Modélisation probabiliste de collections textuelles et distributions de mots

Stéphane Clinchant*†, Eric Gaussier†

*Xerox Research Centre Europe
Stephane.Clinchant@xrce.xerox.com
†Laboratoire d'Informatique de Grenoble
Université J. Fourier, Grenoble 1
Eric.Gaussier@imag.fr

Résumé. Nous examinons dans cet article les liens entre modèles probabilistes de documents textuels et observations empiriques sur la distribution des mots au sein d'une collection. Nous proposons une caractérisation formelle de ces observations, et introduisons la distribution beta négative binomiale. Cette distribution (connue sous diverses dénominations mais dont la dérivation que nous proposons est nouvelle) permet de rendre compte des observations empiriques et fournit un modèle non paramétrique dont le bon comportement est validé en catégorisation de textes.

1 Introduction

Plusieurs modèles probabilistes de collections textuelles ont récemment été proposés, dans un cadre apprentissage automatique ou recherche d'information Hofmann (1999); Blei et al. (2003); Buntine et Jakulin (2004); Madsen et al. (2005); Elkan (2006). Ces modèles sont souvent définis d'un point de vue statistique. Au cours des années, cependant, plusieurs observations empiriques sur la façon dont les mots se comportent dans les documents ont été faites (depuis le travail de G. Zipf en 1949 jusqu'à des études plus récentes, comme par exemple celles décrites dans Church et Gale (1995); Katz (1996)). Dans quelle mesure les modèles probabilistes sont en accord avec ces observations empiriques est une question toujours ouverte, et que nous abordons dans cet article. En particulier, nous commençons par passer en revue les observations empiriques sur les distributions de mots, en les caractérisant formellement (section 2). Nous examinerons ensuite quelques-uns des modèles probabilistes les plus populaires au regard de ces caractéristiques formelles (section 3), avant de dériver une loi (BNB) qui rend compte de toutes ces caractéristiques (section 4). La section 5 valide expérimentalement cette nouvelle loi sur une tâche de catégorisation de textes.

2 Observations empiriques sur les distributions des mots dans une collection de documents

Nous passons en revue dans cette section trois observations empiriques majeures sur la façon dont les mots sont distribués dans une collection de documents. La première de ces observations concerne la distribution du nombre de documents, au sein d'une collection donnée, dans lesquels un mot donné apparaît un certain nombre de fois. La deuxième observation concerne le phénomène de *burstiness*[1], pour lequel nous fournissons ici une définition formelle et une propriété caractérisante. Enfin, la troisième observation concerne la loi de Zipf. Une partie des développements présentés ici est reprise de Clinchant et Gaussier (2008).

2.1 Un comportement de type "négatif binomial"

Church et Gale (1995) ont été les premiers, à notre connaissance, à fournir une étude complète du nombre de documents, au sein d'une collection donnée, dans lesquels un mot i apparaît exactement x_i fois (une quantité que nous noterons $\#(d, x_i)$). Leur étude portait sur la comparaison de plusieurs distributions, afin de déterminer celle qui modélisait au mieux la probabilité de $\#(d, x_i)$. Parmi toutes les distributions comparées, la distribution négative binomiale est celle qui ressort le plus favorablement de l'étude, c'est-à-dire celle qui fournit la meilleure adéquation aux données. Parmi les différentes paramétrisations (équivalentes) de la distribution négative binomiale, nous présentons ici celle qui nous semble la plus employée, qui repose sur deux paramètres réels, β et r, avec $0 < \beta < 1$ et $0 < r$, et qui correspond à la distribution :

$$\text{NegBin(x ;r,}\beta) = \frac{\Gamma(r+x)}{x!\Gamma(r)}(1-\beta)^r\beta^x$$

pour $x = 0, 1, 2, \cdots$ (Γ est la fonction gamma). x désigne ici le nombre d'occurrences d'un mot dans un document.

Le bon comportement de la distribution négative binomiale a également été observé dans des travaux plus récents. Rigouste (2006), par exemple, reproduit les expériences décrites dans Church et Gale (1995) sur différentes collections, pour finalement aboutir à la même conclusion : la meilleure adéquation aux données est fournie par la distribution négative binomiale. Ces résultats montrent que $P(\#(d, x_i = n))$ est mieux modélisée par une négative binomiale que par une binomiale, une Poisson ou un mélange de Poisson. Ils n'impliquent pas, bien sûr, que $P(\#(d, x_i = n))$ est distribuée selon une négative binomiale, mais seulement que la validité d'une nouvelle distribution pour la modélisation textuelle doit pouvoir rendre compte de la distribution empirique $P(\#(d, x_i = n))$.

2.2 Le phénomène de *rafale*

Un phénomène important mis en avant dans Church et Gale (1995); Katz (1996) est celui du *comportement en rafale* (en anglais *burstiness*), qui décrit le fait que les mots, dans un document, tendent à apparaître par "paquets". En d'autres termes, une fois que l'on a observé l'occurrence d'un mot dans un document, il est bien plus probable d'observer de nouvelles

[1] Il n'y a pas, à notre connaissance, d'équivalent français pour ce terme. Nous utiliserons *comportement en rafale* pour rendre compte de ce phénomène.

occurrences. La notion de *burstiness* est identique à celle d'*effet de post-échantillonnage*[2], décrite par exemple dans Feller (1968), et qui se traduit par le fait que plus un mot est observé dans un document, plus on a de chances de l'observer par la suite. Plusieurs modèles ont tenté de rendre compte de ces phénomènes. En revanche, peu de descriptions opérationnelles de ce phénomène (c'est-à-dire permettant de qualifier une distribution de probabilité au regard du comportement en rafale) existent.

Une des premières descriptions a été proposée dans Church et Gale (1995), et repose sur la quantité :

$$B_P = \frac{E_P[x_i]}{P(x_i \geq 1)}$$

où x_i désigne le nombre d'occurrences du mot i, et E_P l'espérance par rapport à la distribution P. Si cette mesure fournit bien une méthode pour comparer deux distributions de mots, elle ne permet pas, en revanche, de caractériser directement la distribution P. Pour ce faire, nous avons récemment introduit (Clinchant et Gaussier (2008)) la définition suivante :

Définition 1 *Nous disons qu'un mot i est en rafale au niveau n_0 sous une distribution P ssi il existe un entier $n_0, 1 \leq n_0$, tel que pour tout couple d'entiers $(n', n), n' \geq n \geq n_0$:*

$$P(x_i \geq n' + 1 | x_i \geq n') \geq P(x_i \geq n + 1 | x_i \geq n)$$

Cette définition traduit directement le fait qu'un mot est *en rafale* s'il est plus facile de le générer une fois qu'il a été généré un certain nombre de fois. L'introduction d'un niveau dans cette définition permet de rendre compte de phénomènes plus fins, où le même mot peut être *en rafale* dans un contexte et pas dans d'autres. En pratique toutefois, il n'est pas toujours facile de calculer $P(x_i \geq n+1 | x_i \geq n)$ et donc de déterminer si une distribution rend compte ou non du phénomène de rafale, sur la base de la définition précédente. La propriété suivante (dont la démonstration est donnée en annexe) permet de résoudre en partie ce problème :

Propriété 2 Caractérisation du comportement en rafale
Soit $P(x_i)$ une distribution pour le mot i, et soit $a_n = \frac{P(x_i=n+1)}{P(x_i=n)}$.

(i) *S'il existe n_o tel que la suite a_n est croissante à partir du rang n_0, alors i est en rafale (au niveau n_0) sous P.*

(ii) *S'il existe n_o tel que la suite a_n est décroissante à partir du rang n_0, alors i n'est pas en rafale sous P.*

Cette propriété permet de caractériser un certain nombre de distributions. Ainsi, pour les distributions univariées et discrètes standard nous avons :

• $P(x_i) =$ **Binomial**(N, p_i)

$$P(x_i = n) = \binom{N}{n} p_i^n (1 - p_i)^{N-n}$$

$$\forall n, a_n = \frac{(N-n)p_i}{(n+1)(1-p_i)}$$

a_n est donc strictement décroissante, ce qui montre que la distribution binomiale ne rend pas compte du phénomène de rafale, une conclusion établie par ailleurs dans Elkan (2006).

[2]Nous utilisons ici ce terme pour traduire le terme anglais *aftereffect of future sampling*.

•$P(x_i) =$ **Poisson**(λ_i)

$$P(x_i = n) = e^{-\lambda_i} \frac{\lambda_i^n}{n!}$$

$$\forall n, a_n = \frac{\lambda_i}{n+1}$$

qui est strictement décroissante pour toutes les valeurs de λ_i. La distribution de Poisson ne permet donc pas de rendre compte du phénomène de rafale.

•$P(x_i) =$ **Geometric**(p_i) Nous obtenons :

$$P(x_i = n) = p_i(1 - p_i)^n$$

$$\forall n, a_n = (1 - p_i)$$

a_n est donc constante. La distribution géométrique est donc neutre par rapport au phénomène de rafale.

•$P(x_i) =$ **NegBin**(r_i, β_i)

$$\forall n, a_n = \frac{\beta_i(r_i + n)}{n+1}$$

a_n est strictement croissante *ssi* $r_i < 1$, strictement décroissante *ssi* $r_i > 1$ et constante sinon. Ceci montre que la distribution négative binomiale peut rendre compte à la fois des mots *en rafale* et des mots qui ne le sont pas, suivant la valeur du paramètre r[3].

2.3 La loi de Zipf

La loi de Zipf est certainement une des lois de distribution de mots les plus connues. Faisant l'hypothèse que les mots sont ordonnés dans une liste par nombre d'occurrences décroissant, et en notant z_i le rang du mot i dans cette liste, la loi de Zipf (Zipf (1949)) établit que : $x_i = \frac{C}{z_i^a}$, où a est un paramètre du modèle, et C une constante de normalisation[4].

Malgré de nombreuses observations empiriques la concernant, la loi de Zipf est rarement citée dans les travaux sur la modélisation probabiliste de documents. Ce fait n'est pas forcément surprenant, dans la mesure où la loi de Zipf ne modélise pas directement le nombre d'occurrences d'un mot dans un document ou une collection. Toutefois, la loi de Zipf a un certain nombre d'implications sur la distribution du nombre d'occurrences d'un mot dans un document (ou une collection) qu'il est bon de conidérer. En particulier, on peut montrer (cf. par exemple Baayen (2001)) que la loi de Zipf implique que le nombre de types (i.e. de mots diffférents) qui apparaissent exactement m fois dans un échantillon de L mots et M types[5], nombre noté $V(m, L)$, vaut :

$$V(m, L) = \frac{M}{m(m+1)}$$

[3]Une analogie intéressante peut être établie ici avec la biologie. Ainsi, Anscombe (1948) mentionne que le paramètre de forme d'une distribution négative binomiale, c'est-à-dre r, dépend de la capacité intrinsèque d'une espèce à se reproduire.

[4]Il est à noter que depuis l'élaboration de cette loi, plusieurs tentatives ont été menées pour l'expliquer et l'améliorer (cf. par exemple Baayen (2001) pour une étude détaillée de ces tentatives).

[5]La phrase *"le chat mange le gâteau"*, par exemple, contient cinq mots et quatre types.

La mesure ci-dessus peut être généralisée à travers la quantité :

$$\alpha(m, L) = \frac{E[V(m, L)]}{M}$$

(E désigne l'espérance). Pour la loi de Zipf :

$$\alpha(m, L) = \frac{1}{m(m+1)} \tag{1}$$

Dans la lignée des travaux de Orlov Orlov et Chitashvili (1983), il est possible de généraliser l'équation précédente en considérant la famille généralisée des lois de Zipf, qui est définie par :

$$\alpha(m, L) = \frac{\int_0^{+\infty} \frac{(log(1+x))^{\gamma-1} x^{\delta}}{(1+x)^{m+\beta+1}} dx}{\int_0^{+\infty} \frac{(log(1+x))^{\gamma-1} x^{\delta-1}}{(1+x)^{\beta+1}} dx} \tag{2}$$

où δ, β et γ sont des paramètres qui dépendent de la distribution considérée (ainsi $\delta = \beta = \gamma = 1$ fournit l'équation 1).

Les développements précédents montrent que les distributions Zipfiennes ont des caractéristiques particulières, traduites par l'équation 2. Ceci nous conduit finalement à proposer la définition suivante pour la compatibilité d'une loi donnée avec la loi de Zipf :

Définition 3 [Compatibilité avec la loi de Zipf]
Nous dirons qu'une distribution de mots P est compatible avec la loi de Zipf si elle vérifie l'équation 2.

3 Modèles probabilistes de documents

Nous voulons, dans cette section, fournir un rapide panorama des prinicpaux modèles génératifs de documents proposés au cours de ces dernières années.

La plupart des modèles probabilistes de textes reposent soit sur une distribution de Poisson (ou des mélanges de Poisson), soit sur une distribution multinomiale. Le modèle probabiliste Okapi et le modèle de langue (Ponte et Croft (1998)) illustrent bien l'utilisation de ces deux types de distribution. Toutefois, dans la mesure où une distribution multinomiale sur un vocabulaire se réduit à une distribution binomiale sur chaque mot individuel, aucune de ces distributions n'est en accord avec les observations empiriques, en particulier avec le comportement en rafale[6]. Il en va de même pour bon nombre de modèles développés lors de cette dernière décennie. Ainsi, un des modèles les plus populaires en classification supervisée est le modèle Bayes naîf construit sur une distribution multinomiale (cf. Mccallum et Nigam (1998)). Ce modèle ne permet pas d'expliquer le comportement en rafale des mots dans une collection. Un autre problème lié au modèle Bayes naïf est sa tendance à n'affecter un document qu'à une seule classe ou thème, ce qui va à l'encontre de la polythématicité inhérente aux documents textuels. Ce fait a conduit au développement de *PLSA* (*Probabilistic Latent Semantic Analysis*, Hofmann (1999)), qui permet une afffectation souple des mots et des documents aux classes.

[6] Il faut cependant noter que le lissage utilisé dans l'approche modèle de langue modifie le comportement de la multinomiale sous-jacente. Notre observation ne tient pas compte de cette modification.

Toutefois, ce modèle reposant également sur une hypothèse multinomiale, le comportement en rafale n'est toujours pas pris en compte. De plus, *PLSA* introduit des variables fictives qui le rendent en partie déficient d'un point de vue statistique, le modèle n'étant pas "génératif" au sens strict du terme. Ce défaut génératif a conduit au développement du modèle *LDA* (*Latent Dirichlet Analysis*, Blei et al. (2003)), qui accorde, comme ses prédécesseurs, une place centrale à la distribution multinomiale. Ces différents modèles sont en fait fortement reliés, et peuvent également être rapprochés d'autres modèles introduits dans d'autres communautés. Ainsi, Girolami et Kabán (2003) montre une relation formelle entre *PLSA* et *LDA*, alors que Gaussier et Goutte (2005) présente une équivalence entre *PLSA* et *NMF* (*Non-negative Matrix factorization*, Lee et Seung (1999, 2000)). Buntine et Jakulin (2004) propose une vue générale de ces modèles sous la forme d'une analyse en composante principale discrète.

Plus récemment, la volonté de prendre en compte le comportement en rafale des mots a conduit au développement du modèle *DCM* (*Dirichlet Compound Multinomial*), tout d'abord introduit dans Minka (2003), puis repris dans Madsen et al. (2005) et étendu dans Elkan (2006) sous le nom *EDCM*, puis au développement du modèle *SD* (*Smoothed Dirichlet*, Nallapati et al. (2006)). La bonne adéquation des modèles *DCM* et *EDCM* au comportement en rafale est validée, dans les travaux cités, expérimentalement. Nous donnons ici, sur la base des notions introduites dans la section précédente, une preuve formelle de cette adéquation pour le modèle *EDCM* (la preuve est identique pour les modèles *DCM* et *SD*).

Le modèle est défini par :

$$Q(d|\beta) = n! \frac{\Gamma(s)}{\Gamma(s+n)} \prod_{i:x_i^d \geq 1} \frac{\beta_i}{x_i^d}$$

où n est la longueur du document d, x_i^d désigne le nombre d'occurrences du mot i dans le document d et $s = \sum_{i:x_i^d \geq 1} \beta_i$. Comme noté dans Elkan (2006), la quantité ci-dessus ne définit pas une distribution de probabilité (d'où la notation Q), mais peut être normalisée à cette fin. La marginalisation de Q fournit :

$$P(x_i^d|\beta) \propto \frac{\Gamma(s - \beta_i + n - x_i^d)}{\Gamma(s - \beta_i)} \frac{\beta_i}{x_i^d}$$

et donc :

$$\begin{aligned} \frac{a_{k+1}}{a_k} &= \left(\frac{k^2 + 2k + 1}{k^2 + 2k}\right) \times \\ &\quad \left(\frac{s - \beta_i + n - k - 1}{s - \beta_i + n - k - 2}\right) > 1 \end{aligned}$$

ce qui montre la bonne adéquation de *EDCM* pour modéliser le comportement en rafale. Toutefois, l'adéquation de ces trois modèles (*DCM, EDCM, SD*) au comportement négatif binomial et leur compatibilité avec la loi de Zipf restent des questions ouvertes.

Un autre modèle récent introduit en recherche d'information et visant à rendre compte de l'effet de post-échantillonnage, est le modèle *DFR* (*Divergence from Randomness*, Amati et Rijsbergen (2002)), qui utilise la loi de succession de Laplace (appelée *normalisation L* dans la terminologie *DFR*), ainsi qu'un ratio de processus de Bernouilli (appelé *normalization B*).

Dans le cas de la loi de Laplace, la quantité a_n vaut :

$$a_n = \frac{n}{(n+1)}$$

et est donc croissante. On voit donc ici formellement que la loi de Laplace rend bien compte du comportement en rafale. Pour le normalisation B, a_n est définie par :

$$a_n = 1 - \frac{F+1}{D(n+1)}$$

où F et D sont respectivement définies dans Amati et Rijsbergen (2002) comme le nombre total d'occurrences du mot et le nombre de documents dans lesquels le mot apparaît. a_n étant croissante, on peut conclure à la bonne adéquation de la normalisation *B* par rapport au comportement en rafale. Ici encore, l'adéquation aux autres observations empiriques reste à déterminer.

Nous voulons maintenant introduire une distribution qui rend bien compte de l'ensemble des observations empiriques.

4 La distribution Beta Negative Binomiale

Un désavantage de la distribution négative binomiale réside dans le fait que les estimateurs du maximum de vraisemblance de ses paramètres n'existent que lorsqu'on dispose de plusieurs observations. En d'autres termes, on ne peut estimer β et r par maximum de vraisemblance pour un seul document, ce qui rend cette distribution inadaptée aux modèles de langue (Ponte et Croft (1998)) ou modèle de divergence (Amati et Rijsbergen (2002)).

Une extension intéressante, toutefois, de la négative binomiale consiste à considérer le paramètre β comme issu d'une loi Beta. Dans ce cas, la distribution résultante a la forme :

$$BNBGen(x; r, a, b) = \frac{\Gamma(r+x)\Gamma(a+x)}{x!\Gamma(r)\Gamma(a)\Gamma(b)} \times \frac{\Gamma(a+b)\Gamma(r+b)}{\Gamma(a+b+r+x)} \tag{3}$$

où $x = 0, 1, 2, \cdots$, et a et b représentent les deux paramètres de la loi Beta. En faisant l'hypothèse que cette loi est uniforme (i.e. $a = b = 1$), on obtient une distribution à un paramètre que nous désignerons dans la suite par **Beta négative binomiale BNB**[7] :

$$BNB(x; r) = \frac{r}{(r+x+1)(r+x)} \tag{4}$$

Comme précédemment, cette distribution est définie pour $x = 0, 1, 2, \cdots$. La figure 1 montre les distributions de probabilité obtenues pour plusieurs valeurs du paramètre. Comme nous allons le voir, l'estimateur du maximum de vraisemblance existe cette fois même en présence d'un seul document.

Nous examinons maintenant la bonne adéquation de la BNB aux observations empiriques décrites dans la section précédente.

[7] Cette distribution est parfois appelée distribution de Johnson car elle est étudiée dans Johnson et al. (1993). Elle correspond aussi à la loi de Yule-Simon (Baayen (2001)).

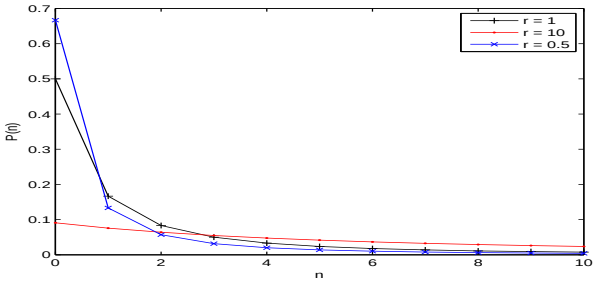

FIG. 1 – *Distribution de probabilité pour la BNB pour plusieurs valeurs de r.*

1. **Comportement négatif binomial**
   Ce comportement est étudié plus en détail dans la section 5. Nous voulons juste mentionner ici que la BNB, construite sur la négative binomiale, fournit une bonne adéquation aux données.
2. **Comportement en rafale**
   Pour la distribution BNB, $a_n = \frac{r+n}{r+n+2}$ est strictement croissante. La distribution BNB modélise donc le comportement en rafale et l'effet de post-échantillonnage. Plus généralement, la famille de dsitributions donnée par l'équation 3 peut être utilisée pour rendre compte de ces phénomènes si le paramètre r vérifie $0 < r \leq 1$.
3. **Compatibilité avec la loi de Zipf** La BNB est compatible avec la loi de Zipf au sens de la définition 3 donnée dans la section précédente. En effet, rappelons que cette compatibilité est fondée sur l'équation :

$$\alpha(m, L) = \frac{\int_0^{+\infty} \frac{(log(1+x))^{\gamma-1} x^{\delta}}{(1+x)^{m+\beta+1}} dx}{\int_0^{+\infty} \frac{(log(1+x))^{\gamma-1} x^{\delta-1}}{(1+x)^{\beta+1}} dx}$$

   En choisissant : $\delta = \gamma = 1$ et $\beta = r$, nous obtenons l'équation 4 de la BNB (nous ne donnons pas les détails, techniques, de cette dérivation ici ; le lecteur intéressé peut se reporter à Baayen (2001) par exemple).

4.1 Estimation des paramètres

Nous faisons l'hypothèse ici que chaque mot $i (1 \leq i \leq M)$ d'une collection de N documents est modélisé, indépendamment des autres mots, par une distribution BNB de paramètres r_i. Comme précédemment, le nombre d'occurrences du mot i dans le document d sera noté x_i^d. L'estimateur du maximum de vraisemblance pour chaque r_i est défini par :

$$\hat{r}_i = \text{argmax}_{r_i} L(r_i) = \text{argmax}_{r_i} \prod_d \frac{r_i}{(r + x_i^d)(r_i + x_i^d + 1)}$$

Or :

$$\frac{\partial \log L}{\partial r_i} = \sum_d \frac{1}{r_i} - \frac{1}{r_i + x_i^d} - \frac{1}{r_i + x_i^d + 1}$$

L'annulation de cette dérivée fournit :

$$r_i = \frac{1}{N} \frac{1}{\sum_d \frac{1}{r_i + x_i^d} - \frac{1}{r_i + x_i^d + 1}}$$

qui définit une équation du point fixe pour r_i. De plus, dans le cas d'un seul document, ou dans le cas où $\forall d, x_i^d = x_i$, la solution de l'équation ci-dessus se réduit à :

$$r_i = \sqrt{x_i(x_i + 1)} \tag{5}$$

4.2 Relations avec d'autres distributions

La probabilité de présence d'un mot dans un document sous le modèle BNB est donnée par :

$$P_{\text{BNB}}(x_i \geq 1 |\, r_i) = \frac{r_i}{r_i + 1}$$

qui, dans le cas où $r_i = x_i$, correspond à la loi de succession de Laplace. On peut donc réinterpréter la quantité Prob_2 de la normalisation L de Amati et Rijsbergen (2002) comme la probabilité de présence du mot dans un document, sous un modèle BNB dont le paramètre est réglé à x_i. Ce réglage est en fait très proche de la valeur obtenue par maximum de vraisemblance. En effet, un développement de Taylor de l'équation 5 fournit :

$$r_i \sim x_i(1 + \frac{1}{2x_i}) = x_i + 0.5$$

Ainsi, pour x_i suffisamment grand, $r_i \sim x_i$ (l'erreur d'approximation est de l'ordre de 5% pour $x_i = 10$).

Soit F_i le nombre total d'occurrences du mot i dans la collection : $F_i = \sum_d x_i^d$. En prenant $r_i = \frac{F_i}{N}$ (i.e. le nombre d'occurrences de i obtenu en considérant que ce mot est distribué aléatoirement dans la collection), on obtient :

$$P_{\text{BNB}}(x_i \geq 1 | r_i) = \frac{F}{F + N} \approx \frac{F}{N}, \text{ for } \frac{F}{N} \text{ small or moderate}$$

qui est la quantité Prob_1 utilisée dans le modèle *tf-itf I(F)* de Amati et Rijsbergen (2002). On voit donc ici que le modèle *tf-itf I(F)* au complet (c'est-à-dire pour les deux quantités Prob_1 et Prob_2) s'interprète simplement à partir de distributions BNB.

Enfin, sur la base de la représentation : $P(d) \propto \prod_{1 \leq i \leq M} P(x_i \geq 1)^{x_i^d}$, dans le contexte de la catégorisation, la règle de décision pour un classifieur probabiliste est d'affecter le document d à la catégorie c qui maximise :

$$Q(d, c) = \log(P(c)) + \sum_{1 \leq i \leq M} x_i^d \log(P(x_i \geq 1 | c))$$

où $P(x_i \geq 1 | c)$ désigne la probabilité de présence du mot i dans la catégorie c. Dans l'hypothèse où cette probabilité est fournie par un modèle BNB dont le paramètre r_i^c est pris égal à $\frac{x_i^c}{l_c}$ (c'est-à-dire au nombre d'occurrences de i dans c, normalisé par la longueur de c), nous obtenons :

$$Q(d, c) = \log(P(c)) + \sum_{1 \leq i \leq M} x_i^d (\log(\frac{x_i^c}{l_c}) - \log(\frac{x_i^c}{l_c} + 1))$$

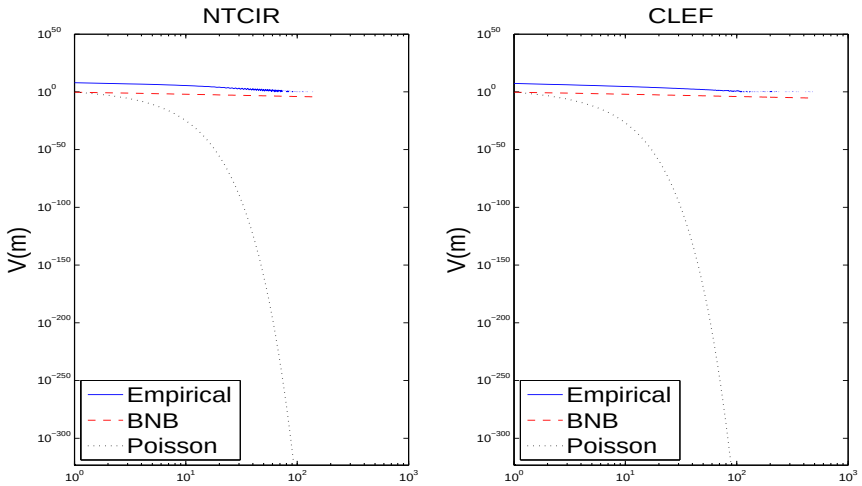

FIG. 2 – *Courbe $(m, V(m))$ en échelle logarithimique pour les corpus CLEF 2003 et NTCIR-PAJ.*

En général, $x_i^c < l_c$, de telle sorte que : $Q(d,c) \sim \log(P(c)) + \sum_{1 \leq i \leq M} x_i^d \log(\frac{x_i^c}{l_c})$, qui est exactement la fonction utilisée par un classifieur de type Bayes naïf multinomial dont les paramètres sont estimés par maximum de vraisemblance (cf. par exemple Mccallum et Nigam (1998)). Ainsi, le modèle Bayes naïf multinomial peut être approché par un modèle fondé sur des distributions de mots de type BNB, dont les paramètres sont égaux aux nombrex d'occurrences des mots dans la catégorie, normalisés par la longueur de la catégorie.

5 Validation expérimentale

Dans un premier temps, nous étudierons le comportement négatif binomial de la BNB. Nous verrons que la BNB modélise mieux les données que la distribution de Poisson. Dans un deuxième temps, nous evaluerons le comportement de distributions BNB pour la classification de textes. Dans ce dernier cadre, nous nous concentrerons sur des modèles probabilistes (modèle multinomial et modèle Dirichlet lissé), d'une part car certains de ces modèles fournissent des résultats parmi les meilleurs en catégorisation de textes (Nallapati et al. (2006)), et d'autre part car c'est le cadre des modèles probabilistes qui nous intéresse ici.

5.1 Comportement négatif binomial

Les courbes des figures 2 montrent la distribution empirique de $V(m)$, celle prédite par une BNB et une Poisson sur deux corpus. Le premier corpus est le corpus CLEF 2003, composé de 150000 documents environ. Le deuxième est le corpus NTCIR composé de 2 millions de documents. Pour rappel, $V(m)$ est le nombre de types qui apparaissent exactement m fois dans la collection. Les paramètres des distributions sont définis, en prenant $\lambda_i = \frac{F_i}{l_c}$ (Poisson) et

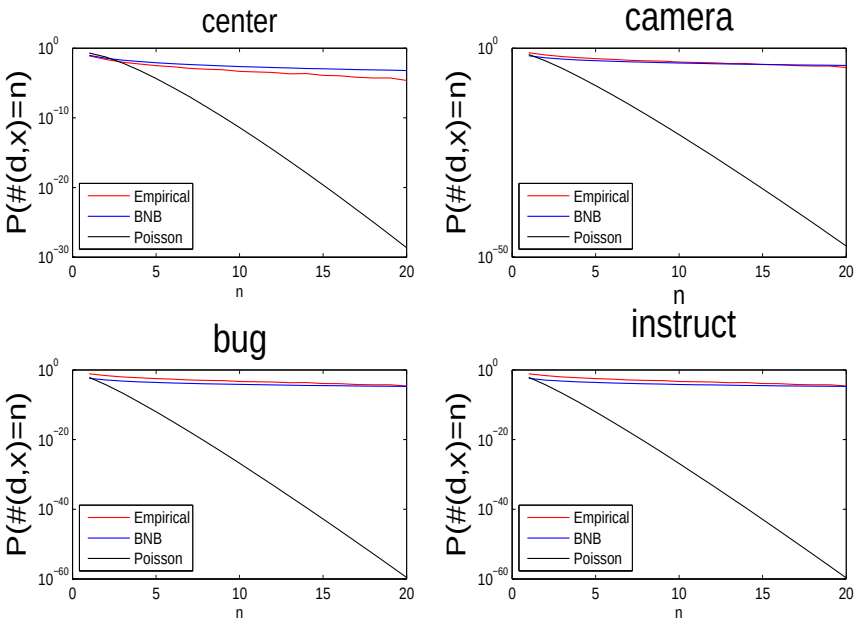

FIG. 3 – *Courbe* $(m, V(m))$ *pour quatre termes du corpus CLEF 2003.*

$r_i = \frac{F_i}{l_c}$ (BNB). Nous utilisons ici la distribution de Poisson car elle approche la binomiale pour des echantillons de taille importante. D'une manière générale, les 3 courbes sont proches pour les basses fréquences (m petit), ce qui indique que les distributions BNB et Poisson sont aptes à modéliser correctement cette gamme de fréquences. Cependant, plus la fréquence augmente et plus la loi de Poisson diverge de la distribution empirique. Au contraire, la BNB reste proche de la distribution empirique sur toute la gamme de fréquences.

La figure 3 montre la distribution du nombre de document qui ont m occurences d'un mot en particulier, pour 4 mots différents. Les distributions sont estimées de la même manière que précédemment. Le mot $center$ est le 95ieme mot le plus fréquent du corpus avec 44665 occurrences dans 21448 documents. Le mot $camera$ est un peu moins fréquent avec un rang de 1500 et et 5140 occurrences dans 3306 documents. Les deux autres mots, bug et $instruct$, ont un rang d'environ 4000 avec une fréquence totale de 1250 environ, bug apparaissant dans un peu moins de documents (850) que $instruct$ (1151). Les courbes révèlent le comportement déjà remarqué, à savoir : une modélisation correcte pour les basses frequences pour les lois Poisson et BNB, et une meilleure modelisation pour la BNB sur des gammes de fréquences plus élevées.

5.2 Catégorisation de textes

Nous nous intéressons dans cette section à l'utilisation d'un modèle fondé sur la distribution BNB pour la catégorisation de documents, modèle que nous comparons à d'autres plus clas-

siques. De manière générale, chaque catéorie est représentéée par un vecteur de paramètres, et un nouveau document est affecté à la catégorie qui maximise une fonction de type $Q(d, c)$. Nous expliciterons l'estimation des paramètres ainsi que la fonction Q pour chacun des modèles. Dans la suite (et comme précédemment), d désigne un document, c une catégorie, l_d la longueur d'un document et l_c la longueur d'une catégorie. $\mathcal{D}$ désigne la collection de documents et $l_\mathcal{D}$ sa longueur, N le nombre de document dans la collection. Enfin, x_i^d désigne le nombre d'occurrences du mot i dans le document d et $F_i = \sum_{d \in \mathcal{D}} x_i^d$.

5.2.1 Modèle multinomial

Ce modèle est fondé sur la décomposition suivante :

$$P(d|c) = P(d = (x_1^d, \ldots, x_n^d)|c) = \frac{l_d!}{\prod_{i=1}^n x_i!} \prod_{i=1}^n p_{c,i}^{x_i^d}$$

Les paramètres $p_{c,i}$ sont estimés par maximum de vraisemblance, ce qui fournit :

$$p_{c,i} = \frac{\sum_{d \in c} x_i^d}{\sum_{i=1}^M \sum_{d \in c} x_i^d}$$

En pratique, ces paramètres sont lissés afin d'en avoir une meilleure estimation. Nous utilisons dans ce travail un lissage de type Dirichlet, défini par :

$$p_{c,i} = \frac{\sum_{d \in c} x_i^d + \mu p_{g,i}}{\sum_{i=1}^M \sum_{d \in c} x_i^d + \mu} \tag{6}$$

où $p_{g,i} = \frac{F_i}{l_\mathcal{D}}$ et μ est le coefficient de lissage. Ce coefficient est estimé automatiquement sur chaque ensemble d'apprentissage en optimisant la *vraisemblance moins 1* suivant l'approche proposé par Zhai et Lafferty (2004). Il s'agit en fait de chercher la valeur de μ qui maximise la fonction l_{-1} ci-dessous, qui possède un unique maximum :

$$l_{-1}(\mu) = \sum_d \sum_{1 \leq i \leq M} x_i^d \log(\frac{x_i^d - 1 + \mu p_{g,i}}{l_d - 1 + \mu})$$

Enfin, la fonction Q pour ce modèle est définie à partir de $\log(P(c|d))$, sans tenir compte des termes constants, ce qui donne :

$$Q(d, c) = \log(P(c)) + \sum_{1 \leq i \leq M} x_i^d \log(p_{c,i})$$

où $p_{c,i}$ est donné par l'équation 6 et $P(c)$ correspond à la proportion de documents de la collection affectés à c.

5.2.2 Modèle Dirichlet lissé

Le deuxième modèle que nous considérons ici est fondé sur une distribution de Dirichlet. Ce modèle est décrit dans Nallapati et al. (2006).

Introduisons tout d'abord la distribution lissée $p^s_{d,i}$, representant la distribution lissée du mot i dans le document d :

$$p^s_{d,i} = \lambda \frac{x^d_i}{\sum_i x^d_i} + (1-\lambda) p_{g,i} = \lambda p_{d,i} + (1-\lambda) p_{g,i}$$

où $p_{d,i} = \frac{x^d_i}{\sum_i x^d_i}$ est la distribution empirique du mot dans le document. Le paramètre λ est le paramètre de lissage global, comme μ l'était pour la multinomiale.

La distribution de Dirichlet permet de modéliser des données multivariées sur un simplexe, et peut être utilisée ici pour modéliser la distribution d'un document dans une catégorie :

$$P(p^s_{d,1}, \ldots, p^s_{d,M}|c) = \frac{\Gamma(\sum_{i=1}^M \alpha^c_i)}{\prod_{i=1}^M \Gamma(\alpha^c_i)} \prod_{i=1}^M (p^s_{d,i})^{\alpha^c_i - 1}$$

Les paramètres α^c_i sont donnés par (Nallapati et al. (2006)) :

$$\log(\alpha^c_i) = \frac{1}{|l_c|} \sum_{d \in c} \log(p^s_{d,i})$$

Cette équation peut encore s'écrire sous la forme :

$$\log(\alpha^c_i) = \frac{1}{|l_c|} \left(\log(1-\lambda) + \log(p_{g,i}) + \sum_{d \in c} \log(1 + \frac{\lambda p_{d,i}}{(1-\lambda) p_{g,i}}) \right)$$

Afin de mieux comparer des catégories de longueur différentes, les paramètres α^c_i sont renormalisés pour sommer à 1 ($\sum_{i=1}^M \alpha^c_i = 1$), ce qui fournit une distribution de probabilité. Enfin, la fonction Q de ce modèle est définie à partir de l'entropie croisée entre la distribution α^c et le modèle de langue lissé d'un document :

$$Q(d,c) = \sum_{1 \leq i \leq M} \alpha^c_i \log(p^s_{d,i})$$

ce qui est égal, à une constante près, à :

$$Q(d,c) = \sum_{1 \leq i \leq M} \alpha^c_i \log(1 + \frac{\lambda}{1-\lambda} \frac{p_{d,i}}{p_{g,i}})$$

Le paramètre de ce modèle Dirichlet lissé est le coefficient λ. La valeur de λ a été fixé a 0.0001 dans nos experiences. Cette valeur est une valeur optimale pour certains corpus, comme observé par Nallapati et al. (2006).

5.3 BNB

Le dernier modèle que nous considérons est fondé sur des distributions BNB (equation 4). Nous adoptons ici le cadre *DFR* de Amati et Rijsbergen (2002) qui repose sur une fonction Q de la forme :

$$Q(d,c) = - \sum_{1 \leq i \leq M} \hat{P}(x_i|d) \log(P(x_{c,i}|r_{g,i}))$$

Cette fonction représente l'espérance, sur le document, de la quantité $-\log(P(x_{c,i}|r_{g,i}))$ qui correspond à l'information apportée par l'observation de $x_{c,i}$ occurrences du mot i dans la catégorie c par rapport à la distribution globale (sur toute les catégories) de i. En particulier, la quantité $P(x_{c,i}|r_{g,i})$ représente la probabilité d'observer $x_{c,i}$ occurrences du mot i dans la catégorie c suivant une loi BNB dont le paramètre est appris sur l'ensemble des catégories. Plus $x_{c,i}$ diffère du nombre d'occurrences moyen de i sur la collection, plus cette quantité est faible et plus l'information qu'apporte i est importante.

Le paramètre $r_{g,i}$ est défini : $r_{g,i} = \frac{F_i}{N}$.De plus, afin de normaliser les occurrences entre les catégories (certaines catégories qui contiennent plus de documents que d'autres "tirent" les nombres d'occurrences des mots vers le haut), nous utilisons la normalisation logarithmique suivante, adoptée dans plusieurs modèles de recherche d'information :

$$\hat{x}_{c,i} = x_{c,i} \log(1 + \frac{\text{avg}_l}{\sum_i x_{c,i}})$$

où avg_l correspond à la longueur moyenne d'une catégorie. En utilisant cette normalisation et la définition de la loi BNB (cf. équation 4), nous obtenons finalement :

$$Q(d,c) = -\sum_{1 \leq i \leq M} \frac{x_i^d}{l_d} \log(\frac{r_{g,i}}{(r_{g,i} + \hat{x}_{c,i})(r_{g,i} + \hat{x}_{c,i} + 1)})$$

5.3.1 Résultats

Nous avons utilisé trois corpus pour nos expériences de catégorisation : *Industry Sector*, *20Newsgroup* et *WebKb*[8]. Ces trois corpus ont suivi le même prétraitement : indexation par Lemur[9] avec filtrage des mots vides. Le corpus Industry est composé d'environ 10000 documents et comporte 104 classes équilibrées en taille. Le corpus 20Newsgroup est organisé en 20 categories d'environ 1000 documents chacune. Enfin le corpus Webkb, comporte environ 8000 documents et 7 catégories. Chaque corpus a été divisé en deux parties, une réservée à l'apprentissage des modèles, l'autre au test. Pour 20Newsgroup, la séparation a été de 80/20 (80% des documents pour l'apprentissage et 20% pour le test) et de 50/50 pour les deux autres corpus. Nous utilisons deux mesures de performance : la précision, qui correspond au rapport du nombre de dcouments bien catégorisés par un modèle sur le nombre de documents catégorisés (cette mesure correspond donc à l'*accuracy*) et la MAP (*Mean Average Precision*) qui correspond à la moyenne, sur toutes les catégories, de la précision moyenne aux différents points de rappel. Cette dernière mesure, inspirée de la recherche d'information, permet d'évaluer dans quelle mesure un système "classe" des documents corrects en tête de liste, c'est-à-dire dans quelle mesure un modèle fournit de bons scores pour des documents corrects, et de moins bons scores pour des documents incorrects. Pour chaque catégorie, les documents affectés à cette catégorie (soit n le nombre de ces documents) sont triés en fonction du score donné par le modèle. La précision moyenne aux différents points de rappel est alors donnée par :

$$\frac{1}{n} \sum_{r=1}^{n} \text{Prec}(r) \times \text{rel}(r)$$

[8] Ces trois corpus sont disponibles sur internet

[9] http ://www.lemurproject.org/lemur/

	20NewsGroup	Industry	WebKB
Multinomial	0.87	0.77	0.51
Dirichlet Lissé	**0.89**	**0.87**	0.49
BNB	0.88	0.77	**0.66**
tf-itf	0.84	0.81	0.52

TAB. 1 – *Précision, moyennée sur 5 splits, des différent modèles (meilleurs resultats en gras).*

	20NewsGroup	Industry	WebKB
Multinomial	0.92	0.82	0.64
Dirichlet Lissé	**0.94**	**0.91**	0.69
BNB	0.93	0.82	**0.80**
tf-itf	0.90	0.85	0.70

TAB. 2 – *Mean Average Precision, moyennée sur 5 splits, pour les différents modèles (meilleurs resultats en gras).*

où r désigne le rang dans la liste des documents, Prec(r) la précision de la liste de documents jusqu'au rang r et où rel(r) est une fonction qui vaut 1 si le document au rang r appartient à la catégorie et 0 sinon. La MAP est obtenue en prenant la moyenne sur les catégories de cette quantité.

Les tableaux 1 et 2 montre les performances des modèles sur les 3 corpus considérés. Nous montrons aussi les résulats d'un autre modèle DFR *tf-itf* (cf Amati et Rijsbergen (2002)) afin d'illustrer le fait que la difference de performance ne s'explique pas uniquement par la famille de modèles considérés. D'une manière génerale, les modèles Dirichlet lissé et BNB tendent à produire des résultats meilleurs que le modèle multinomial et que le modèle *tf-itf*, même si la différence est moins marquée dans ce dernier cas. Sur le corpus Industry, le modèle BNB est très en dessous du modèle Dirichlet lissé, alors qu'il est très en dessus sur le corpus WebKB et que les performances sont équivalentes sur le corpus 20Newsgroup. Une différence fondamentale existe cependant entre ces deux modèles, liée au fait que le modèle Dirichlet lissé repose sur un paramètre qui doit être réglé sur chaque collection, alors que le modèle BNB ne fait appel à aucun paramètre et est plus robuste de ce point de vue. Nous illustrons ce comportement dans ce qui suit.

5.4 Paramétrique vs. non-paramétrique

Nous voulons mettre en évidence que les modèles Dirichlet lissé et multinomiaux subissent des variations de performance importantes en fonction de leur paramètre. La figure 4 présente l'évolution de la MAP pour le modèle multinomial sur le corpus Reuters, et l'evolution des performances sur 20newsgroup en fonction de λ pour le modèle Dirichlet lissé. Comme on peut le constater, les résultats obtenus varient énormément en fonction des paramètres sous-jacents au modèle, avec des performances parfois au-dessus et parfois en-dessous de celles du modèle BNB. Certes, les paramètres de ces modèles peuvent être appris automatiquement, par validation croisée par exemple, mais encore faut-il connaître un intervalle *a priori* pour la gamme de valeurs à tester. Par exemple, le paramètre μ, en recherche d'information, prend des

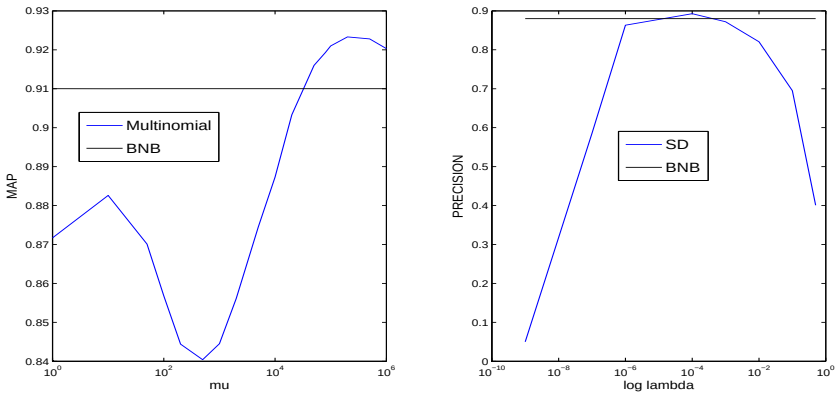

FIG. 4 – *Evolution de la MAP en fonction de μ sur le corpus Reuters ModApte split et de la précision en fonction de λ sur le corpus 20Newsgroup*

valeurs entre $[0, 2000]$; en catégorisation des valeurs comme 1 million ou 500,000 donnent des performances acceptables. De même pour le modèle Dirichlet lissé, on constate que les valeurs optimales de λ sont autour de 0.0001, c'est à dire que le lissage introduit gomme presque entièrement l'influence des données initiales. Pour des valeurs de λ autour de 0.5, les performances du modèle Dirichlet lissé s'effondrent alors que l'interprétation de ce paramètre est naturelle et que cette gamme de valeurs est utilisée en recherche d'information. En comparaison, le modèle BNB n'est pas soumis à ce genre de fluctuation et offre une certaine robustesse au niveau des performances.

6 Conclusion

Nous avons examiné dans cet article les liens entre modèles probabilistes de documents textuels et observations empiriques sur la distribution des mots au sein d'une collection. En particulier, nous avons proposé une caractérisation formelle de ces observations, caractérisation qui permet d'évaluer dans quelle mesure une distribution de probabilité est en accord avec les observations empiriques. De la distribution négative binomiale mise en avant dans plusieurs travaux antérieurs, nous avons dérivé la famille de distribution beta négative binomiale, qui contient la distribution de Yule-Simon. Notre dérivation fournit donc une nouvelle interprétation de la loi de Yule-Simon.

Nous avons ensuite montré la bonne adéquation de la distribution obtenue aux observations empiriques, et avons montré comment cette distribution permettait de réinterpréter plusieurs modèles existant (modèle *DFR* pour la recherche d'information, et modèle Bayes naïf pour la classification). Nous avons enfin montré comment utiliser notre modèle en catégorisation. Dans ce dernier cadre, nous avons montré comment notre modèle se comparait au modèle multinomial et au modèle Dirichlet lissé, ce dernier étant considéré comme un des meilleurs

modèles de catégorisation de textes. Sur une des collections testées (*20NewsGroup*), les différents modèles fournissent des résultats équivalents. Sur *Industry*, le modèle Dirichlet lissé fournit les meilleurs résultats alors que sur *WebKB* c'est le modèle que nous avons introduit qui fournit les meilleurs résultats. Toutefois, à la différence des autres modèles, notre modèle ne nécessite le réglage d'aucun paramètre et offre ainsi une certaine robustesse au niveau des performances.

Références

Amati, G. et C. J. V. Rijsbergen (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst. 20*(4), 357–389.

Anscombe, F. (1948). The transformation of poisson, binomial and negative-binomial data. *Biometrika 35*.

Baayen, R. (2001). *Word Frequency Distributions*. Dordrecht : Kluwer Academic.

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

Buntine, W. et A. Jakulin (2004). Applying discrete pca in data analysis. In *AUAI '04 : Proceedings of the 20th conference on Uncertainty in artificial intelligence*, Arlington, Virginia, United States, pp. 59–66. AUAI Press.

Church, K. W. et W. A. Gale (1995). Poisson mixtures. *Natural Language Engineering 1*, 163–190.

Clinchant, S. et É. Gaussier (2008). The bnb distribution for text modeling. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, et R. W. White (Eds.), *ECIR*, Volume 4956 of *Lecture Notes in Computer Science*, pp. 150–161. Springer.

Elkan, C. (2006). Clustering documents with an exponential-family approximation of the dirichlet compound multinomial distribution. In W. W. Cohen et A. Moore (Eds.), *ICML*, Volume 148 of *ACM International Conference Proceeding Series*, pp. 289–296. ACM.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications, Vol. I*. Wiley, New York.

Gaussier, É. et C. Goutte (2005). Relation between PLSA and NMF and implications. In R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, et J. Tait (Eds.), *SIGIR*, pp. 601–602. ACM.

Girolami, M. et A. Kabán (2003). On an equivalence between plsi and lda. In *SIGIR*, pp. 433–434. ACM.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *SIGIR*, pp. 50–57. ACM.

Johnson, N., A. Kemp, et S. Kotz (1993). *Univariate Discrete Distributions*. John Wiley & Sons, Inc.

Katz, S. M. (1996). Distribution of content words and phrases in text and language modelling. *Nat. Lang. Eng. 2*(1), 15–59.

Lee, D. D. et H. S. Seung (1999). Learning the parts of objects by non-negative matrix factorization. *Nature 401*(6755), 788–791.

Lee, D. D. et H. S. Seung (2000). Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, et V. Tresp (Eds.), *NIPS*, pp. 556–562. MIT Press.

Madsen, R. E., D. Kauchak, et C. Elkan (2005). Modeling word burstiness using the dirichlet distribution. In L. D. Raedt et S. Wrobel (Eds.), *ICML*, Volume 119 of *ACM International Conference Proceeding Series*, pp. 545–552. ACM.

Mccallum, A. et K. Nigam (1998). A comparison of event models for naive bayes text classification. In *The Fifteenth National Conference on Artificial Intelligence (AAAI)*.

Minka, T. (2003). *Estimating a Dirichlet Distribution*. Ph. D. thesis, Unpublished paper available at http ://research.microsoft.com/m̃inka.

Nallapati, R., T. Minka, et S. Robertson (2006). The smoothed-dirichlet distribution : a new building block for generative models. In *CIIR Technical Report - http ://www.cs.cmu.edu/ nmramesh/sd_tc.pdf*.

Orlov, J. et R. Chitashvili (1983). Generalized z-distribution generating the well-known "rank-distributions". *Bulletin of the Academy of Sciences, Georgia 110*.

Ponte, J. M. et W. B. Croft (1998). A language modeling approach to information retrieval. In *SIGIR*, pp. 275–281. ACM.

Rigouste, L. (2006). *Modéthodes probabilistes pour l'analyse exploratoire de donnÃ©es textuelles*. Ph. D. thesis, Thèse de l'ENST, Télécom Paris.

Zhai, C. et J. Lafferty (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst. 22*(2), 179–214.

Zipf, G. (1949). *Human behaviour and the principle of least effot : An introduction to human ecology*. Addison-Wesley.

Annexe A - Démonstration de la propriété 2

Rappelons la propriété 2 :

Soit $P(x_i)$ une distribution pour le mot i, et soit $a_n = \frac{P(x_i=n+1)}{P(x_i=n)}$.

(i) *S'il existe n_o tel que la suite a_n est croissante à partir du rang n_0, alors i est en rafale (au niveau n_0) sous P.*

(ii) *S'il existe n_o tel que la suite a_n est décroissante à partir du rang n_0, alors i n'est pas en rafale sous P.*

Preuve

Pour $n \geq n_0$, nous avons :

$$P(x_i \geq n+1 | x_i \geq n) = \frac{P(x_i \geq n+1)}{P(x_i \geq n)} = \frac{1}{\frac{P(x_i=n)}{P(x_i \geq n+1)} + 1}$$

Mais :

$$\frac{P(x_i \geq n+1)}{P(x_i = n)} = a_n + a_n a_{n+1} + a_n a_{n+1} a_{n+2} + \cdots$$

Et de façon similaire :

$$\frac{P(x_i \geq n+2)}{P(x_i = n+1)} = a_{n+1} + a_{n+1}a_{n+2} + a_{n+1}a_{n+2}a_{n+3} + \cdots$$

La comparaison terme à terme des membres droits des équations ci-dessus fournit pour (i), sachant que la suite a_n est croissante :

$$\forall n \in N, n \geq n_0\,, \frac{P(x_i \geq n+2)}{P(x_i = n+1)} \geq \frac{P(x_i \geq n+1)}{P(x_i = n)}$$

et donc :

$$\forall n \in N, n \geq n_0\,, P(x_i \geq n+2 | x_i \geq n+1) \geq P(x_i \geq n+1 | x_i \geq n)$$

ce qui établit (i).
De façon similaire, nous obtenons pour (ii) :

$$\forall n \in N, n \geq n_0\,, P(x_i \geq n+2 | x_i \geq n+1) \leq P(x_i \geq n+1 | x_i \geq n)$$

ce qui prouve (ii). □

Summary

In this paper, we study the links between probabilistic models of text collections and empirical observations concerning word frequency distributions. In particular, we propose formal characterizations of these empirical observations, and introduce the beta negative binomial distribution. This distribution, already known under different terms, is derived here from different considerations, and is shown to behave well wrt empirical observations. We also show that a categorization model based on the beta negative binomial distribution yields robust results on several standard collections.