across documents. For the Reuters corpus, each subset of chunks includes the top 30% with the highest variability across documents.

As can be seen from the results, the best stability index is achieved with a subset of chunks where the variance is normalised with the document occurrence ratio. The chunking strategy <JJ.*>*<NN.*>+, <VB.*>+ achieves the best results, and the differences observed are highly significant statistically.

VI. Conclusion

In this paper, we discuss the structuring of features to be included in the input matrix for a topic model. We also introduce an evaluation measure that measures performance more consistently than existing measures such as perplexity. The performance of topic model outputs can be measured in qualitative and quantitative terms. Qualitatitve measures relate to the intelligibility of the inferred topics: topics are described by vocabulary distribution with the top-n words being a good indication of what the topic is about. Quantitative measures reflect the model's ability to reproduce the results given different initialisation conditions. We show that the structuring of features contributes to both the qualitative and quantitative performance of a topic model.

References

[1] D. Newman, C. Chemudugunta, P. Smyth, and M. Steyvers, "Analysing Entities and Topics in News Articles Using Statistical Topic Models," in *LNCS-IEEE Conference on Intelligence and Security Informatics*, San Diego, USA, 2006, pp. 93–104.

[2] T. Griffiths, M. Steyvers, and J. Tenenbaum, "Topics in Semantic Representation," *Psychological Review*, vol. 114, no. 2, pp. 211–244, 2007.

[3] L. Rigouste, O. Cappé, and F. Yvon, "Inference and Evaluation of the Multinomial Mixture Model for Text Clustering," *Inf. Process. Manage.*, vol. 43, no. 5, pp. 1260–1280, 2007.

[4] S. T. Dumais, "Using SVMs for Text Categorization," *IEEE Intelligent Systems Magazine, Trends and Controversies*, vol. 13, no. 4, pp. 21–23, 1998.

[5] N. Slonim and N. Tishby, "Document Clustering using Word Clusters via the Information Bottleneck Method," in *SIGIR '00: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2000, pp. 208–215.

[6] N. Tishby, F. Pereira, and W. Bialek, "The Information Bottleneck Method," in *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, 1999, pp. 368–377.

[7] X. Wang, A. McCallum, and X. Wei, "Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval," in *ICDM '07: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 697–702.

[8] H. M. Wallach, "Topic Modelling: Beyond Bag-of-Words," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[9] T. Lange, V. Roth, M. L. Braun, and J. M. Buhmann, "Stability-based Validation of Clustering Solutions," *Neural Comput.*, vol. 16, no. 6, pp. 1299–1323, 2004.

[10] J. B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, 1967, pp. 281–297.

[11] B. V. Dasarathy, *Nearest neighbor (NN) norms: NN Pattern Classification Techniques*. Los Alamitos: IEEE Computer Society Press, 1990, 1990.

[12] A. Webb, *Statistical Pattern Recognition*. John Wiley and Sons, Ltd., 2002.

[13] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, 2009.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

TABLE I
AVERAGE STABILITY INDEX - CRAN CORPUS

	<NN.*>+		<NN.*>+<VB.*>+		<JJ.*>*<NN.*>+, <VB.*>+	
	Average	Variance	Average	Variance	Average	Variance
Bag-of-words	0.79	3.8×10^{-05}	0.79	3.8×10^{-05}	0.79	3.8×10^{-05}
All chunks	0.7	0.0001	0.65	6.6×10^{-05}	0.60	0.0004
No normalization	0.75	4.1×10^{-05}	0.76	4.1×10^{-05}	0.73	4.1×10^{-05}
$mean(p$(chunk))	0.81	0.0001	0.81	6.6×10^{-05}	0.86	7.5×10^{-05}
$ratio$(documents containing chunk)	0.82	9.1×10^{-05}	0.83	5.8×10^{-05}	0.87	3.5×10^{-05}

TABLE II
AVERAGE STABILITY INDEX - REUTERS CORPUS

	<NN.*>+		<NN.*>+<VB.*>+		<JJ.*>*<NN.*>+, <VB.*>+	
	Average	Variance	Average	Variance	Average	Variance
Bag-of-words	0.75	2.3×10^{-05}	0.75	2.3×10^{-05}	0.75	2.3×10^{-05}
All chunks	0.69	1.7×10^{-05}	0.69	7.4×10^{-06}	0.62	1.6×10^{-05}
No normalization			0.64	1.6×10^{-05}	0.69	1.6×10^{-05}
$mean(p$(chunk))	0.78	1.8×10^{-05}	0.83	1.9×10^{-05}	0.81	1.9×10^{-05}
$ratio$(documents containing chunk)	0.52	0.0009	0.83	5.3×10^{-05}	0.85	8.3×10^{-06}

TABLE III
SOME INTERESTING TOPICS: CRAN CORPUS, <JJ.*>*<NN.*>+, <VB.*>+

Topic 6	Topic 44	Topic 94
jet	method	vortex
thrust	integral equation	wake
jet speed	digital computer	growth
nose jet	approximate treatment	free shear layer
adjustment	boundary layer body revolution	constant velocity
interaction	circular cylinder	exerted
shallow shell analysis	field flow	basic equation
shock conditions	problem heat transfer	vortex cancellation
theory plastic	rod	relation
air	additional span	shockwave equation

TABLE IV
SOME INTERESTING TOPICS: REUTERS CORPUS, <JJ.*>*<NN.*>+, <VB.*>+

Topic 14	Topic 74	Topic 56
terms	year	production
letter	unemployment	produced
signed	ratio	year
disclosed	averaged	estimated
acquire	increased	increase
intent	unemployment rate	energy
definitive agreement	consumer price	cover
transaction	capital spending	agriculture ministry
approval	residual fuel demand	industrial production index base
financial corp	rate	lake