Evaluation of a Probabilistic Method for Unsupervised Text Clustering

Loïs Rigouste, Olivier Cappé, and François Yvon

Ecole Nationale Supérieure des Télécommunications (GET / CNRS UMR 5141)
46 rue Barrault, 75634 Paris Cedex 13, France
(e-mails: `rigouste, cappe, yvon` at `enst.fr`)

Abstract. In this contribution, we investigate the use of a simple probabilistic model for unsupervised document clustering in large collections of texts. The model consists of a mixture of multinomial distributions over the word counts, each component corresponding to a different theme.

The evaluation corpus is a medium size subset of the Reuters news feed, which comes with a manual categorization. The similarity between the clustering produced and this existing categorization is computed in terms of mutual information, and compared to the variations of log-likelihood and perplexity. We analyze the influence of the smoothing parameter, of the size of the vocabulary and of the addition of supervised information.

Our results, which are somewhat more pessimistic than those usually found in the literature, show that it is difficult to reach the quality of the manual categorization when no hint is given at the initialization step. We also show that a side effect of the so-called "curse-of-dimensionality" is that this probabilistic model yields the same results as a simpler, hard clustering algorithm.
Keywords: Text Mining, Unsupervised Clustering, Evaluation.

1 Introduction

Due to the wide availability of huge collections of text documents (news corpora, e-mails, web pages, scientific articles...), unsupervised clustering has emerged as an important text mining task. Several probabilistic models, performing a soft (non-deterministic) clustering of the data, such as Probabilistic Latent Semantic Analysis [Hofmann, 2001] or Latent Dirichlet Allocation [Blei *et al.*, 2002], have been introduced for that purpose. In this contribution, we study the simpler model [Nigam *et al.*, 2000, Clérot *et al.*, 2004] in which the corpus is represented by a mixture of multinomial distributions, each component corresponding to a different "theme". Dirichlet priors are set on the parameters and we use the Expectation-Maximization (EM) algorithm to obtain maximum a posteriori (MAP) estimates of the parameters.

To get a deeper understanding of the potentials of this approach, we consider a reasonably simple corpus, consisting of 5000 Reuters news stories taken from five different categories (as defined by Reuters). After introducing the two measures used for evaluation (perplexity and mutual information

between the obtained themes and the Reuters categorization), we investigate the influence of several aspects of the model. An interesting experimental outcome of this study is to show that, due to the high dimensionality of the problem, the model behaves almost like a hard clustering algorithm (with a specific distance measure).

2 The Model

We denote by n_D, n_W and n_T, respectively, the number of documents, the size of the vocabulary and the number of themes (that is, the number of components of the mixture model). Since we use a bag-of-words representation, the corpus is fully determined by the count matrix $C = (C_d(w))_{d=1\ldots n_D, w=1\ldots n_W}$, where the notation C_d is used to refer to the word counts of a specific document d. The multinomial mixture model is such that:

$$\mathrm{P}(C_d; \alpha, \beta) = \sum_{t=1}^{n_T} \alpha_t \frac{l_d!}{\prod_{w=1}^{n_W} C_d(w)!} \prod_{w=1}^{n_W} \beta_{wt}^{C_d(w)} \tag{1}$$

which corresponds to the following probabilistic generative mechanism:

- sample a theme t in $\{1, \ldots, n_T\}$ with probabilities $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{n_T})$;
- sample l_d (length of document d) words from a multinomial distribution with parameter $(l_d;\ \beta_{1t}, \beta_{2t}, \ldots, \beta_{n_W t})$.

The notation β is used to denote the collection of theme-specific word frequencies. Note that the document length itself is taken as an exogenous variable and its distribution is not accounted for in the model. As all documents are assumed to be independent, the corpus log-likelihood $\mathcal{L}$ is given by $\sum_{d=1}^{n_D} \log \mathrm{P}(C_d; \alpha, \beta)$.

To estimate the model parameters, we use the Expectation-Maximization (EM) algorithm with independent noninformative Dirichlet priors on α (with hyperparameter θ_α) and on the columns $\beta_{\bullet t}$, for $t = 1, \ldots, n_T$ (with hyperparameter θ_β). Denoting the current estimates of the parameters by α' and β' and the latent (unobservable) theme of document d by T_d, it is straightforward to check that each iteration of the EM algorithm updates the parameters according to:

$$\mathrm{P}(T_d = t | C; \alpha', \beta') = \frac{\alpha'_t \prod_{w=1}^{n_W} \beta'^{C_d(w)}_{wt}}{\sum_{t'=1}^{n_T} \alpha'_{t'} \prod_{w=1}^{n_W} \beta'^{C_d(w)}_{wt'}} \tag{2}$$

$$\alpha_t \propto \theta_\alpha - 1 + \sum_{d=1}^{n_D} \mathrm{P}(T_d = t | C; \alpha', \beta') \tag{3}$$

$$\beta_{wt} \propto \theta_\beta - 1 + \sum_{d=1}^{n_D} C_d(w)\, \mathrm{P}(T_d = t | C; \alpha', \beta') \tag{4}$$

where the normalization factors are determined by the constraints $\sum_{t=1}^{n_T} \alpha_t = 1$ and $\sum_{w=1}^{n_W} \beta_{wt} = 1$, for t in $\{1, \ldots, n_T\}$. It turns out that θ_α has little, if any, influence and we set $\theta_\alpha = 1$ in the following. For obvious reasons, we refer to $\theta_\beta - 1$ as the smoothing parameter. We set it to 0.1 to begin with.

3 Evaluation

To evaluate the performance of the model for unsupervised document clustering we use two different measures. The *perplexity*

$$\widehat{\mathcal{P}}^\star = \exp[-\frac{1}{l^\star} \sum_{d=1}^{n_D^\star} \log(\sum_{t=1}^{n_T} \alpha_t \prod_{w=1}^{n_W} \beta_{wt}^{C_d^\star(w)})]$$

quantifies how much the model is able to predict new data, denoted generically by the star superscript. The normalization by the total number of word occurrences $l^\star$ in the test corpus $C^\star$ is conventional and used to allow comparison with simpler models such as the unigram model, which ignores the document level. A second indicator is the *mutual information* between the clustering produced by the model and the Reuters categories, which is more directly related to our ability to accurately cluster the data. It is defined as:

$$\widehat{\mathcal{MI}}^\star = \sum_{c=1}^{n_C} \sum_{t=1}^{n_T} (\frac{1}{n_D^\star} \sum_{d=1}^{n_D^\star} \mathrm{P}(\Gamma_c|C_d^\star)\, \mathrm{P}(T_d = t|C_d^\star)) \times \log \frac{n_D^\star \sum_{d=1}^{n_D^\star} \mathrm{P}(\Gamma_c|C_d^\star) p(T_d = t|C_d^\star)}{(\sum_{d=1}^{n_D^\star} \mathrm{P}(\Gamma_c|C_d^\star))(\sum_{d=1}^{n_D^\star} \mathrm{P}(T_d = t|C_d^\star))}$$

where $\mathrm{P}(\Gamma_c|C_d)$ is the "probability" that document d belongs to category Γ_c (usually 0 or 1, as most documents belong to a unique Reuters category) and $\mathrm{P}(T_d = t|C_d)$ is the output of the model (probability that the document d belongs to theme t). The estimated mutual information is then normalized, respectively, by the marginal entropies of the themes and categories. The harmonic average of those scores (between 0 and 1) is referred to as the *(MI) F-Score*.

3.1 Baseline Performance

We selected 5,000 texts from the 2000 Reuters Corpus, from five well-defined categories (arts, sports, health, disasters, employment). All experiments are performed using ten-fold cross-validation (with 10 random splits of the corpus), with 30 iterations of the EM algorithm for each run and with five themes ($n_T = 5$). As will be seen below, initialization of the EM algorithm does play a very important role in obtaining meaningful document clusters. After a bit

of experimentation, we found that a good option is to make sure that, initially, all clusters overlap significantly and that none of the theme-dependent word probabilities is too small. The "Dirichlet" initialization thus consists in sampling an initial (fictitious) configuration of posterior probabilities in (2) which is close to equiprobability*.

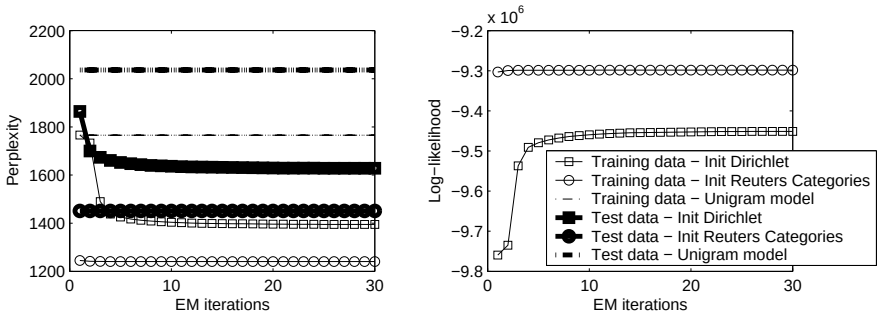

Fig. 1. Evolution of Perplexity and Log-likelihood over EM iterations.

To get an idea about the best achievable performance, we also used the Reuters categories as initialization. We establish a one-to-one mapping between the mixture components and the Reuters categories, setting for every document the initial posterior probability in (2) to 1 for a given theme. Figure 1 displays the corresponding training data likelihood (right) and perplexity as a function of the number of iterations. The first striking observation is that the gap between both initializations is huge. With the "Dirichlet" initialization, we are able to predict the word distribution more accurately than with the unigram model but much worse than with the somewhat ideal initialization. This gap is also patent for the training data log-likelihood. In the following, we report only the values obtained after the last EM iteration, since the variations after the first few iterations are small (note that this phenomenon is particularly marked for the Reuters initialization). Also, we no more report the perplexity on the training data since it conveys the same information as log-likelihood.

The Mutual Information F-Score is similarly oriented with a final value of 0.87 for the Reuters initialization and 0.25 for the "Dirichlet" one. To get an idea of the signification of these numbers, we randomly perturbated a certain amount of the Reuters tags and computed the MI F-Score with the original

* It is not possible to start with exact equiprobability, or, else, it can be seen from the update equations that all word distributions are similar and the clusters never separate from one another. Hence we sample from a Dirichlet distribution with the same parameter for every component. This variable controls the variance of the probabilities sampled. It also has an interesting influence on the results that we do not develop here.

categorization. Proceeding this way, perturbing (respectively) 5%, 15% and 50% of the document labels gives F-Score of 0.9, 0.7 and 0.25. Hence 0.25 corresponds to a rather poor performance. Now we check if this gap between both initializations can be reduced when tuning the smoothing parameter.

3.2 Influence of the Smoothing Parameter

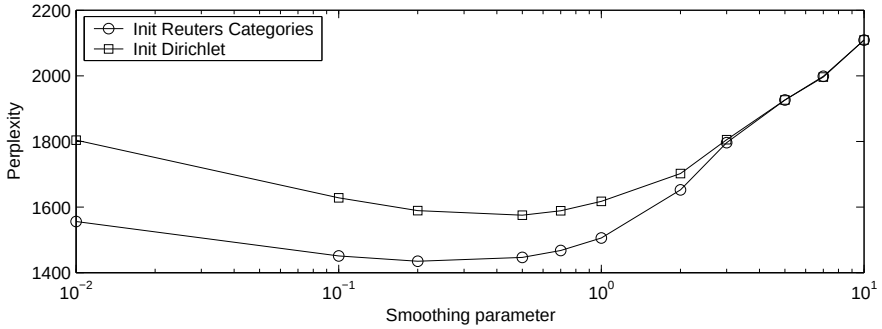

Fig. 2. Perplexity as a function of smoothing.

Figure 2 depicts the influence of the smoothing parameter $\theta_\beta - 1$ in terms of perplexity. For both initializations, the best performances are obtained for smoothing parameters between 0.1 and 2, with an optimum at 0.5. Clearly using some prior information about the fact that word probabilities should not get too small helps to fit the distribution of new data, even for words that are rarely (or even never) seen in association with a given theme.

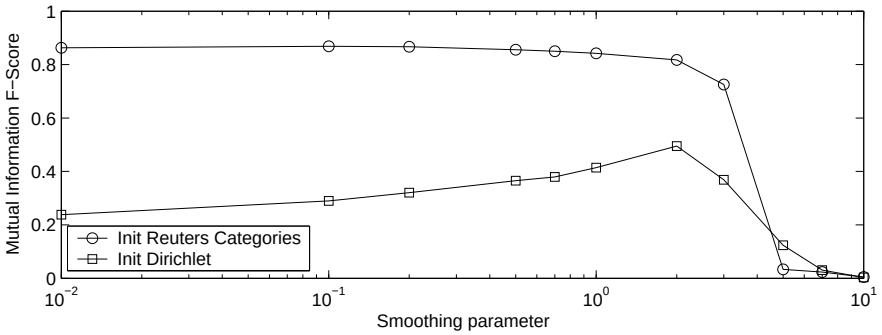

Fig. 3. Evolution of mutual information as a function of smoothing

Figure 3 reveals a slightly different behavior for the MI F-Score. First, except when using very large (5 or more) values of the smoothing parame-

ters, which yields a serious drop in performance, the categorization accuracy is rather insensitive to smoothing for the Reuters initialization. Of more practical interest however is the behavior for the "Dirichlet" initialization, which is roughly consistent with what is observed in Figure 2, except for the fact that the optimum is obtained for higher values of the smoothing parameter (around 2). A possible explanation of this observation that more smoothing improves categorization capabilities (even if it slightly degrades distribution fit) is that the model is so coarse and the data so sparse that only quite frequent words are helpful in categorizing; the other words are essentially misleading, unless properly initialized. This suggests that removing rare words from the vocabulary should improve the classification accuracy.

As an aside, it is interesting to observe, in figure 4, that the variations of the MI F-Score is highly dependent on the initialization and the smoothing parameter. For large (unrealistic) values, the more iterations we conduct, the more inaccurate prior information we give to the model and the worst the performances get. For the initialization "Dirichlet", the optimal value of $\theta_\beta - 1$ (2) clearly corresponds to the higher increasing curve. From 3, the clustering begins to degrade after 5 or 6 iterations.

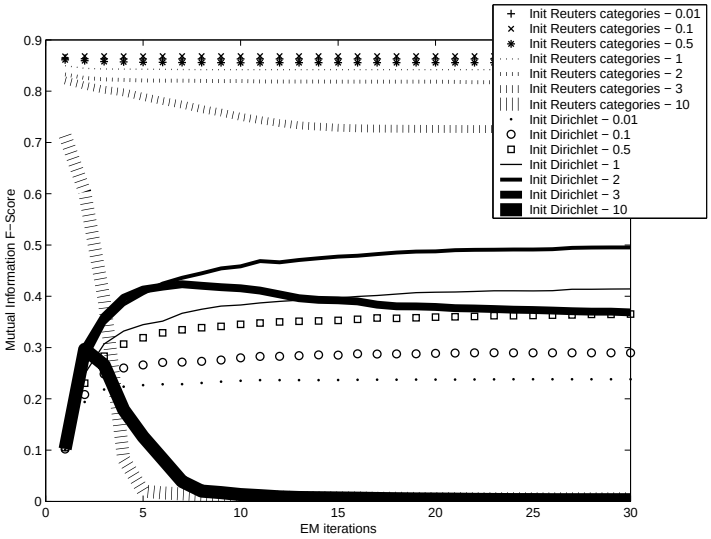

Fig. 4. Evolution of mutual information as a function of EM iterations, with different smoothing values

3.3 Adjusting the Vocabulary Size

A valid question, after having decided to ignore part of the vocabulary, is if we should rather cut rare words (hapax) or frequent words (stop-words). We

try both strategies, removing consecutevely tens, hundreds and thousands of terms from the vocabulary. The words discarded are simply not taken into account in the count matrix**.

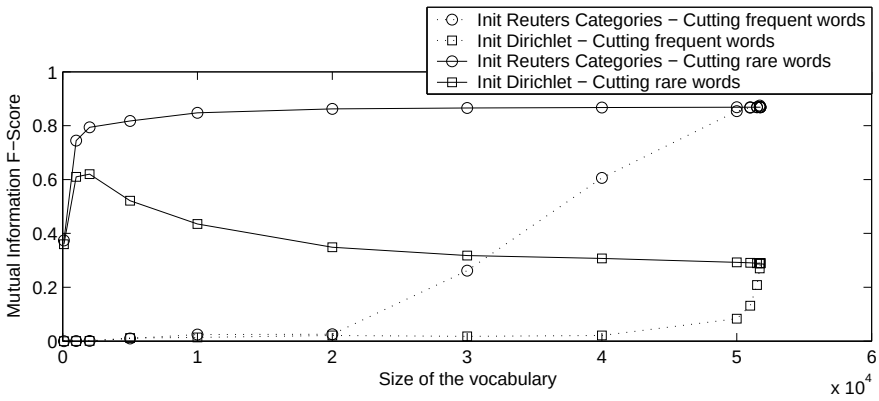

Fig. 5. Evolution of mutual information when removing rare words.

Results in term of perplexity are not helpful, since the size of the vocabulary has an impact on perplexity which is hard to distinguish from the variations due to a possible better fit of the model. The MI F-Score, on the other hand, is meaningful even when the vocabulary sizes are different. The results in Figure 5 suggest that we can substantially improve the performance of the model with the "Dirichlet" initialization, by keeping a very limited number of frequent words (around 2,000). Note that the obtained F-Score is still far from reaching the performance attained with the Reuters initialization. This agrees with our previous observation that even the rarest word may be informative, when properly initialized.

On the other hand, removing frequent words almost always hurts as one can see when reading the dashed curves from right (full vocabulary) to left (all words removed from vocabulary). Only in the case of the "Reuters Categories" initialization, discarding the 50 or 100 most frequent words leads to a slightly better performance but it is hardly visible on the figure. Then the MI F-Score steadily decreases when cutting frequent words. The score is almost 0 with 20,000 rare words, which is not surprising, since, in this case, the vocabulary only consists of words with 1 occurence in the whole corpus and a text is therefore reduced to at most a dozen of terms.

** We do not study here the effect of another common trick: grouping all unknown words under the token "Out Of Vocabulary".

3.4 Adding Supervised Information

Clearly none of the variants discussed so far is susceptible of bridging the gap between the ideal results, obtained using Reuters categories, and the results achievable in practice. To this aim, we consider using a limited number of texts (2, 5, 10, 20 or 50) from each theme to initialize the theme-dependent word frequency parameters. Note that in this case, the EM algorithm is used in "semi-supervised" mode, updating only the posterior probabilities for the texts whose category is truly unknown. In each case and each repetition (we are still using ten-fold cross validation), we repeat the experiment ten times to make up for the chances of picking "unrepresentative" texts.

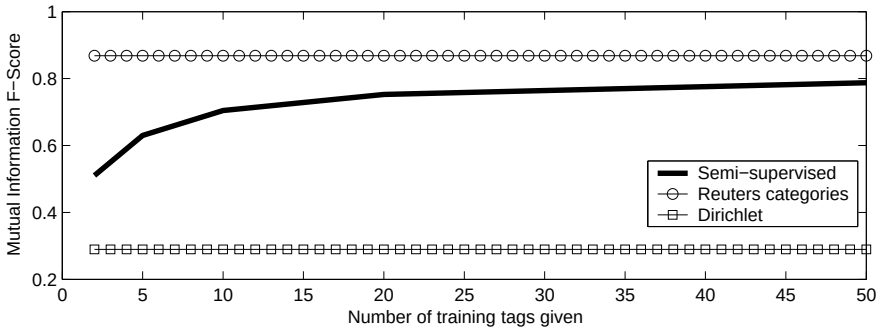

Fig. 6. Evolution of mutual information when using partial category information.

Figure 6 shows that, as expected, results improve with the number of known text tags and that acceptable values are obtained quite fast: with 10 tags per theme (that is 1.1% of the training documents labeled), the obtained F-Score is already about 0.7 (to be compared with 0.8 when 5.5% of the labels are known and 0.9 when all the labels are known).

Figure 7 conveys the same impression and suggests that knowing 20 or 50 labels per category is almost equivalent in terms of perplexity and log-likelihood. Hence, knowing a few percents of the document labels is enough to catch up on word distribution modelling (perplexity) and a few additional percents suffice to obtain very good categorization performance.

3.5 Equivalence with a Non-Probabilistic Algorithm

A surprising fact, when working with this model, is the huge fraction of posterior probabilities (that a document belongs to a given theme) dramatically close to 0 or 1. Indeed, when starting from Reuters categories, the proportion of texts classified in only one given theme (that is, with probability one up to machine precision) is almost 100%. Since we start from the opposite point of "extreme fuzziness", this effect is not as strong with the "Dirichlet"

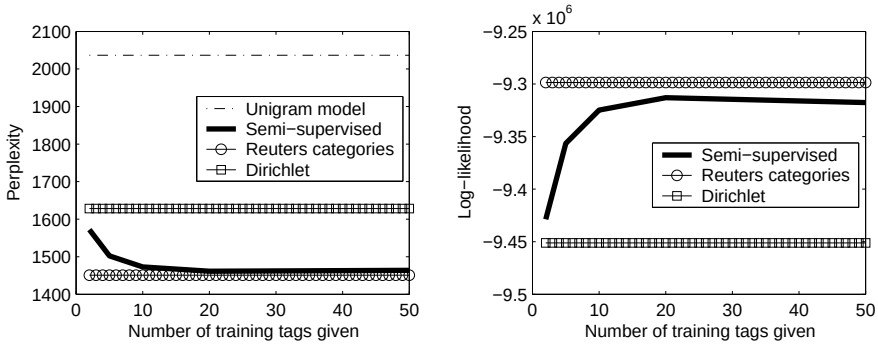

Fig. 7. Evolution of perplexity and log-likelihood when using partial category information.

initialization. Still, after the fifth iteration, more than 90% of the documents are categorized with absolute certainty.

Therefore, we compare the results obtained with an algorithm similar to EM but based on hard clustering. This is in fact a version of K-means, with the following distance between a text $d \in \{1, \ldots, n_D\}$ and theme (or cluster) $t \in \{1, \ldots, n_T\}$:

$$dist(d,t) = \frac{1}{\alpha_t \prod_{w=1}^{n_W} \beta_{wt}^{C_d(w)}}$$

This distance is computed for every document and every theme and each document is assigned to its closest theme. The reestimation of the parameters β_{wt} is done according to (4) where the posterior "probabilities" are always either 0 or 1. α_t simply becomes the proportion of documents in theme t and β_{wt} the ratio of the number of occurrences of w in theme t over the total number of occurrences in documents in theme t.

$$\alpha_t = \frac{1}{n_D} \sum_{d=1}^{n_D} \delta_{\{d \in t\}}$$

$$\beta_{wt} = \frac{\sum_{d \in t} C_d(w)}{\sum_{w=1}^{n_W} \sum_{d \in t} C_d(w)}$$

We applied this algorithm to the same dataset, with the same initialization procedures as above. At the end of each iteration, we compute the Mutual Information F-Score between the fuzzy clustering produced by EM and the hard clustering produced by this version of K-means.

- With the "Reuters Categories" initialization, the Mutual Information F-Score between the clusterings produced is 1 after one iteration.
- With the "Dirichlet" initialization, which is somehow the opposite of a hard clustering, the F-Score between the soft and hard clustering converges very fast to 1 and is greater than 0.99 after five iterations.

In both cases, the different outputs of the fuzzy and hard methods become indiscernible after very few iterations. We believe that this behavior of EM can be partly explained by the large dimensionality of the space of documents*** . This assumption can be verified with experiments on artificially simulated datasets.

4 Conclusion

In this article, we study a mixture model of thematic multinomial distributions for corpus clustering. We show that, even though some parameters have a real influence and actually help reduce the gap, there exists a large difference between the best achievable performance and the ones we are able to obtain without prior supervised information. Eventually, we note that in this case, a fuzzy clustering approach is just uselessly time consuming since we get exactly the same results with a hard clustering version of the algorithm.

In future work, it would be interesting to check if the same conclusions apply to more complicated models such as PLSA and LDA. Besides, we are still looking for ways to improve the performances of the model with the "Dirichlet" initialization, for example using other inference methods.

5 Acknowledgment

This work has been supported by France Télécom, Division R&D, under contract n°42541441.

References

[Blei *et al.*, 2002]David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 14, pages 601–608, Cambridge, MA, 2002. MIT Press.

[Clérot *et al.*, 2004]Fabrice Clérot, Olivier Collin, Olivier Cappé, and Eric Moulines. Le modèle "monomaniaque" : un modèle statistique simple pour l'analyse exploratoire d'un corpus de textes. In *Colloque International sur la Fouille de Texte (CIFT'04)*, La Rochelle, 2004.

[Hofmann, 2001]Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning Journal*, 42(1):177–196, 2001.

[Nigam *et al.*, 2000]Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, and Tom M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.

*** The vocabulary contains more than 50,000 words.