Evaluating Topic Models with Stability

Alta de Waal and Etienne Barnard

Human Language Technologies, Meraka Institute
Faculty of Engineering, North West University
adewaal@csir.co.za, ebarnard@csir.co.za

Abstract

Topic models are unsupervised techniques that extract likely topics from text corpora, by creating probabilistic word-topic and topic-document associations. Evaluation of topic models is a challenge because (a) topic models are often employed on 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. Perplexity has often been used as a performance measure, but can only be used for fixed vocabularies and feature sets. We turn to an alternative performance measure for topic models – topic stability – and compare its behaviour with perplexity when the vocabulary size is varied. We then evaluate two topic models, LDA and GaP, using topic stability. We also use labelled data to test topic stability on these two models, and show that topic stability has significant potential to evaluate topic models on both labelled and unlabelled corpora.

1. Introduction

The vast amount of electronic text available has stimulated the development of novel processing techniques in order to extract, summarise and understand the information contained therein. Topic modelling is a technique for extracting topics from a text collection by creating probabilistic word-topic and topic-document associations [1]. The most successful topic models are generative models,using the assumption that documents are generated from a mixture of latent topics. A variety of topic models with different generative assumptions about how the documents are generated have been proposed. The documents do not need labels, implying that topic modelling is an unsupervised technique [2]. 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. However, current measures suffer from a number of shortcomings. Perplexity, for example, 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. In this paper, we investigate an alternative, namely topic stability, which overcomes some of these deficiencies.

The objective of this study is threefold. First, we compare the behaviour of perplexity and topic stability as two alternative performance metrics for topic models. Secondly, we compare the performance of two topic models, namely Latent Dirichlet Allocation (LDA) and Gamma-Poisson (GaP), using topic stability. Finally, we investigate the relationship between stability and classification accuracy when labels are available. The rest of the paper is outlined as follows. First we put our work in context with the literature. Two topic models, LDA and GaP are described in section 3. Then, we give an overview of perplexity as well as the process to derive topic stability in sections 4 and 5. Two text corpora that we use in experimentation and data preprocessing are described in section 6. The experimental setup and results follow in section 6.1

2. Related Work

In this study we focus on evaluation techniques for unsupervised methods, specifically topic models. In the field of topic modelling, the majority of studies use perplexity as an evaluation method [1, 3, 4]. Rigouste further suggests [1] a document co-occurrence score that is not dependent on feature dimensionality reduction in the way that perplexity is. The document co-occurrence method demands an equal number of topics in two independent sets. The use of this method to evaluate unsupervised algorithms is described in detail in [5]. Information-based measures, such as relative information gain are also used to evaluate topic models, but are difficult to interpret [1, 6].

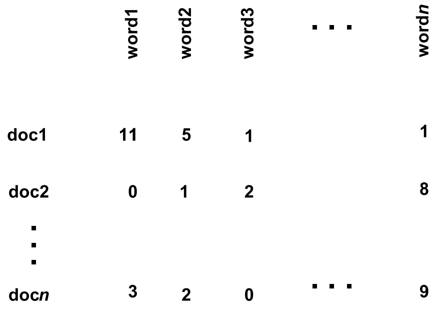
The concept of topic stability was introduced by Steyvers and Griffiths [2], where stability between aligned topics for two independent topic solutions is measured using the symmetrized Kullback Leibler (KL) distance between the two topic distributions. Classification of documents is another way to test the performance of topic models [3, 7]: the *document* $\times$ *topic* matrix is used as the feature matrix to classify the documents of a labelled corpus using a classifier such as a support vector machine. The topic model is thus measured in terms of the quality of features that it produces.

We focus on comparing perplexity and topic stability as evaluation methods for topic models. Our approach to measuring topic stability is a hybrid between the document co-occurrence of Rigouste and the topic stability of Steyvers and Griffiths. Instead of using the Kullback Leibler divergence between two topic distributions over words (Steyvers and Griffiths), or the document co-occurrence score (Rigouste), we calculate the document correlation between two aligned topics. This allows us to compute a stability measure which is somewhat insensitive to the specific words chosen to describe each topic.

3. Topic Models

For the purpose of topic modelling, a large matrix is constructed from a text corpus (consisting of a number of distinct documents), with rows representing the documents and columns representing the word frequencies (for words in the corpus vocabulary – see figure 1).

In this view, a document is represented as a high-



	word1	word2	word3	. . .	word*n*
doc1	11	5	1		1
doc2	0	1	2		8
. . .					
doc*n*	3	2	0	. . .	9

Figure 1: Document × Word Matrix

dimensional vector, containing the counts of each word in the document. This representation of a text corpus is widely used by a number of clustering techniques, where documents are associated based on their semantic or 'thematic' similarity [1]. 'Thematic' similarity or meaning is extracted by applying statistical computations on the large *document* × *word* matrix [8]. Many approaches to text clustering exist [1, 3, 7, 9], using different sets of assumptions on how the documents in a text corpus are generated. We focus on probabilistic approaches that result in probabilistic topic-document associations [1] by assuming a probabilistic generative process for documents. This section describes two popular topic models with different generative assumptions, namely Latent Dirichlet Allocation (LDA) and Gamma-Poisson (GaP).

3.1. Terminology and notation

We define the following terms and their associated notation:

- A *corpus* is a collection of *M* documents denoted by $\mathcal{C} = \{\, \mathbf{w}_1, \mathbf{w}_2\ , \ldots, \mathbf{w}_M\}$. The first dimension of the *document* × *word* matrix in figure 1 is of size *M*.
- A *word* w is the basic unit of discrete data.
- A *document* is a sequence or passage of *N* words denoted by $\mathbf{w}_d = \{\, w_1, w_2\ , \ldots, w_N\}$.
- A *vocabulary* is subset of unique words (denoted by w_l) in the text corpus and indexed by $\{1, \ldots, V\}$. The second dimension of the *document* × *word* matrix is of size *V*.
- We define *T* latent semantic components or *topics* to approximate the *document* × *word* matrix with $T \ll V$.
- The *bag-of-words* representation of a document is the matrix representation illustrated in Fig.1; it neglects word order and only stores the word counts in each document. The quantity $C_{w_i d}$ is the word count of word w_i in document d.

When relating this terminology to machine learning theory, a word is a feature, a bag is a data vector and a document is a sample [7].

3.2. Latent Dirichlet Allocation (LDA)

The basic idea of LDA is that a document is represented as a random mixture over latent topics and a topic is a distribution over words in the vocabulary. LDA assumes that the mixture of topics for a document originates from a Dirichlet distribution and assigns a Dirichlet prior to the mixture of topics for a document. The Dirichlet prior is chosen because of its conjugacy to the multinomial distribution, a property which is crucial in simplifying the statistical inference problem [1, 3]. LDA assumes the following generative process for documents in a corpus $\mathcal{C}$ [3]:

For each document $\mathbf{w} = 1, \ldots\ , M$

1. Choose $\theta \sim$ Dirichlet(α), θ and α are of dimension *T*.
2. For each word w_i in the document,
   (a) Choose a topic $z_i \sim$ Multinomial(θ).
   (b) Choose a word $w_i \sim$ Multinomial($\boldsymbol{\beta}_{\boldsymbol{z_i}}$). $\boldsymbol{\beta}$ is a *V* × *T* matrix.

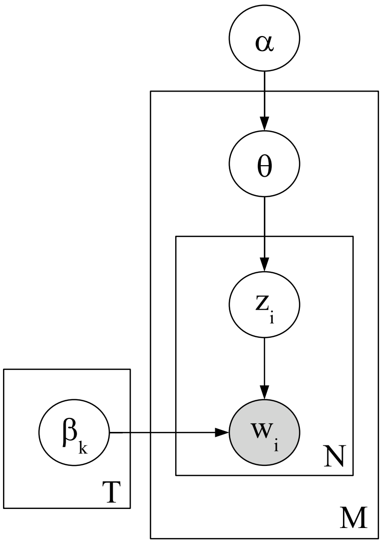


Figure 2: LDA graphical model

Topic models can be described graphically using directed graphs. In such a graphical model, variables are represented by *nodes*, dependencies between variables by *edges* and replications by *plates* [3]. Plates can be nested within one another. Observable nodes are shaded whereas latent variables are unshaded. In figure 2 the plate surrounding θ indicates that θ is a document level variable (with *M* replications) and the plate surrounding z and w indicates that they are word-level variables (with *N* replications). The plate surrounding β indicates that one topic must be chosen from *T* topics. The parameter β indicates which words are important for which topic and θ indicates which topics are important for a particular document [2].

3.3. Gamma-Poisson (GaP)

In [4], Canny introduces the Gamma-Poisson model (GaP), which uses a combination of Gamma and Poisson distributions to infer latent topics. It presents an approximate factorisation of the document × word matrix with matrices β and *X* (see figure 3). The word × topic matrix β represents the global topic information of the corpus $\mathcal{C}$ and each column β_k can be thought of as a probability distribution over the corpus vocabulary for a specific theme k. Each column $\mathbf{x}_d$ in the topic × document matrix *X* represents the topic weights for the document d. The Gamma distribution generates the topic weights vector $\mathbf{x}_d$ in each document independently. The Poisson distribution generates the vector of observed word counts **n** from expected counts **y**. The relation between $\mathbf{x}_d$ and **y** is a linear matrix $\mathbf{y} = \beta\mathbf{x}_d$. The topic weights $\mathbf{x}_d$ represent the topic content for each document and encodes the total length of passages about topic k in the document. GaP differs from LDA in this regard: LDA chooses topics independently per word in a document, according to the Dirichlet distribution [3], whereas GaP chooses words according to this topic weighting. GaP assumes the following generative process:

For each document $\mathbf{w}_d = 1, \ldots, M$

1. Choose $\mathbf{x}_d \sim$ Gamma(a,b)
2. For each word $w_i = 1, \ldots, N$
   (a) Generate $n_{w_i} \sim$ Poisson($\beta\mathbf{x}_d$)

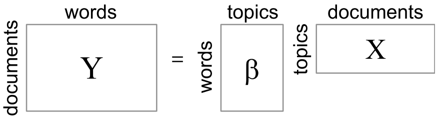


Figure 3: Matrix factorisation of *GaP*

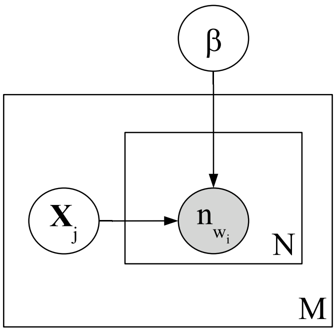


Figure 4: GaP graphical model

The Gamma distribution has two parameters: The first parameter a is called the shape parameter and the second parameter b is called the scale parameter. The mean value of $\mathbf{x}_k$ is $c_k = a_k b_k$ [4]. The plates in figure 4 further illustrate topics as passages of text in a document, as the $\mathbf{x}_d$ parameter does not reside in the N-plate.

4. Perplexity

Perplexity is a standard performance measure used to evaluate models of text data. It measures a model's ability to generalise and predict new documents: the perplexity is an indication of the number of equally likely words that can occur at an arbitrary position in a document. A lower perplexity therefore indicates better generalisation. We calculate perplexity on the test corpus $\mathcal{C}^*$ containing M^* documents as follows:

$$p(\mathcal{C}^*) = \exp\left\{ - \frac{\sum_{d=1}^{M^*} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M^*} N_d} \right\} \quad (1)$$

Perplexity is therefore the exponent of the mean log-likelihood of words in the test corpus. Consequently, it exhibits similar behaviour to log-likelihood: a reduction in feature dimensionality (in our case, vocabulary) reduces the perplexity, regardless of whether an improved fit to the data has been achieved [1]. This argument will be extended below.

5. Topic Stability

One of the key attributes of a useful topic model is that it should model corpus contents in a stable fashion. That is, useful topics are those that persist despite changes in input representation, model parametrization, etc. We therefore propose topic stability under such perturbations as an alternative performance indicator.

For probabilitstic models such as LDA and GaP, a natural perturbation method presents itself: since these models rely on the iterative optimization of a likelihood function from a random initial condition, they invariably converge to different local solutions from different starting points. We therefore measure stability as the document correlation between two topics that were generated in two independent algorithmic runs from different initial conditions.

In unsupervised learning, there is no way to order or lable topics prior to model estimation [2]. Thus, topics will in general be assigned to unrelated lables in separate runs. When the numbers of topics in the two algorithmic runs are the same, the Hungarian method (also known as Kuhn's method [10], [11]) can be used to align the topics. The Hungarian method is an algorithm for determining a complete weighted bipartite matching that minimises the distance between the two sets in the graph [11], [12]. First, a weight matrix must be set up to indicate the similarities of all pairs resulting from different runs; the algorithm then calculates the optimal overall matching between the two runs.
Two algorithm runs of a topic model can be represented in a bipartite graph (figure 5), where each set represents a run. Once a weight matrix is calculated for the graph, the best matched pairs can be calculated using the Hungarian method. Greedy matching is an alternative method that does not guarantee optimal matching [12].

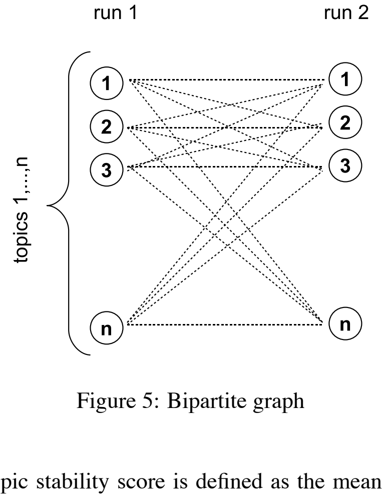


Figure 5: Bipartite graph

The topic stability score is defined as the mean document correlation over all topics, after topics have been aligned with the Hungarian method. The process of obtaining the topic stability scores is described in more detail in the following subsections.

5.1. Weighting

The first step in matching the bipartite graph is to obtain a weighting matrix that represents the weighting of all possible edges between topics in the two runs. Topic models result in two outputs, namely a *topic* × *document* matrix and a *word* × *topic* matrix. We use the *topic* × *document* matrix to calculate the weighting *l*.
Given two algorithmic runs, represented as sets *A* and *B* in a bipartite graph with an equal number of topics *k* in both sets, the weighting between two topics in the respective sets is calculated as follows:

$$l_{ab} = \sum_{d=1}^{M} P(\tau_a|\mathbf{w}_d)P(\Gamma_b|\mathbf{w}_d), \qquad (2)$$

where *M* is the number of documents in the corpus, $\mathbf{w}_d$ represents document *d*, and τ_a and Γ_b are the topic distributions from the respective sets *A* and *B*, over document *d*. In order to find the best matched pairs between *A* and *B*, the quantity $\sum_{i=1}^{T} l_{a_i b_i}$ is maximised.
Alternatively, the Kullback-Leibler divergence can be used as a weighting scheme [13].

5.2. Topic Alignment

The Hungarian method searches for the match with maximum weight, i.e., the set of edges that touches each topic in the two sets exactly once, so that $\sum_{i=1}^{T} l_{a_i b_i}$ is maximised [13].
Let $\mathcal{G}$ = (*A*,*B*;*E*) be a bipartite graph, with sets *A* and *B* as in figure 5.The algorithm starts with an empty matched set $\mathcal{M}$. Given the current matching $\mathcal{M}$, $D_{\mathcal{M}}$ is a directed graph where each edge *e* in $\mathcal{M}$ is oriented from *B* to *A* with length $\lambda_e = w_e$. Each edge *e* not in $\mathcal{M}$ is oriented from *A* to *B*, with length $\lambda_e = -w_e$. Let $A_{\mathcal{M}}$ and $B_{\mathcal{M}}$ be the set of topics in *A* and *B*, missed by $\mathcal{M}$. If there is an alternating path from $A_{\mathcal{M}}$ to $B_{\mathcal{M}}$, find the shortest one *P*, and replace $\mathcal{M}$ with the set difference of $\mathcal{M}$ and the edges of *P*. We iterate this process until no alternating path from $A_{\mathcal{M}}$ to $B_{\mathcal{M}}$ can be found.

5.3. Document Correlation

Once the topic alignment is completed, the correlation of documents between matching topics in the respective sets gives a good indication of the model stability. The document correlation is calculated using the *topic* × *document* matrix where each row represent the topic assignment to documents. Figures 7 and 8 are graphical representations of the document correlation between the topics from the first run and matching topics from the second run, for two different topic models. The dark diagonal line in figure 7 indicates a strong correlation between documents in matching topics.

6. Data Description and Experimental Setup

We used two text collections for the purpose of this research:

- The Cranfield collection [14] of aerodynamic abstracts has 1397 documents. The Cranfield (CRAN) collection is not labelled.
- The *20 Newsgroup* (NEWS) corpus, a large collection of approximately 20,000 newsgroup documents from 20 different newsgroups, collected by Kevin Lang [15]. Each document in this corpus is labelled according to its newsgroup. Cross-posts (duplicates) were removed from the corpus. Some of the newsgroups are closely related, whereas others cover completely unrelated domains.

As part of the data pre-processing step, all non-alphabetic characters were removed as well as words containing only consonants, or words with a sequence of three and more of the same alphabetic character. All words occurring only once were removed, and lastly, documents containing fewer than five words were also removed. From the NEWS corpus, email headings and group information were also removed. After the preprocessing step, the NEWS corpus contained 18705 documents with 52416 unique words and the CRAN corpus 1397 documents with a vocabulary of size 4437.
Both datasets were split into a 80% - 20% training and test set and words occurring only in the test set were ignored.

6.1. Experiments

6.1.1. Perplexity vs Document Correlation

As mentioned in section 4, perplexity as a performance metric is influenced by the feature dimensionality: it invariably improves with a reduction in input dimensionality, regardless of the quality of the fit obtained. To demonstrate this behaviour, we compare perplexity and document correlation against feature dimensionality. Using the CRAN corpus, we gradually reduce the vocabulary by randomly removing columns from the *word* × *topic* matrix. Thus, the number of vocabulary words is systematically reduced from 100% to 30%, keeping the number of documents the same. The document correlation was calculated on both the training and test set and perplexity was calculated on the test set.

Figure 6 displays the results. The lower graph represents the perplexity scores on the y-axis against the vocabulary dimension on the x-axis. The perplexity scores decrease (i.e. improve) every time dimensionality is reduced, even though there is no reason to believe that the random deletion of words will improve the topic model. The document correlation (upper graph) on the training and test set changes less dramatically, and the correlation on the test set becomes somewhat worse (lower) when words are removed, as would be expected.

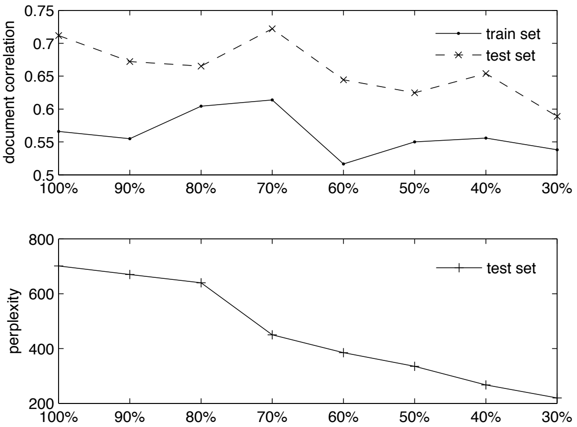


Figure 6: Perplexity vs Document Correlation

6.1.2. LDA vs GaP

In this set of experiments, we compare the performance of the two topic models, LDA and GaP, using document correlation.
In the first experiment, we conduct the straightforward document correlation method as described in section 5 on LDA and GaP, using the full CRAN and NEWS corpora. The resulting document correlation is displayed in table 1. It is clear from the results that LDA has a somewhat more stable topic assignment (indicated by the document correlation) than GaP. Figures 7 and 8 are graphical representations of the topic stability of the two respective models. The dark diagonal line in figure 7 indicates that the aligned topics generally have high document correlation. On the other hand, figure 8 has a less pronounced diagonal line, indicating more instability in topic assignment for the GaP model.

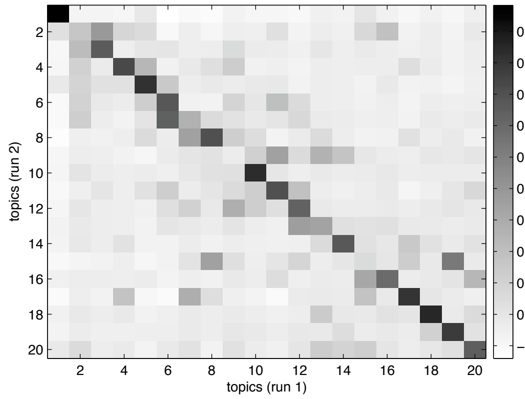


Figure 7: Document correlation matrix for 2 LDA topic solutions

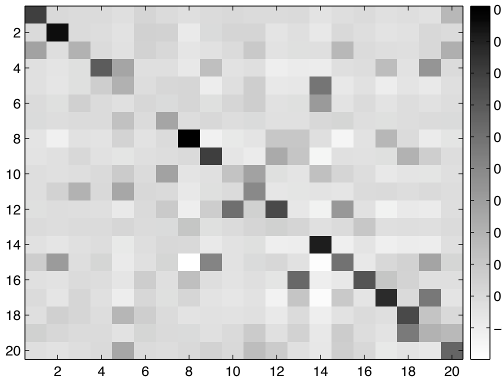


Figure 8: Document correlation matrix for 2 GaP topic solutions

Table 1: *Document correlation for two topic solutions*

	CRAN	NEWS
LDA	0.591	0.757
GAP	0.488	0.527

In the second experiment, instead of performing two independent executions of the algorithm, we run each algorithm once on the labelled NEWS data. We then use the document labels to populate the second set in the bipartite graph. Table 2 displays the results. Although neither LDA nor GaP result in a very good correlation between inferred topics and document labels, LDA has a slightly better correlation than GaP. The relatively low correlation values are not surprising, given that these algorithms make continuous-valued "soft" assignments between documents and topics, whereas the NEWS lables consist of binary assignments. It is encouraging to see that the stability and correlation results nevertheless agree in their preference for the LDA algorithm in this instance.

Table 2: *Document correlation for a topic solution and labelled data*

	NEWS
LDA	0.246
GAP	0.197

7. Conclusions

The two biggest challenges when measuring the performance of a topic model, are the unsupervised nature of the data and the creation of probabilistic 'soft' document clusters, rather than 'hard' clusters. The most common measure used to evaluate topic models, perplexity, solves these problems by using a word-predictability criterion. However, perplexity values computed with different feature sets are not comparable. We have shown that a modified version of topic stability is a useful alternative performance measure for topic models. At the core of topic stability is the ability to align topics from two independent topic assignments. For this purpose, the Hungarian method guarantees an optimal one-on-one alignment of topics.
We present a topic stability method that uses the average document correlation between topics as the performance metric. Our method does not suffer from the vocabulary dependency of perplexity. We also tested two topic models, LDA and GaP using this method and found that LDA performs better than GaP in terms of topic stability; this agrees with the assessment that arises from the use of document lables when those are available.
In future work, we would like to confirm that stability is a useful comparative measure, by studying other forms of perturbation, other corpora, and additional modelling algorithms. We also plan to perform a systematic comparison of our document correlation technique for topic stability with other techniques, such as the document co-occurrence scores used by Rigouste *et al.* [1]. Furthermore, we used the topic × document matrix to align the topics and indicate the topic stability. This is in contrast with Steyvers and Griffiths [2], who used the topic × word matrix for the same tasks. More work is needed to understand the respective properties of these two matrices in evaluating the performance of the topic model. (Our preliminary results suggest that word correlation is less reliable than document correlation, since closely related words may take on widely varying weights without affecting document classification.) Finally, we are in the process of implementing a suite of evaluation methods that address different aspects of topic models in order to describe the properties of these models more comprehensively.

8. References

[1] Rigouste, L., Cappé, O., and Yvon, F., "Inference and evaluation of the multinomial mixture model for text clustering." Inf. Process. Manage. 43(5), 1260-1280, 2007.

[2] Steyvers, M. and Griffiths, T., "Probabilistic topic models", In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), Latent Semantic Analysis: A Road to Meaning, pp 427-448. Laurence Erlbaum New Jersey, 2007.

[3] Blei, D. M., Ng, A. Y., and Jordan, M. I., "Latent dirichlet allocation", Journal of Machine Learning Research. 3:993-1022, 2003.

[4] Canny, J., "GAP: A Factor Model for Discrete Data", ACM Conference on Information Retrieval (SIGIR) Sheffield, England, July 2004.

[5] Lange, T., Roth, V., Braun, M.L. and Buhmann, J.M., "Stability-based validation of clustering solutions.", Neural Computation, 16(6):1299-1323, 2004.

[6] Rigouste, L., Cappé, O., and Yvon, F., "Evaluation of a probabilistic method for unsupervised text clustering" In Proceedings of the International Symposium on Applied Stochastic Models and Data Analysis (ASMDA), Brest, France, 2005.

[7] Buntine, W. and Jakulin, A., "Discrete component analysis", Tech. Rep., Helsinki Institute for Information Technology, July 2005.

[8] Deerwester S., Dumais, S. T Landauer T K., Furnas, G. W and Harshman, R. A., "Indexing by latent semantic analysis", Journal of the Society for Information Science, 41(6), 391-407, 1990.

[9] Hofmann, "Probabilistic latent semantic indexing" in SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, pp 50-57, ACM Press, 1999.

[10] Kuhn, H., "The Hungarian method for the assignment problem.", Naval Research Logistic Quarterly, 2:83-97, 1955.

[11] Frank, A., "On Kuhn's Hungarian method - a tribute from Hungary.", Tech. Rep. TR-2004-14, Egerváry Research Group, Budapest, 2004.

[12] Rosenbaum, P.R., "Optimal matching for observational studies.". Journal of the American Statistical Association, Vol.84, No.408, pp.1024-1032, 1989.

[13] Rigouste, L., "Méthodes probabilistes pour l'analyse exploratoire de données textuelles.", PhD Thesis, 2007.

[14] Glasgow IDOM - Cranfield collection `http://ir.dcs.gla.ac.uk/resources/test_collections/cran`

[15] Lang, K., "News Weeder: Learning to filter Netnews", Proc. 12th Intl Conf. Machine Learning, San Francisco, 1995