Latent Dirichlet Allocation

Alberto Bietti
alberto.bietti@mines-paristech.fr

Mai 2012

1 Introduction

Le modèle *Latent Dirichlet Allocation* (LDA) [2] est un modèle probabiliste génératif qui permet de décrire des collections de documents de texte ou d'autres types de données discrètes. LDA fait partie d'une catégorie de modèles appelés "topic models", qui cherchent à découvrir des structures thématiques cachées dans des vastes archives de documents. Ceci permet d'obtenir des méthodes efficaces pour le traitement et l'organisation des documents de ces archives: organisation automatique des documents par sujet, recherche, compréhension et analyse du texte, ou même résumer des textes. Aujourd'hui, ce genre de méthodes s'utilisent fréquemment dans le web, par exemple pour analyser des ensemble d'articles d'actualité, les regrouper par sujet, faire de la recommendation d'articles, etc. Des modèles de ce type peuvent également s'utiliser sur des images, en utilisant des "mots visuels", par exemple pour regrouper des images par catégorie (voir L. Fei-Fei et P. Perona [3]), ou encore pour les problèmes de filtrage collaboratif (*collaborative filtering*), par exemple la recommendation de films, en assimilant un utilisateur et les films qu'il a vus à un document et les mots qu'il contient.

Le LDA est un modèle Bayesien hiérarchique à 3 couches (voir la Figure 2 pour une représentation graphique): chaque document est modélisé par un mélange de *topics* (thèmes) qui génère ensuite chaque mot du document. La structure des documents du corpus peut être déterminée par des techniques d'inférence approchée basées sur des méthodes variationnelles ou des méthodes de Gibbs sampling. Les paramètres des distributions peuvent être estimés par l'algorithme EM. La librairie `lda-c` de D. Blei (qui utilise des méthodes variationnelles pour l'inférence et l'estimation de paramètres) et le package *R* `lda` (qui utilise le *collapsed Gibbs sampling*) on été testés sur des corpus de textes adaptés.

2 Le modèle

La Figure 2 représente le modèle graphique de LDA et la Figure 1 en donne une intuition. Commençons par expliciter les différents termes et paramètres du modèle:

- Un *mot* w est la donnée discrète, correspondant à l'indice d'un mot dans un vocabulaire fixe de taille V. On peut considérer que w est un vecteur de taille V de composantes toutes nulles sauf pour la composante i où i est l'indice du mot choisi ($w^i = 1$).
- Un *document* est un N-uplet de mots, $\mathbf{w} = (w_1, \dots, w_N)$.

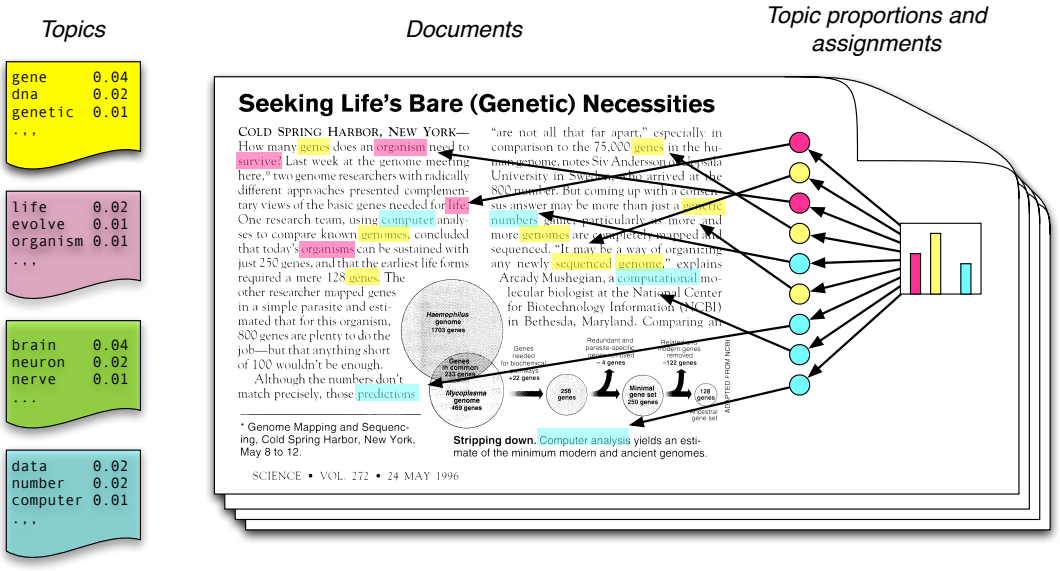

Figure 1: Schéma décrivant LDA. A gauche, on peut voir la structure de chaque *topic*, donnant une probabilité à chaque mot d'un vocabulaire fixe. Pour un document donné, l'histogramme à droite décrit la distribution de *topics* dans ce document. Pour chaque mot du document, on choisit d'abord un sujet depuis cette distribution (les bulles), puis on tire un mot depuis le sujet choisi. Source: [1].

- Un *corpus* est une collection de D documents, $\mathbf{D} = (\mathbf{w}_1, \ldots, \mathbf{w}_D)$.
- Les variables $z_{d,n}$ représentent le topic choisi pour le mot $w_{d,n}$.
- Les paramètres θ_d représentent la distribution de topics du document d.
- α et η définissent les distributions *à priori* sur θ et β respectivement, où β_k décrit la distribution du topic k.

Processus de génération Le processus génératif suivi par LDA pour un document $\mathbf{w}$ est le suivant (voir le modèle graphique de la Figure 2):

1. Choisir $\theta \sim Dirichlet(\alpha)$.
2. Pour chaque mot w_n:
   - Choisir un topic $z_n \sim Multinomial(\theta)$
   - Choisir un mot $w_n \sim Multinomial(\beta_k)$, avec $k = z_n$.

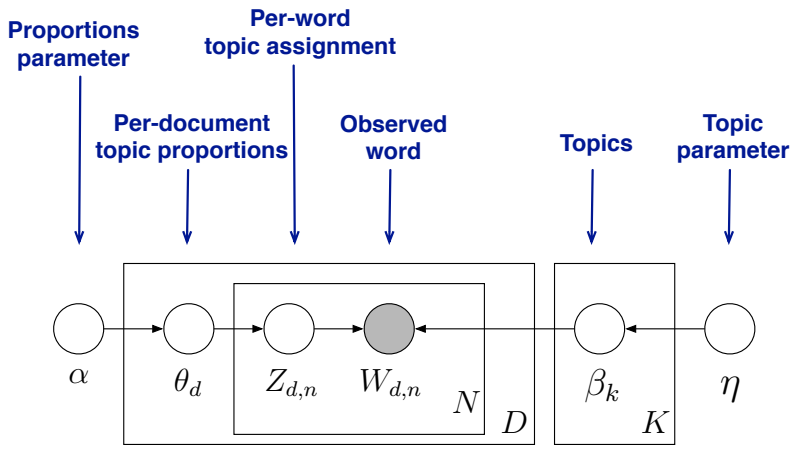

Figure 2: Représentation de LDA sous forme de modèle graphique. Les boîtes représentent des répliques du modèle qu'elles contiennent (par exemple, il y a une boîte N pour chaque document D). Source: [1].

Loi de Dirichlet La loi de Dirichlet permet de tirer une variable θ telle que $\forall i, \theta_i \geq 0$ et $\sum_{i=1}^{k} \theta_i = 1$ (θ est dans le $(k-1)$-simplexe). Sa densité est de la forme:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$$

avec $\alpha \in \mathbb{R}^k, \alpha_i > 0$ et $\Gamma(x)$ la fonction Gamma. Cette distribution permet donc d'obtenir une distribution multinomiale de paramètre θ, correspondant pour LDA au mélange de topics d'un document w. La Figure 3 montre la densité d'une telle distribution pour 3 topics. Chaque sommet du triangle correspond à un topic, et chaque point du triangle représente donc une pondération des 3 topics. Le paramètre $\alpha = \alpha_1 + \cdots + \alpha_k$ contrôle l'homogénéité des θ_i: lorsque α est grand, les θ_i sont proches et homogènes, lorsque α est petit, la plupart des θ_i sont proches de 0 sauf quelques uns. Dans les cas extrêmes, tous les θ_i sont égaux ($\alpha \to \infty$) ou tous les θ_i sont nuls sauf un ($\alpha \to 0$).

L'un des avantages de la loi de Dirichlet est qu'elle est conjuguée à la loi multinomiale, c'est à dire que si $z_1, \ldots, z_N$ sont des variables multinomiales de paramètre θ, alors la variable $\theta|z_1, \ldots, z_N$ donnée par $p(\theta|z_1, \ldots, z_N) \propto p(z_1, \ldots, z_n|\theta)p(\theta|\alpha)$ suit également une loi de Dirichlet. Ceci permettra de simplifier les calculs au moment de l'inférence.

Probabilité jointe et loi marginale Etant donnés les paramètres α et β, la probabilité jointe du mélange de topics θ, des N topics $\mathbf{z}$ et de N mots $\mathbf{w}$ est donnée par:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)p(w_n|\beta_{z_n})$$

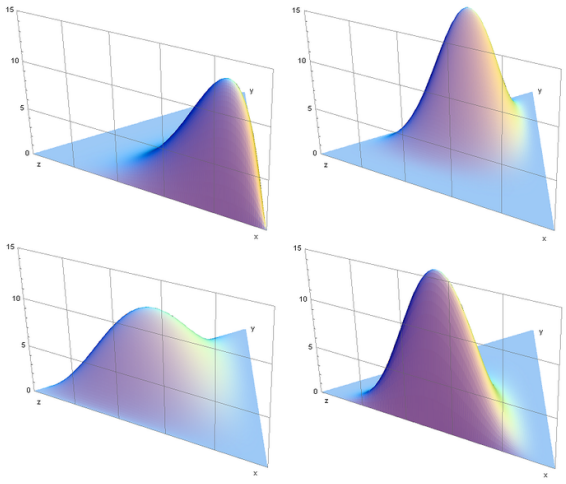

Figure 3: Fonction de densité de loi de Dirichlet à 2 dimensions ($k = 3$ topics) sur le triangle. Source: Wikipedia.

La loi marginale d'un document $\mathbf{w}$ est alors:

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|\beta_{z_n}) \right) d\theta \tag{1}$$

et il suffit de prendre le produit de cette quantité pour chaque document $\mathbf{w}$ du corpus pour obtenir la probabilité de ce corpus.

3 Inférence et estimation de paramètres

Nous avons décrit le modèle du LDA au paragraphe précédent et montré son utilité pour l'analyse de corpus de documents. Mais les variables et paramètres du modèle ne sont pas connus initialement, et il faut essayer de les apprendre à partir des données observables, c'est à dire les mots des documents. Dans la représentation graphique de la Figure 2, on peut voir que les seules variables observées sont les mots $w_{d,n}$, alors que toutes les autres variables sont cachées. Etant donnés les paramètres α et β, le rôle de l'inférence est de déterminer les variables cachées θ et z_n d'un document $\mathbf{w}$, étant donnée la liste des mots w_n du document. Les principales méthodes d'inférence (approchée) pour LDA sont les méthodes de sampling (notamment le *collapsed Gibbs sampling*) et les méthodes variationnelles (particulièrement les méthodes *mean-field*, qui peuvent se faire en batch ou en ligne). On peut ensuite faire recours à ce procédé d'inférence pour estimer les paramètres α, β et η du modèle grâce à l'algorithme EM.

3.1 Inférence

Le problème principal de l'inférence pour LDA est celui de déterminer la distribution *à posteriori* des variables cachées étant donné le document (et les paramètres α et β):

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}$$

Cette distribution est malheuresement très difficile à calculer, comme on peut déjà le remarquer sur l'expression du dénominateur donnée à l'équation 1 (qui peut s'exprimer en fonction des paramètres du modèle). Une inférence exacte utilisant la distribution *à posteriori* des paramètres est donc inenvisageable. Mais il existe plusieurs méthodes d'inférence approchée qui peuvent être utilisées pour LDA, utilisant par exemple l'approximation variationnelle ou le *Markov chain Monte Carlo*.

3.1.1 Markov chain Monte Carlo et Gibbs sampling

Les méthodes *Markov chain Monte Carlo* (MCMC) consistent à construire une chaîne de Markov sur les variables cachées dont la loi stationnaire est la distribution *à posteriori* cherchée. Pour rappel, une chaîne de Markov est définie par une loi de transition $p(\theta^{t+1}, \mathbf{z}^{t+1}|\theta^t, \mathbf{z}^t)$ d'un état $(\theta^t, \mathbf{z}^t)$ au suivant $(\theta^{t+1}, \mathbf{z}^{t+1})$, et peut converger vers une loi stationnaire qui est laissée invariante par la transition. Une fois l'état stationnaire atteint, les échantillons donnés par la chaîne de Markov suivent cette loi stationnaire qui est la distribution voulue.

Le *Gibbs sampling* est une méthode MCMC où la transition de la chaîne est donnée par la loi conditionnelle d'une variable cachée étant données les observations et l'état courant des autres variables cachées. Pour LDA, on peut successivement obtenir des échantillons $\theta|\mathbf{w}, \mathbf{z}$ puis $z_n|\theta, \mathbf{w}$. Le *collapsed Gibbs sampling* échantillonne sur les topics seulement, depuis $z_n|z_{-n}, \mathbf{w}$, en intégrant les θ.

3.1.2 Inférence variationnelle

L'inférence variationnelle utilise l'optimisation plutôt que l'échantillonnage. L'idée est de commencer par établir une distribution sur les variables cachées avec des paramètres libres (les *paramètres variationnels*), puis d'optimiser les paramètres variationnels pour que la distribution converge vers la distribution *à posteriori* souhaitée.

3.1.3 Estimation des paramètres

Pour estimer les paramètres α et β, on peut utiliser la méthode *empirical Bayes* (ou maximum de vraisemblance marginale), qui consiste à chercher des paramètres α et β qui maximisent la log-vraisemblance (marginale) des données:

$$\ell(\alpha, \beta) = \sum_{d=1}^{D} \log p(\mathbf{w}_d|\alpha, \beta)$$

On a vu (équation 1) que le calcul de $p(\mathbf{w}_d|\alpha, \beta)$ est inenvisageable en pratique, mais l'inférence variationnelle permet d'obtenir la borne inférieure de la log-vraisemblance qui peut être exploité par un algorithme Espérance-Maximisation appelé *variational EM*, qui utilise en plus les paramètres

variationnels de la méthode décrite au paragraphe 3.1.2. L'algorithme itère les deux étapes suivantes: l'étape E qui effectue une inférence variationnelle avec les paramètre α et β courants pour calculer la log-vraisemblance, et l'étape M qui maximise en α et β la borne inférieure de la log-vraisemblance calculée.

4 Résultats

Les résultats de la Figure 4 ont été obtenus en utilisant la librairie C de David Blei `lda-c`[1] sur un corpus de 2246 articles de *Associated Press*. Chaque colonne représente un thème (topic) découvert, et les mots y sont classés par probabilité décroissante. On peut voir par exemple que la première colonne traite de présidents et de politique, la troisième de faits policiers et la septième de marchés financiers.

Un autre exemple est donné dans la Figure 5, qui montre quelques thèmes découverts par LDA sur un corpus de 17000 articles scientifiques du magazine *Science*. La Figure 6 montre un exemple d'explorateur de thèmes qui a été fait à partir de 100000 articles Wikipedia en utilisant LDA. Cet explorateur est disponible à l'adresse `http://www.sccs.swarthmore.edu/users/08/ajb/tmve/wiki100k/browse/topic-`

[1] Disponible à l'adresse `http://www.cs.princeton.edu/~blei/lda-c/`

References

[1] D. M. Blei. Introduction to probabilistic topic models. In *Communications of the ACM*, à paraître.

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

[3] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

bush	i	police	soviet	percent	mecham	stock
dukakis	think	shot	gorbachev	year	keating	market
campaign	dont	man	president	rate	senators	index
jackson	people	arrested	summit	last	lincoln	stocks
president	like	two	reagan	report	deconcini	trading
democratic	get	people	bush	prices	meeting	million
convention	going	city	union	increase	john	dow
presidential	just	shooting	europe	month	barry	issues
republican	say	night	gorbachevs	rose	like	rose
new	know	killed	moscow	price	time	volume
vice	im	yearold	leader	production	office	shares
george	thats	officers	mikhail	department	years	jones
bentsen	see	death	nato	months	nixon	exchange
primary	go	car	meeting	annual	wine	new
sen	things	authorities	world	average	bishops	average
michael	make	found	new	government	church	wall
bushs	back	monday	leaders	expected	five	american
reagan	got	wounded	foreign	new	made	big
told	says	men	visit	inflation	senate	prices
state	am	officer	american	january	told	board

---

bill	communist	court	dollar	budget	israel	police
senate	party	case	cents	bush	israeli	people
house	korea	supreme	late	billion	jewish	students
legislation	south	ruling	gold	congress	arab	demonstrators
sen	korean	judge	lower	spending	palestinian	protesters
measure	north	state	cent	deficit	peace	killed
vote	first	appeals	higher	president	palestinians	government
rep	solidarity	decision	futures	plan	plo	today
congress	walesa	rights	yen	security	minister	two
president	war	federal	bid	cuts	west	protest
committee	two	courts	ounce	new	east	violence
law	president	law	london	fiscal	gaza	building
amendment	leader	appeal	pound	cut	occupied	state
voted	government	order	trading	administration	bank	security
new	people	lawyers	new	programs	jews	injured
debate	talks	attorney	prices	social	shamir	condition
approved	th	ruled	troy	bushs	middle	protests
republican	congress	souter	fell	house	territories	capital
veto	union	civil	francs	federal	meeting	student
year	years	lawyer	york	tax	peres	moslem

Figure 4: Les 20 premiers mots de quelques topics sur les 100 obtenus avec la librairie `lda-c` sur des articles de Associated Press.

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Figure 5: 4 topics obtenus sur des articles scientifiques.

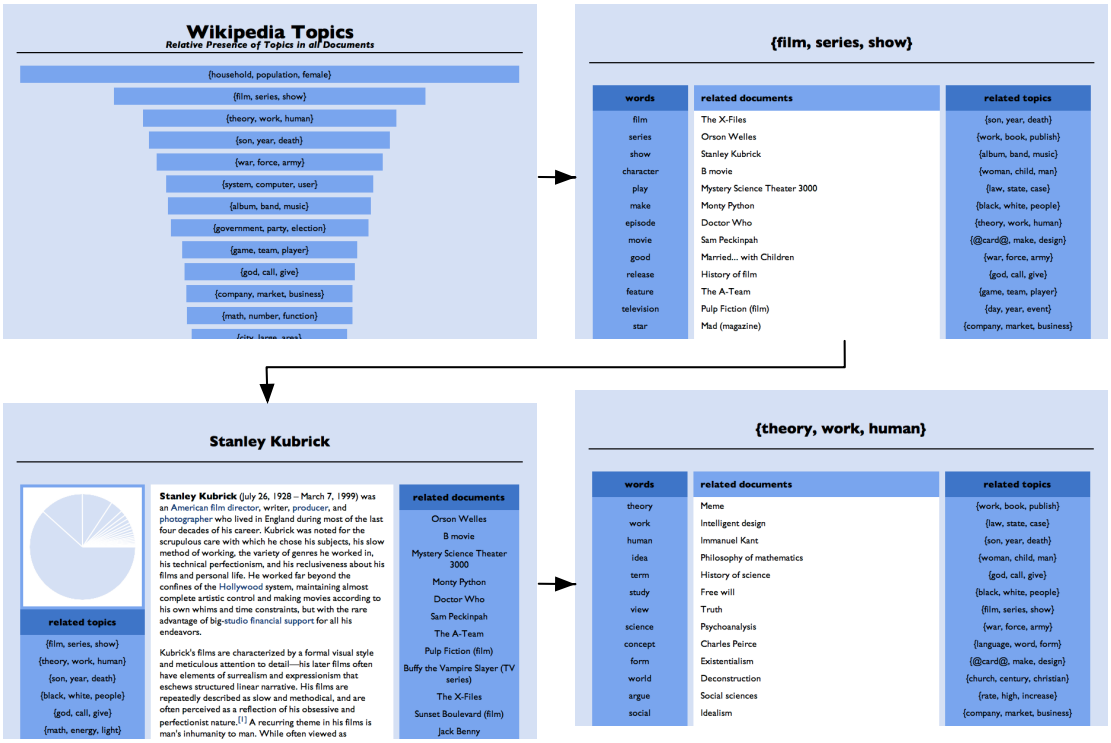

Figure 6: Explorateur de documents Wikipedia basé sur LDA.