Quelques observations sur le modèle LDA

Loïs Rigouste, Olivier Cappé et François Yvon
{*rigouste,cappe,yvon*} *at enst.fr*

GET – Télécom Paris & CNRS – LTCI (UMR 5141)
46 rue Barrault - F75013 Paris

Abstract

Unsupervised text clustering is a basic task of text mining, which consists in building thematically homogeneous groups or hierarchies of documents in a collection. Non-deterministic class assignments can also be used to derive numerical representations of documents in a semantic vector space.

In this contribution, we present a detailed study of a probabilistic model for unsupervised clustering, the Latent Dirichlet Model (LDA), originally introduced in (Blei et al., 2002; Griffiths and Steyvers, 2002; Minka and Lafferty, 2002). More specifically, we study one particular estimation technique for this model, as well as methods for computing the likelihood of a document. Experimental results obtained with this model are contrasted with those obtained with the simpler mixture of multinomials model.

Résumé

Une des tâches de la fouille de données textuelles consiste à construire de manière non-supervisée des classes ou des hiérarchies de documents thématiquement homogènes à partir d'un corpus. Une variante consiste à envisager des associations non-déterministes entre classes et documents, permettant de dériver des représentations numériques synthétiques des documents, qui sont alors vus comme des points dans un espace sémantique latent.

Dans cet article, nous étudions en détail le modèle proposé simultanément dans (Blei et al., 2002; Griffiths and Steyvers, 2002; Minka and Lafferty, 2002) sous le nom de *Allocation Dirichlet Latente* (*Latent Dirichlet Allocation* ou LDA). Nous nous intéressons ici plus particulièrement à l'étude de l'estimation des paramètres de ce modèle et au calcul de la vraisemblance d'un document. Les résultats obtenus sont comparés avec celui d'un modèle plus simple, le modèle de mélange de multinomiales.

Keywords: Clustering, Modèles Génératifs, Allocation Dirichlet Latente

1 Introduction

Le problème qui nous intéresse est celui de la *catégorisation non-supervisée* de documents, soit, dit en d'autres termes, celui du repérage de classes thématiquement homogènes dans un corpus textuel. Si cette problématique est relativement ancienne en lexicométrie (Benzécri et al., 1981), la disponibilité de larges corpus numérisés et les besoins convergents d'applications de recherche d'information et de fouille de textes ont suscité de nouvelles propositions : en particulier le modèle LSI (Deerwester et al., 1990), son équivalent probabiliste (Hofmann, 2001), ou encore tout un ensemble de modèles probabilistes génératifs : le modèle de mélange de lois multinomiales ou poissoniennes (Nigam et al., 2000; Clérot et al., 2004), le modèle GAP (Canny, 2004), etc. (Voir Buntine and Jakulin, 2004 pour une revue de ces différents modèles).

Dans cet article, nous étudions en détail le modèle probabiliste proposé simultanément dans (Blei et al., 2002; Griffiths and Steyvers, 2002; Minka and Lafferty, 2002) sous le nom d'*Allocation Dirichlet Latente* (*Latent Dirichlet Allocation* ou LDA). Le trait principal de ce modèle est qu'il

Ce travail a été soutenu par France-Télécom, R&D, dans le cadre du contrat n° 42541441.

permet indirectement d'établir une association réellement non-déterministe entre documents et thèmes sous-jacents.

Après une présentation du modèle (Section 2), nous nous intéressons plus particulièrement à l'estimation des paramètres de ce modèle (Section 3) et au calcul de la vraisemblance d'un document (Section 4). Nous présentons ensuite divers résultats expérimentaux (Section 5) en les contrastant avec ceux obtenus avec un modèle de mélange de lois multinomiales.

2 Le modèle LDA

Avant de présenter formellement le modèle, il n'est pas inutile de donner un éclairage plus intuitif sur ses motivations principales. LDA partage avec de nombreux modèles une représentation en « sac-de-mots », selon laquelle chaque document est représenté par un profil d'occurrences. Un corpus est donc représenté par un ensemble (une matrice) de vecteurs de comptes, qu'il s'agit de modéliser.

Dans le modèle de mélange de multinomiales (Nigam et al., 2000), on suppose que chaque document du corpus appartient à un thème latent (à une classe) *unique* : conditionnellement au thème, le vecteur représentant les profils d'occurrences dans un document est vu comme la réalisation d'un tirage sous une loi multinomiale, dont les paramètres dépendent du thème. Dans la mesure où les thèmes sont inconnus, l'estimation de ces paramètres repose sur la mise en œuvre de l'algorithme EM (*Expectation-Maximisation*), qui va permettre de déterminer (i) les paramètres des lois multinomiales associées aux différents thèmes ; (ii) pour chaque document, sa probabilité d'appartenir à chacun des thèmes. Ce vecteur de probabilité peut être également vu comme une représentation numérique résumée du document, qui peut être exploitée à des fins variées (comparaison de documents, visualisation du corpus etc.).

LDA vise à assouplir ce cadre, en proposant que l'association non-déterministe entre thèmes et documents soit médiatisée par les occurrences : *ce sont donc les occurrences qui sont ventilées par thème*, et non plus les documents. Les différentes occurrences au sein d'un même document restent toutefois globalement liées par une variable latente qui contrôle la distribution des thèmes au sein du document.

2.1 Le modèle de génération

Nous notons n_D, n_W, n_T respectivement le nombre de documents dans le corpus, la taille du vocabulaire d'indexation et le nombre de thèmes. C_{wd} est le terme général de la matrice de comptes ; l_d désigne la longueur (le nombre total d'occurrences) du document $d \in \{1, \dots, n_D\}$, qui peut être également représenté comme un vecteur d'occurrences $(W_{d1}, \dots, W_{dl_d})$; on note enfin les hyperparamètres du modèle $\lambda_\alpha, \lambda_\beta \in \mathbb{R}_+^*$.

Le modèle de génération du corpus proposé par LDA est alors le suivant :

– Pour chaque thème $t \in \{1, \dots, n_T\}$, tirer les paramètres des lois discrètes probabilisant les occurrences des mots du vocabulaire selon une loi de Dirichlet[1] $\beta_t = (\beta_{t1}, \dots, \beta_{tn_W}) \sim \text{Dir}(\lambda_\beta, \dots, \lambda_\beta)$. β_{tw} s'interprète comme la probabilité de l'occurrence du mot w dans un document du thème t.
– Pour chaque document $d \in \{1, \dots, n_D\}$:
  – Tirer la distribution des thèmes dans d selon $\alpha_d = (\alpha_{d1}, \dots, \alpha_{dn_T}) \sim \text{Dir}(\lambda_\alpha, \dots, \lambda_\alpha)$. Chaque α_{dt} indique donc la proportion des occurrences du document d qui sont associées

[1] Cette loi est présentée succinctement en annexe.

au thème t.
- Pour chaque position i dans d, $i \in \{1, \ldots, l_d\}$:
  - Tirer un thème selon une loi discrète $T_{di} \sim \text{Disc}(\alpha_d)$.
  - Tirer un mot conditionnellement au thème selon : $W_{di} \sim \text{Disc}(\beta_{T_{di}})$.

LDA voit donc chaque document comme un ensemble d'occurrences apparaissant dans un ordre arbitraire, ce qui le rapproche du modèle du « sac-de-mots ». Le choix d'un thème est effectué *indépendamment* pour chaque occurrence du document, sous la contrainte du respect global de la distribution des thèmes fixée par α_d : il est donc tout à fait possible de considérer des changements de thème à chaque position du document.

Comme le note (Buntine and Jakulin, 2004), on peut également voir ce modèle comme un modèle de mélange de lois discrètes, dont les coefficients de mélange α_d sont tirés indépendamment pour chaque document. Chaque document résulte alors de l_d tirages indépendants selon une loi discrète de paramètres $(\sum_{t=1}^{n_T} \alpha_{dt}\beta_{t1}, \ldots, \sum_{t=1}^{n_T} \alpha_{dt}\beta_{tn_W})$, c'est à dire que le profil d'occurrences caractéristique de chaque document s'obtient comme une combinaison barycentrique, contrôlée par la variable latente α_d, des n_T profils de base $\beta_1, \ldots, \beta_{n_T}$.

2.2 Calcul de la vraisemblance complète

Nous introduisons ici deux notations supplémentaires. On suppose fixée une configuration de variables indicatrices de thèmes $T = (T_{11}, \ldots, T_{n_D l_{n_D}})$ pour chaque occurrence de chaque document du corpus : T_{di} est le thème associé à l'occurrence i du document d. T étant connue, on note N_{td} le nombre d'occurrences de d qui sont assignées au thème t ; de manière équivalente, K^d_{tw} désigne le nombre d'occurrences du mot w assignées au thème t dans le document d, $K_{tw} = \sum_{d=1}^{n_D} K^d_{tw}$ la même quantité pour l'ensemble du corpus et $K_t = \sum_{w=1}^{n_W} K_{tw}$ le nombre total d'occurrences dans le thème t.

Conditionnellement au vecteur α_d et à la matrice β, la probabilité jointe de W_d et T_d s'exprime alors par (les occurrences étant indépendantes)[2] :

$$\begin{aligned} \mathrm{P}(W_d, T_d|\alpha_d, \beta) &= \prod_{i=1}^{l_d} \mathrm{P}(W_{di}|T_{di}, \alpha_d, \beta)\, \mathrm{P}(T_{di}|\alpha_d, \beta) \\ &= \prod_{t=1}^{n_T} \left(\alpha_{dt}^{N_{td}} \prod_{w=1}^{n_W} \beta_{tw}^{K^d_{tw}} \right) \end{aligned}$$

Les documents étant également supposés indépendants, la vraisemblance du corpus et des indicatrices latentes de thèmes s'exprime comme le produit :

$$P(W, T|\alpha, \beta) = \prod_{d=1}^{n_D} \mathrm{P}(W_d, T_d|\alpha_d, \beta)$$

L'observation principale de (Griffiths and Steyvers, 2002) est qu'il est possible d'intégrer cette vraisemblance sous la loi *a priori* des paramètres α et β, permettant d'aboutir à la forme analytique suivante :

$$\mathrm{P}(W, T) = \prod_{d=1}^{n_D} \left(\frac{\Gamma(n_T\lambda_\alpha)}{\Gamma(\lambda_\alpha)^{n_T}} \frac{\prod_{t=1}^{n_T} \Gamma(N_{dt} + \lambda_\alpha)}{\Gamma(l_d + n_T\lambda_\alpha)} \right)$$

Les deux facteurs de ce produit correspondent respectivement à $\mathrm{P}(W|T)$ et à $\mathrm{P}(T)$.

[2] Les détails de ce calcul, comme ceux du calcul suivant figurent dans l'annexe.

3 Estimation du modèle

Estimer le modèle LDA consiste essentiellement à déterminer les valeurs de β à partir des observations. Plusieurs techniques d'estimation ont été proposées : l'inférence variationnelle, dans (Blei et al., 2002) ; la méthode *expectation-propagation* (Minka and Lafferty, 2002) et l'utilisation d'un échantillonneur de Gibbs (Griffiths and Steyvers, 2002). Suivant les recommandations convergentes de (Griffiths and Steyvers, 2004; Buntine and Jakulin, 2004), c'est cette dernière méthode que nous avons étudiée.

Le principe général consiste à construire une séquence de configurations d'indicatrices de thèmes dont la loi stationnaire soit $\mathrm{P}(T|W)$. Pour cela, partant d'une configuration aléatoire, on modifie itérativement les assignations thématiques T_{di} de chaque occurrence du corpus en simulant sous la loi conditionnelle

$$\mathrm{P}(T_{di}|T_{-di}, W) = \frac{\mathrm{P}(T, W)}{\mathrm{P}(T_{-di}, W_{-di})} \frac{1}{\mathrm{P}(W_{di}|T_{-di}W_{-di})}$$

où T_{-di} désigne le vecteur T privé de l'élément T_{di}. Le second facteur de l'équation ci-dessus ne dépend pas de T_{di} et peut donc être vu comme un facteur de normalisation. Comme le montre (Griffiths and Steyvers, 2002), en utilisant l'expression précédente de la vraisemblance complète pour (W, T) et (W_{-di}, T_{-di}), on aboutit, après simplifications, à :

$$\mathrm{P}(T_{di}|T_{-di}, W) \propto \frac{(K_{T_{di}W_{di}} + \lambda_\beta - 1)}{(K_{T_{di}} + n_W\lambda_\beta - 1)} \frac{(N_{dT_{di}} + \lambda_\alpha - 1)}{(l_d + n_T\lambda_\alpha - 1)} \tag{1}$$

Cette expression peut être retrouvée plus simplement, en utilisant le fait que T_{di} ne dépend des autres observations qu'à travers les indicatrices de thèmes des autres mots du même document. Il vient alors :

$$\begin{aligned} \mathrm{P}(T_{di}|T_{-di}, W) &\propto \mathrm{P}(T_{di}, W_{di}|T_{-di}, W_{-di}) &(2)\\ &\propto \mathrm{P}(W_{di}|T, W_{-di})\,\mathrm{P}(T_{di}|T_{-di}) &(3)\end{aligned}$$

Chaque terme du produit (1) peut être vu comme un estimateur des probabilités impliquées dans (3). Ainsi $\mathrm{P}(T_{di}|T_{-di})$ est estimé par le nombre de fois où le thème T_{di} a été vu dans le document d (non compris l'occurrence di), auquel s'ajoute un terme d'*a priori*. Après renormalisation par la longueur du document tronqué de l'occurrence courante ($l_d - 1$), on retrouve un estimateur habituel pour $\mathrm{P}(T_{di}|T_{-di})$:

$$\frac{N_{dT_{di}} - 1 + \lambda_\alpha}{l_d - 1 + n_T\lambda_\alpha}$$

Le même argument s'applique pour l'occurrence W_{di} : connaissant les variables latentes de thèmes de tous les autres mots du corpus, $\mathrm{P}(W_{di}|T, W_{-di})$ s'estime simplement par :

$$\frac{K_{T_{di}W_{di}} - 1 + \lambda_\beta}{K_{T_{di}} - 1 + n_W\lambda_\beta}$$

L'algorithme d'estimation consiste à faire évoluer l'échantillonneur selon (1) à partir d'une configuration initiale, puis à collecter des valeurs des paramètres β, qui sont finalement moyennées. Ce moyennage est problématique, car il est en théorie possible qu'au fil des simulations, les thèmes soient renumérotés. Il présuppose donc la capacité d'apparier la numérotation des thèmes de l'échantillon M avec ceux d'un autre échantillon M'.

4 Calcul de la vraisemblance, classification des documents

Dans cette section, nous supposons les paramètres connus et nous intéressons plus directement à deux questions liées à la tâche initiale de construction d'une classification des documents : quelle est la vraisemblance d'un document ? Comment calculer la distribution des thèmes associée à un document ? Les différents articles concernant LDA sont relativement peu diserts sur ces questions, hormis (Minka and Lafferty, 2002), qui propose une méthode fondée sur l'algorithme *expectation propagation.* (Griffiths and Steyvers, 2004) mentionne une méthode s'appuyant sur la technique de l'échantillonnage d'importance, mais ne détaille pas sa mise en œuvre.

4.1 Calcul de la vraisemblance

Considérons le document d et supposons que la distribution de thèmes α_d pour ce document est connue. Il vient :

$$\mathrm{P}(W_d|\alpha_d) \;=\; \prod_{w=1}^{n_W}(\sum_{t=1}^{n_T}\alpha_{dt}\beta_{tw})^{C_{wd}}$$

On remarque que $\log(\mathrm{P}(W_d|\alpha_d))$ est une fonction concave des α_{dt} : après reparamétrisation par $\alpha_{d,1}, \ldots \alpha_{d,n_T-1}$, il vient en effet :

$$\log(\mathrm{P}(W_d|\alpha_d)) = \sum_{w=1}^{n_W} C_{wd}\log\left(\sum_{t=1}^{n_T-1}\alpha_{dt}\beta_{tw} + (1-\sum_{t'=1}^{n_T-1}\alpha_{dt'})\beta_{n_T w}\right)$$

dont le Hessien est semi-defini négatif. Cette fonction atteint donc un maximum unique sur le simplexe dans lequel évolue $\alpha_d = (\alpha_{d,1}, \ldots, \alpha_{d,n_T})$.

Le calcul de la vraisemblance demande de marginaliser cette probabilité conditionnelle par rapport à α_d, soit de calculer :

$$\mathrm{P}(W_d) \;=\; \int_{\alpha_d}\prod_{w=1}^{n_W}(\sum_{t=1}^{n_T}\alpha_{dt}\beta_{tw})^{C_{wd}}p(\alpha_d)d\alpha_d$$

Cette intégrale n'admet pas de résolution analytique ; il est en revanche possible de mettre en œuvre une approche de Monte Carlo, consistant à tirer M valeurs $\alpha_d^{(m)}$ sous la loi *a priori* (Dirichlet) et à approximer $\mathrm{P}(W_d)$ par : $\frac{1}{M}\sum_{m=1}^{M}\mathrm{P}(W_d|\alpha_d^{(m)})$.

4.2 Classification d'un document

Estimer la distribution des thèmes dans un document revient à calculer l'espérance conditionnelle de α_d. Pour ce faire, on peut utiliser le même échantillon de Monte Carlo que précédemment en approximant $E(\alpha_d|W_d, \beta)$ par :

$$\frac{\sum_{m=1}^{M}\alpha_d^{(m)}\mathrm{P}(W_d|\alpha_d^{(m)})}{\sum_{m=1}^{M}\mathrm{P}(W_d|\alpha_d^{(m)})}$$

5 Expérimentations

5.1 Choix et préparation des données

Pour cette série d'expériences, nous avons utilisé un petit sous-ensemble d'articles en langue anglaise publiés en 2000 et tirés de la collection (Reuters, 2000). Notre corpus contient un total de 5 000 documents, répartis équitablement parmi cinq rubriques Reuters : "Sport", "Emploi", "Art", "Catastrophes" et "Santé". Une centaine de documents sont à cheval sur deux rubriques et possèdent deux étiquettes catégorielles. Aucun pré-traitement particulier n'est appliqué, sinon le passage en minuscule de tous les mots. Au terme de ce traitement, on dénombre 1 400 213 occurrences, correspondant à 42 437 types.

5.2 Mesures

Les mesures utilisées ci-après peuvent être calculées aussi bien sur le corpus d'apprentissage que sur le corpus de test. Nous nous intéressons plus spécifiquement aux valeurs obtenues sur le corpus de test et nous notons les valeurs sur ce corpus avec un exposant $\star$ pour éviter les confusions avec le corpus d'apprentissage.

On considère principalement deux mesures : la *perplexité*, qui correspond à une expression (normalisée par le compte total d'occurrences) de la log-vraisemblance du corpus, laquelle est calculée selon les principes exposés à la section 4. Formellement,

$$\widehat{\mathcal{P}}^{\star} = \exp\left(-\frac{1}{l^{\star}}\sum_{d=1}^{n_D^{\star}} \log(\mathrm{P}(W_d^*))\right)$$

avec $l^{\star} = \sum_{d=1}^{n_D} l_d^*$ le nombre total d'occurrences dans le corpus de test. Cette normalisation est conventionnelle et permet de mettre en perspective les valeurs obtenues avec ce que donnent d'autres modèles comparables, tels que le modèle unigramme.

Nous comparons les variations de cette quantité avec celles du score de classification, en comparant, sur le corpus de test, les thèmes prédits avec les étiquettes Reuters originales.

5.3 Observations

Estimation des paramètres Comme il apparaît sur la figure 1, le comportement de l'échantillonneur de Gibbs se caractérise par (i) une convergence rapide (100 à 200 simulations) vers une configuration dont les simulations suivantes ne s'écarteront qu'à la marge[3] ; (ii) une forte dépendance vis-à-vis des conditions initiales des valeurs de perplexité finalement atteintes. Ce résultat est confirmé par l'observation de la figure 2, qui donne pour chacune des 20 initialisations le pourcentage d'accord avec les catégories initiales. Suivant les cas, l'algorithme parvient à retrouver entre 60% et 80% des catégories Reuters.

Distribution des classes Le modèle de mélange de lois multinomiales ne construit qu'en apparence des associations floues entre thèmes et documents. Dans la pratique (Rigouste et al., 2005), la très grande majorité des documents est affectée avec probabilité 1 à un thème unique. Pour contraster ce comportement avec celui de LDA, nous avons, pour chacun de ces deux

[3] En particulier, on n'observe pas de renumérotation des classes sur une trajectoire de l'échantillonneur : il semble donc possible de moyenner directement les valeurs des paramètres sur un ensemble de configurations.

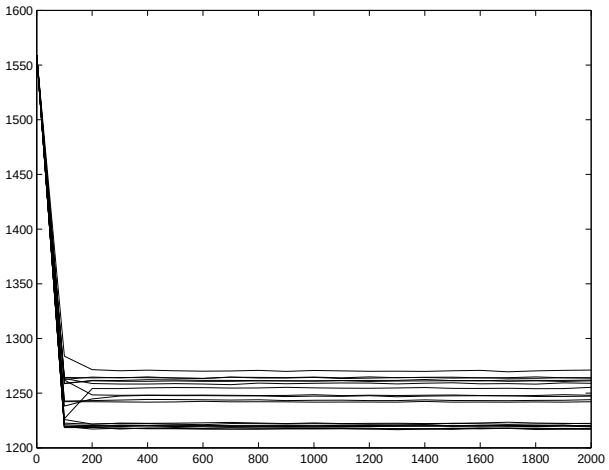

FIG. 1 – Évolution de la perplexité.
Chacune des 20 courbes représente l'évolution de la perplexité calculée sur 500 documents de test en fonction du nombre de simulations. Pour chaque point, les paramètres sont estimés à partir de la configuration d'indicatrices courante.

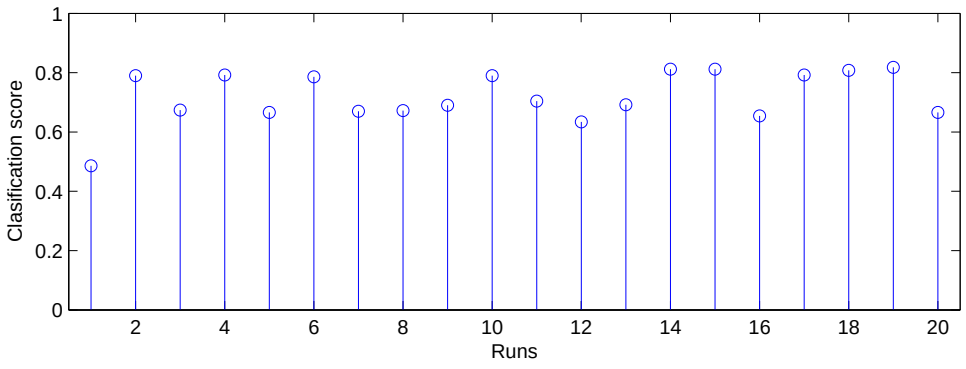

FIG. 2 – Score de classification pour chacune des vingt initialisations.

modèles, calculé l'entropie moyenne de la distribution des thèmes assignés à un document pour un ensemble de 500 documents de test (avec $n_T = 5$). En conservant les notations de la section 4, la grandeur mesurée est donc une moyenne (sur 20 simulations) de :

$$\exp\left(\frac{1}{{n_D}^*}\sum_{d=1}^{{n_D}^*} -\alpha_d \log(\alpha_d)\right)$$

Lorsque l'on construit une classification en n_T thèmes, cette grandeur varie entre 1 (affectation déterministe) et n_T (répartition uniforme). Alors que, pour le modèle de mélange de lois multinomiales, cette grandeur est toujours très proche de 1 (sur les 20 initialisations, la moyenne est : 1.02), pour LDA, cette valeur approche 2 (en moyenne : 1.98), prouvant que ce modèle permet effectivement de construire des classifications non-déterministes.

Calcul de la vraisemblance La figure 3 représente la dispersion statistique de l'estimateur de la vraisemblance de deux documents choisis aléatoirement en fonction du nombre de simulations Monte Carlo, selon que le nombre de thèmes est respectivement fixé à 5 et à 20.

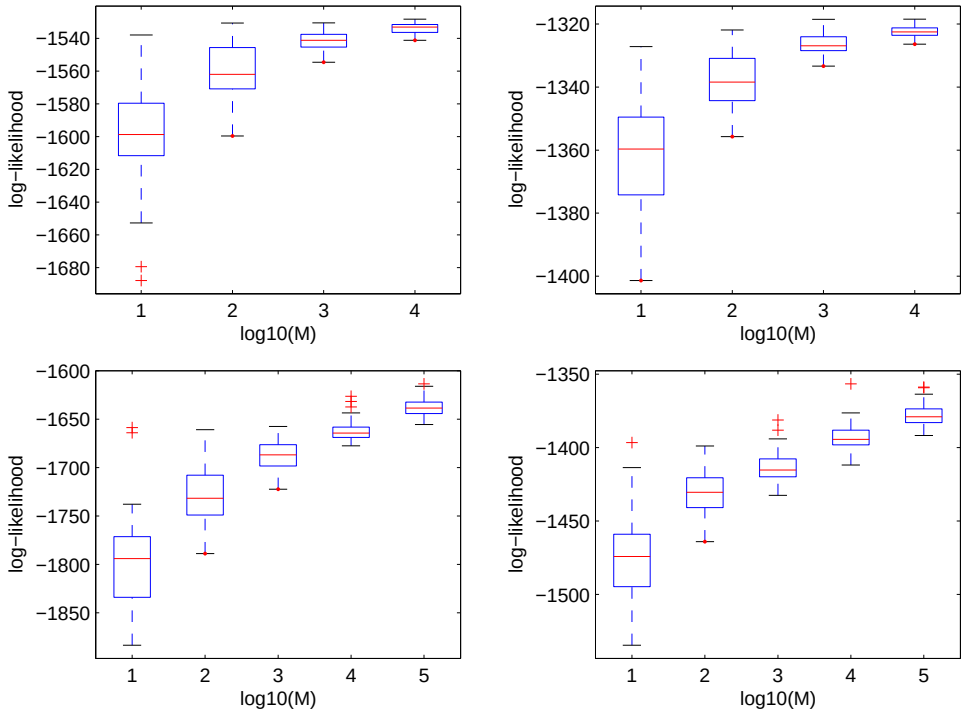

FIG. 3 – Estimation de la log-vraisemblance en fonction du nombre de simulations Monte Carlo pour, en haut $n_T = 5$, en bas $n_T = 20$.

Deux tendances semblent se dessiner :
– lorsque le nombre de thèmes augmente, le nombre de simulations nécessaires pour stabiliser l'estimation de la vraisemblance d'un document s'accroît fortement ;
– utiliser un nombre trop faible de simulations conduit à sous-estimer nettement la vraisemblance.

Les mêmes observations valent pour le calcul des probabilités d'appartenance à un document, ce qui suggère que l'utilisation de LDA pour projeter des documents dans un espace sémantique

devra s'appuyer sur des méthodes plus efficaces et fiables de calcul de ces projections que celle employée ici.

Caractérisation des classes Dans le modèle LDA, une méthode pour identifier les termes caractéristiques d'un thème consiste à ordonner les mots en fonction du rapport entre leur probabilité dans un thème (donnée par β) et leur probabilité unigramme dans l'ensemble du corpus.

La table 1 liste les mots caractéristiques pour les cinq thèmes construits en estimant les paramètres sur l'ensemble du corpus. Seuls les mots apparaissant plus de 10 fois ont été conservés. Si l'on reconnaît, dans les trois premières colonnes, respectivement les catégories Reuters "Art",

cartier	homer	epicentre	impreza	oshawa
matthau	rookie	lahore	nac	aetna
caballe	fours	fireballs	heerenveen	daimler
glamorous	equaliser	millon	breda	flynn
malraux	starter	evacuation	pts	brca
karmitz	ince	runway	cosworth	erm
coggan	header	skidded	caledonia	reimbursement
archive	tendulkar	planta	denotes	camdessus
vulgar	puck	maces	gf	biotech
lover	strawberry	torrents	summaries	sanofi
murdered	coyotes	acari	tabulated	awote
eastwood	schmeichel	spiritus	prefix	outcomes

TAB. 1 – Mots caractéristiques

"Sport" et "Catastrophes", le quatrième thème est artéfactuel et regroupe un ensemble de courtes dépêches sportives énonçant simplement les résultats d'un événement. La cinquième colonne est plus mélangée, contenant des termes typiques des catégories "Emploi" et "Santé".

6 Conclusion et Perspectives

En conclusion de cette brève étude, il apparaît que :
- le modèle LDA permet de construire des classifications qui sont réellement non-déterministes et qui recoupent, avec une bonne précision, les catégories d'origine. Cet effet est atteint par une modélisation toutefois peu naturelle, qui autorise le changement de thème à chaque mot.
- l'estimation du modèle par échantillonnage de Gibbs est possible, quoique computationnellement lourde, mais les estimateurs résultants dépendent fortement des conditions initiales : l'obtention d'échantillons suffisamment décorrélés demande de relancer l'échantillonneur à partir de plusieurs configurations initiales ;
- le calcul de la vraisemblance d'un document, comme celui de sa projection dans les différentes classes, par Monte Carlo conduit à des résultats qui sont d'autant plus fiables que le nombre de thèmes est petit ; il en va de même pour l'estimation des α_d. L'utilisation de LDA pour construire des représentations multi-dimensionnelles dans des espaces de plus grande dimension exigera d'avoir recours à d'autres stratégies d'estimation.

Ces observations restent à affiner en considérant l'effet de prétraitements plus sophistiqués (en particulier pour réduire la dimension du vocabulaire d'indexation) et en examinant d'autres corpus. Nous envisageons également de poursuivre l'étude expérimentale de ce modèle en

considérant des méthodes alternatives de calcul de la vraisemblance et en examinant la question de la détermination automatique du nombre « idéal » de thèmes, abordée dans Griffiths and Steyvers (2004), ainsi que les diverses extensions de LDA proposées plus récemment eg. dans (Blei et al., 2004; Griffiths et al., 2005; Blei and Lafferty, 2005).

Annexes techniques

Loi de Dirichlet La loi de Dirichlet définit des distributions de probabilité sur le simplexe. Chaque observation multi-dimensionnelle $\alpha = (\alpha_1 \dots \alpha_{n_T})$ vérifie donc : $\sum_{i=1}^{n_T} \alpha_i = 1$. En dimension n_T, cette loi est paramétrisée par un vecteur de n_T paramètres $p = (p_1, \dots, p_{n_T})$. La probabilité d'une observation α est :

$$\mathrm{P}(\alpha|p) = \frac{\Gamma(\sum_{i=1}^{n_T} p_i)}{\prod_{i=1}^{n_T} \Gamma(p_i)} \prod_{i=1}^{n_T} \alpha_i^{p_i - 1},$$

où Γ dénote la fonction Gamma d'Euler. Lorsque tous les p_i sont égaux à λ_α, cette expression se simplifie en :

$$\mathrm{P}(\alpha|\lambda_\alpha) = \frac{\Gamma(n_T \lambda_\alpha)}{\Gamma(\lambda_\alpha)^{n_T}} \prod_{i=1}^{n_T} \alpha_i^{\lambda_\alpha - 1}$$

Détails du calcul de la vraisemblance complète

$$\begin{aligned}
\mathrm{P}(W_d, T_d|\alpha_d, \beta) &= \prod_{i=1}^{l_d} \mathrm{P}(W_{di}, T_{di}|\alpha_d, \beta) \\
&= \prod_{i=1}^{l_d} \mathrm{P}(W_{di}|T_{di}, \alpha_d, \beta)\, \mathrm{P}(T_{di}|\alpha_d, \beta) \\
&= \prod_{i=1}^{l_d} (\beta_{T_{di} W_{di}} \alpha_{d T_{di}}) \\
&= \prod_{t=1}^{n_T} \left(\alpha_{dt}^{N_{td}} \prod_{w=1}^{n_W} \beta_{tw}^{K_{tw}^d} \right)
\end{aligned}$$

Détails du calcul de la vraisemblance conditionnelle

$$
\begin{aligned}
\mathrm{P}(W,T) = & \\
&= \int_\beta \int_\alpha \mathrm{P}(W|T,\beta) p(T,\alpha,\beta) d\alpha d\beta \\
&= \int_\beta \int_\alpha \left(\prod_{d=1}^{n_D} \prod_{i=1}^{l_d} \mathrm{P}(W_{di}|T_{di},\beta_{T_{di}}) \right) \left(\prod_{d=1}^{n_D} \prod_{i=1}^{l_d} \mathrm{P}(T_{di}|\alpha_d) \right) \\
&\quad \left(\prod_{t=1}^{n_T} p(\beta_t|\lambda_\beta) \right) \left(\prod_{d=1}^{n_D} p(\alpha_d|\lambda_\alpha) \right) d\alpha d\beta \\
&= \int_\beta \int_\alpha \left(\prod_{d=1}^{n_D} \prod_{i=1}^{l_d} \beta_{T_{di} W_{di}} \right) \left(\prod_{d=1}^{n_D} \prod_{i=1}^{l_d} \alpha_{d T_{di}} \right) \\
&\quad \left(\prod_{t=1}^{n_T} \frac{\Gamma(\sum_{w=1}^{n_W} \lambda_\beta)}{\prod_{w=1}^{n_W} \Gamma(\lambda_\beta)} \prod_{w=1}^{n_W} \beta_{tw}^{\lambda_\beta - 1} \right) \\
&\quad \left(\prod_{d=1}^{n_D} \frac{\Gamma(\sum_{t=1}^{n_T} \lambda_\alpha)}{\prod_{t=1}^{n_T} \Gamma(\lambda_\alpha)} \prod_{t=1}^{n_T} \alpha_{dt}^{\lambda_\alpha - 1} \right) d\alpha d\beta \\
&= \prod_{t=1}^{n_T} \left(\frac{\Gamma(n_W \lambda_\beta)}{\Gamma(\lambda_\beta)^{n_W}} \int_\beta \prod_{w=1}^{n_W} \beta_{tw}^{K_{tw} + \lambda_\beta - 1} d\beta \right) \\
&\quad \prod_{d=1}^{n_D} \left(\frac{\Gamma(n_T \lambda_\alpha)}{\Gamma(\lambda_\alpha)^{n_T}} \int_\alpha \prod_{t=1}^{n_T} \alpha_{dt}^{N_{dt} + \lambda_\alpha - 1} d\alpha \right) \\
&= \prod_{t=1}^{n_T} \left(\frac{\Gamma(n_W \lambda_\beta)}{\Gamma(\lambda_\beta)^{n_W}} \frac{\prod_{w=1}^{n_W} \Gamma(K_{tw} + \lambda_\beta)}{\Gamma(K_t + n_W \lambda_\beta)} \right) \\
&\quad \prod_{d=1}^{n_D} \left(\frac{\Gamma(n_T \lambda_\alpha)}{\Gamma(\lambda_\alpha)^{n_T}} \frac{\prod_{t=1}^{n_T} \Gamma(N_{dt} + \lambda_\alpha)}{\Gamma(l_d + n_T \lambda_\alpha)} \right)
\end{aligned}
$$

Références

Jean-Paul Benzécri et al. *Pratique de l'analyse des données, tome 3. Linguistique et lexicologie*. Dunod, Paris, 1981.

David Blei and John Lafferty. Correlated topic models. In *Advances in Neural Information Processing Systems (NIPS'18)*, volume 18, Vancouver, Canada, 2005.

David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems (NIPS)*, volume 16, Vancouver, Canada, 2004.

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.

Wray Buntine and Alex Jakulin. Applying discrete PCA in data analysis. In M. Chickering and J. Halpern, editors, *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI'04)*, pages 59–66. AUAI Press, 2004.

John F. Canny. GAP : A factor model for discrete data. In *Proceedings of the ACM Conference on Information Retrieval (SIGIR)*, Sheffield, England, 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.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6) :391–407, 1990.

Thomas L. Griffiths and Mark Steyvers. A probabilistic approach to semantic representation. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 2002.

Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 (supl 1) :5228–5235, 2004.

Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua Tenenbaum. Integratic topics and syntax. In *Proceedings of NIPS*17*, Vancouver, CA, 2005.

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

Thomas Minka and John Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the conference on Uncertainty in Artificial Intelligence (UAI)*, 2002.

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.

Reuters. Reuters corpus, 2000. URL `http://about.reuters.com/researchandstandards/corpus/`.

Loïs Rigouste, Olivier Cappé, and François Yvon. 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.