Topic Modeling

Bilel Benbouzid - Alexis Perrier - Décembre 2017

Slack: upem-numi.slack.com

Invitation: bit.ly/2yyUF83

Alexis Perrier - Data Scientist

- Learning analytics @berklee.edu
- Signal processing @splice.com
- **NLP**, healthcare @docenthealth.com

- Twitter: @alexip
- Linkedin.com/in/alexisperrier
- alexis.perrier@gmail.com

Plan de la semaine

2 corpus:

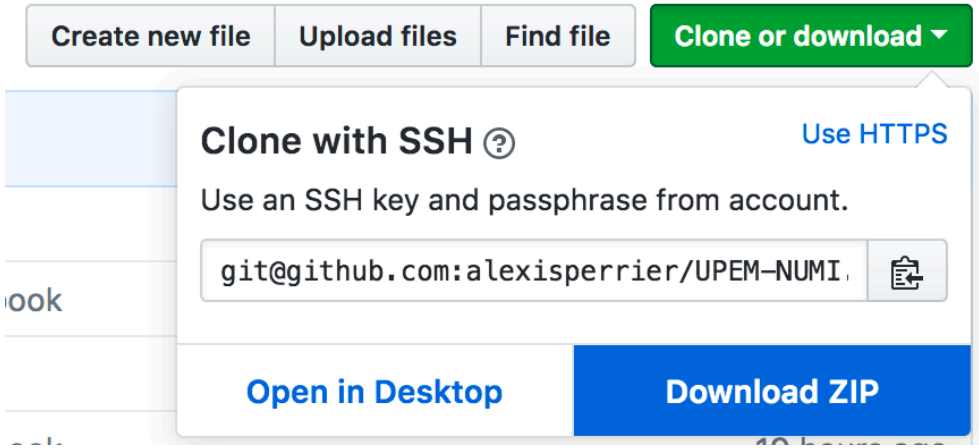
- Est républicain 1999
- God Emperor Trump https://www.facebook.com/GodEmperorTrump/

3 Labs:

- Lundi: Gensim - python - Est Républicain
- Mardi: STM - R - Est Républicain
- MJV (3j): God Emperor Trump

Outils de travail

- Slack: upem-numi.slack.com
- Invitation: bit.ly/2yyUF83
- Notebooks: 35.196.112.120:5000 , jupyter.org
- Code, slides, sur Github: https://github.com/alexisperrier/UPEM-NUMI



Topic Modeling, pourquoi?

Vous avez des milliers de documents,
comment savoir ce qu'ils contiennent ?

- Exemples
  - Article de presse sur une longue période
  - Réseaux sociaux, commentaires, tweets, …
  - Articles scientifiques, textes légaux
  - Débats retranscrit,
  - Littérature, fiction, théâtre, …

Top Down vs Bottom Up

Approche supervisée, top down:

- Vous définissez les sujets a priori comme autant de catégories.
  Vous classez les documents dans ces catégories.
  => Classification de documents

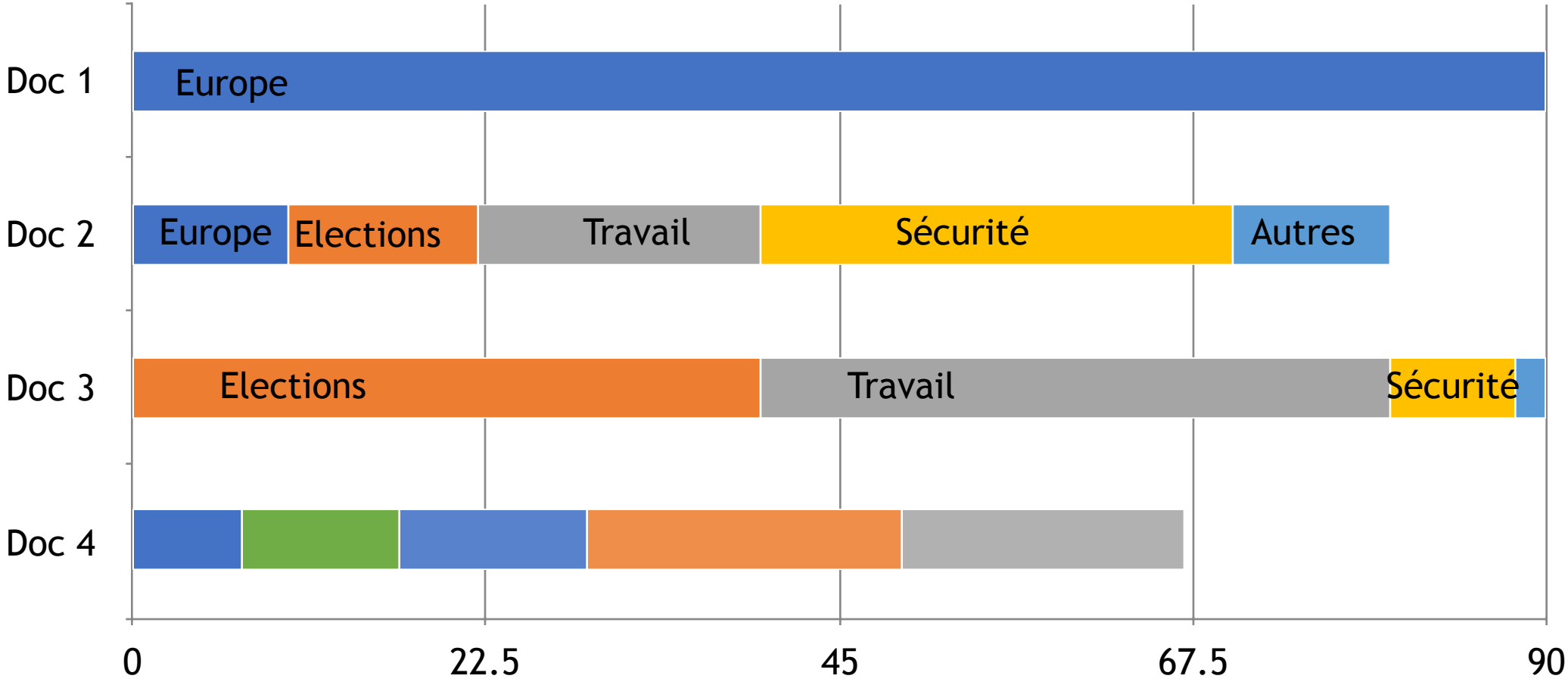
Approche non supervisée, bottom up:

- Aucun a priori préalable sur le contenu des documents
  Inférer les topics
  => Topic Modeling

Topic Modeling

Chaque **document** peut contenir aucun, un ou plusieurs **topics**

Chaque **topic** est composé de plusieurs **mots**



Ensuite

- Croisement avec des **variables externes**
  - Evolution dans le temps
  - Auteur ou Locuteur, parti politique, context,
- Utiliser les topics pour classer les documents
  sous topics, analyse de sentiment, vocabulaire, …
- Réduction de dimension et modèles prédictifs

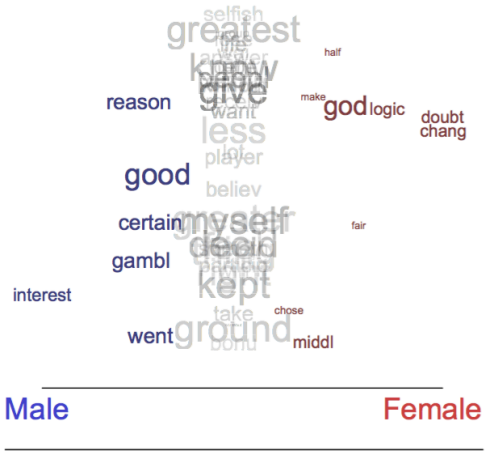
Exemples

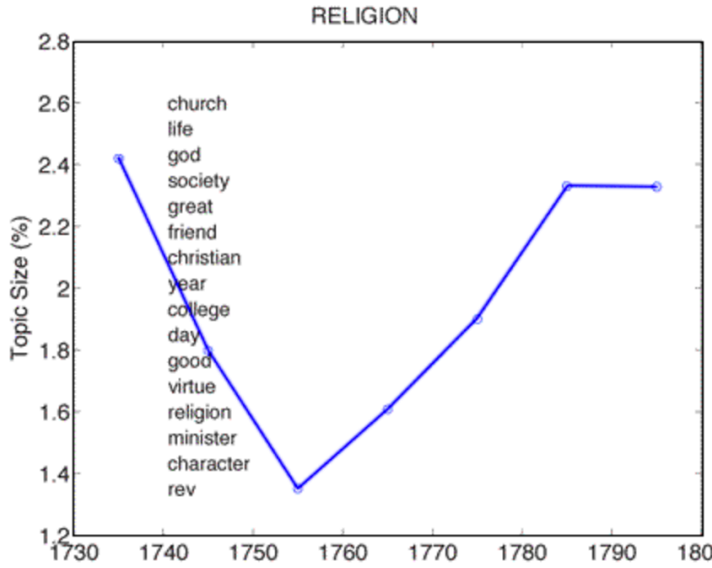
- Followers sur twitter (gensim, LDA, LSA, python)
- Debats présidentiels américains (stm, LDA, R)
- Analyse des notes en milieu hospitalier pour prédire *patient satisfaction* et identifier les problèmes

Autres travaux:

- Gazette de Pensylvanie http://www.common-place-archives.org/vol-06/no-02/tales/
- 3,346 works of 19th-century British, Irish, and American fiction evolution dans le temps
- Analyse graph / réseau des topics dans Proust (Stanford, Mallet, Java)
- Structural Topic Models for Open-Ended Survey Responses (Molly Roberts, STM)
- Topic modeling of French crime fiction novels

FIGURE 15 Intuitive Topic Allowing for Different Vocabularies Based on Gender





Autre exemple

#Datapol

Facebook FN - Commentaires
60 topics

Topic 23: celer, chose, bien, quelque, changer, avis, accord, sûr, savoi
Topic 34: peur, maintenir, perdre, voter, législatif, jamais, besoin, électeu
Topic 42: impression, hauteur, clair, raison, nier, remarquer, attaquer, ma
Topic 29: croire, espérer, essayer, vraiment, tenir, tellement, malheureuseme
Topic 49: commentaire, poser, franchement, autant, question, page, quand, c
Topic 31: vérité, dire, oser, gauche, tomber, leçon, droit, gaulle, haut, extrême,
Topic 27: plaisir, barrage, ménager, gamin, diviser, rire, faire, imaginer, écraser
Topic 19: plaire, exister, remettre, devenir, possible, loin, moins, faible, certaine
Topic 1: travailler, salaire, payer, bosser, mois, loyer, smic, impôt, retraiter, boul
Topic 24: oeil, mieux, comprendre, ouvrir, remplir, gouverner, encore, valoir, cou
Topic 16: chez, pays, ici, bla, guerre, fuir, mondial, retourner, battu, nôtre, install
Topic 45: vive, présidente, marine, président, vivement, coeur, république, excep
Topic 50: taire, connard, gueuler, enculer, couille, putain, dégager, merder, trou,
Topic 56: choix, écrire, fauter, réveiller, choisir, révolution, langue, bouger, major
Topic 17: policier, racaille, flic, théo, attentat, brûler, pompier, criminel, agresser,
Topic 35: patriote, changement, peupler, seul, unir, ensemble, espoir, retrouver, b
Topic 44: propagande, manipuler, manipulation, corrompu, corrompre, chaîner, té
Topic 10: couper, bourdin, émission, journaleux, inviter, parole, écouter, parler, en
Topic 51: aider, prison, carte, abord, clandestin, papier, occuper, dehors, priorité,
Topic 59: intelligent, insulter, truc, aucun, lol, connerie, femme, jeune, inventer, ha
Topic 40: presser, inverser, informer, évidemment, utiliser, souvent, vérifier, object
Topic 12: parent, avenir, petit, mère, enfant, maman, handicaper, gosse, école, fille
Topic 6: chacun, parce, raciste, bleu, respecter, aimer, peau, accepter, habiller, cou
Topic 47: sauver, france, musulmans, jeanne, redresser, erdogan, debout, sauvons
Topic 37: pur, partir, différence, important, tourner, représenter, pis, marquer, prête
Topic 46: jouer, niveau, habitude, juste, trés, raconter, peu, gagner, jeu, importer, p
Topic 15: pleurer, riche, banquier, rigoler, après, macaron, galérer, classer, cdi, plai
Topic 28: espere, arreter, etrangers, etre, apres, marrer, etes, quil, bientot, pere, plu
Topic 18: animal, proposer, idée, programmer, lire, bête, humain, intéresser, chat, s
Topic 32: trop, tard, deux, avant, nul, prochain, perso, jean, national, élection, front,
Topic 8: digne, soir, discourir, douter, prêt, rassurer, attendre, demain, catastrophe,
Topic 21: valls, hollande, sarko, hamon, mélanchon, quinquennat, bébé, pantin, soc
Topic 43: sortir, franc, référendum, européen, économie, parlement, allemagne, suis
Topic 14: rien, voler, pognon, africain, foutre, servir, bordel, million, individu, abrutir,
Topic 22: deuxième, tour, second, premier, mobiliser, dimanche, bulletin, voix, dépou
Topic 20: bol, montrer, cacher, ras, pourquoi, photo, dictature, contrer, lutter, daesh,
Topic 36: islamiste, voiler, coran, musulman, laïcité, crécher, charia, croix, religion, p
Topic 53: marion, placer, pocher, profond, rouer, mettre, dos, bâton, plein, coup, gen
Topic 58: autres, veut, plus, peut, ans, enfants, reste, font, ceux, seule, veulent, droit
Topic 52: courage, bonjour, router, bonne, soutien, croiser, dupont, cordialement, bo
Topic 9: lieu, prier, appliquer, chef, régler, hui, aujourd, prière, état, urgence, loi, inter
Topic 41: identité, égalité, redonner, fierté, souche, fraternité, liberté, française, civilis
Topic 33: yes, back, support, french, take, win, freedom, haver, would, hope, need, w

Etapes

Définition

1. Corpus brut
2. Unité d'analyse: tweet(s), commentaire(s), paragraphe(s), article(s), blog, livre, …
3. Technique utilisée: LSA, LDA, pLSA, HLDA, DTM, Topic2Vec
4. Variable externes?

Etapes:

1. Pre-processing:
   1. Enlever le bruit, **garbage-in garbage-out**
   2. Adapter, transformer le contenu
2. Topic Modeling: faire tourner les modèles
   1. Combien de topics?
3. Interprétation des résultats
   1. Nommer les topics
   2. **Mesurer** leur qualité
4. Visualisation
5. Wordcloud!

Corpus

- Longueur: du tweet (280) à l'article
- Langue: anglais, français, franglais
- Niveau de langue: classique, argot, smiley
- Contenu: images, urls, html, texte,
- Source: sms, tweets, forum, presse, OCR, transcript

Influence la transformation du texte

« *Nous avons des devoirs envers notre pays.*

Nous sommes les héritiers d'une grande histoire et du grand message humaniste adressé au monde.

Nous devons les transmettre d'abord à nos enfants, mais plus important encore, il faut les porter vers l'avenir et leur donner une sève nouvelle. »

Emmanuel Macron, 08 mai 2017

Bag of Words

L’information prise en compte est intégralement contenue dans la liste de mots issue du texte, indépendamment de leur position ou de leur fonction dans la phrase

Le texte original:

« *Il sera une page dans un livre de dix mille pages*
que l'on mettra dans une bibliothèque qui aura un million de livres,
une bibliothèque parmi un million de bibliothèques. »
Ionesco, Le roi se meurt 1962

devient:

aura bibliothèque bibliothèque bibliothèques dans dans de de de
dix il on l livre livres mettra mille million million
page pages parmi que qui sera un un un une une une

Vectorizer le texte

Mots

Documents

	Arbre	Gazon	Été
Doc 1	3	0	0
Doc 2	1	2	1
…			
Doc N	0	2	4

Matrice Mots - Documents

Mots - Documents - TF-IDF

Comment évaluer la fréquence des mots dans un ensemble de documents?

TF-IDF: fréquence du mot dans un document, normalisée par sa présence dans l'ensemble du corpus

Pour un document donné:

- TF: Term Frequency: combien de fois le terme est dans le document
- IDF: Inverse document term frequency:
  # de documents / # documents contenant le terme

On obtient une matrice terme – document plus représentative que un simple comptage des mots

Voir:

- https://fr.wikipedia.org/wiki/TF-IDF
- TfidfVectorizer de scikit-learn

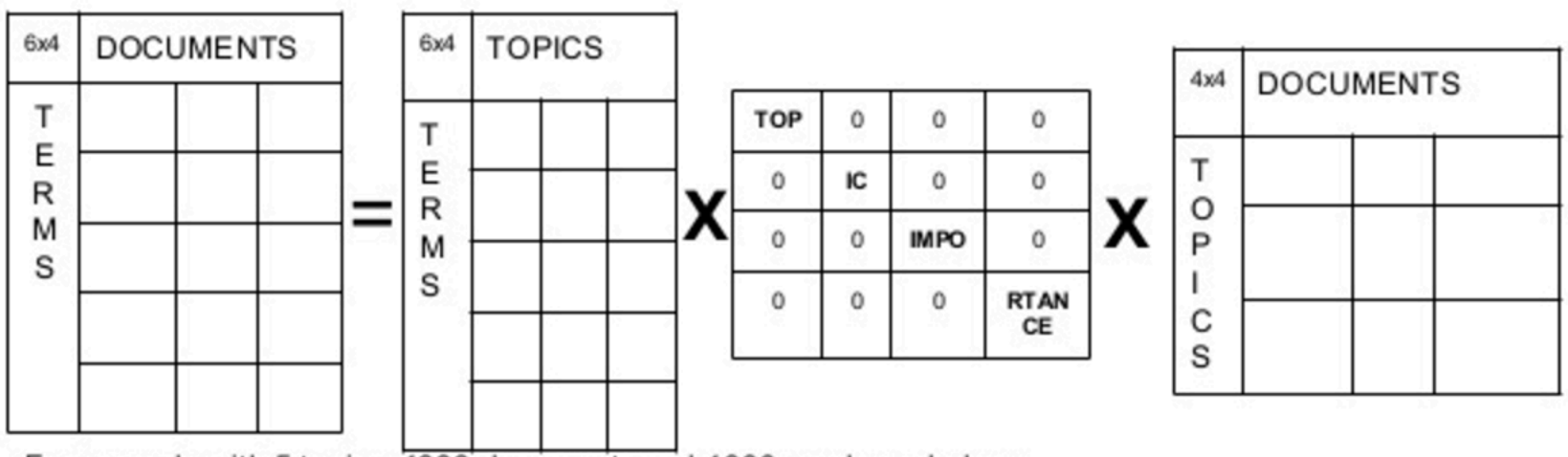
Latent Semantique Analysis

Aproche Deterministe

LSA

Nothing more than a **singular value decomposition (SVD)** of document-term matrix:

Find three matrices U, Σ and V so that: $X = U\Sigma V^t$



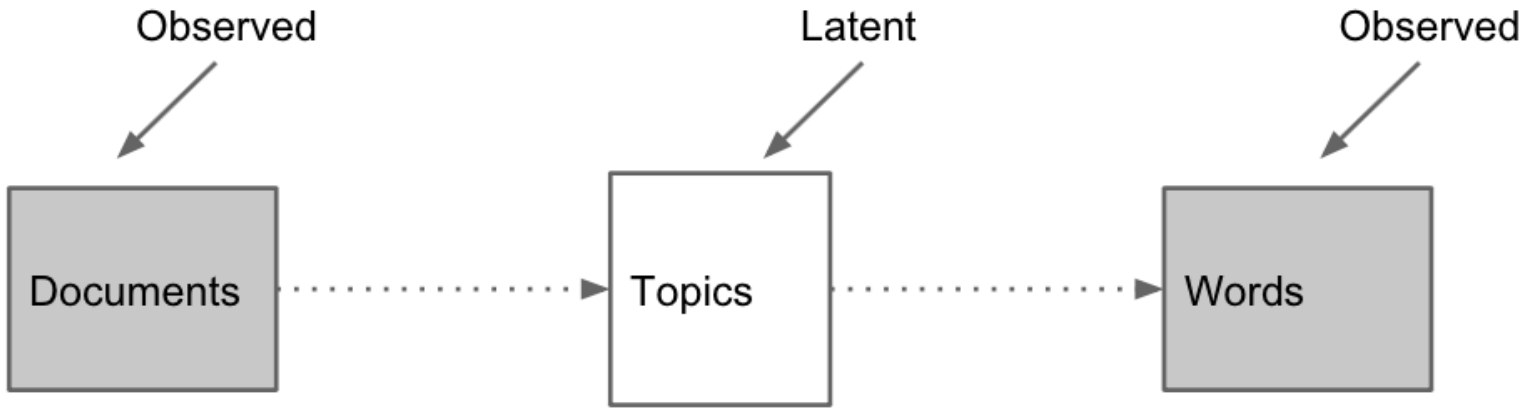
For example with 5 topics, 1000 documents and 1000 word vocabulary:
Original matrix: 1000 x 1000 = 10^6
LSA representation: 5x1000 + 5 + 5x1000 ~ 10^4
-> 100 times less space!

Latent Dirichlet Allocation

Blei 2002:

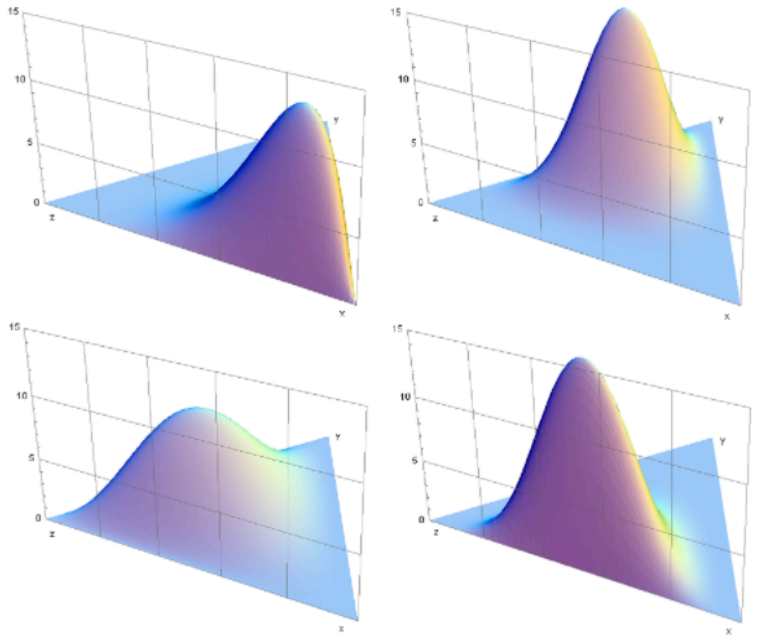
- Les topics sont distribués aléatoirement dans les documents
- Les mots sont distribués aléatoirement dans les topics
- L'aléatoire suit une distribution de Dirichlet

- **K: Nombre de topics**
- α: Nombre de topics par document
- ß: Nombre de mots par topic



Distribution de Dirichlet

Sorte de gaussienne généralisée



The Dirichlet Distribution

- Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$
- We write:

$$\Theta \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_m)$$

$$P(\theta_1, \theta_2, \ldots, \theta_m) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{m} \theta_k^{\alpha_k - 1}$$

- Samples from the distribution lie in the *m-1* dimensional probability simplex

https://fr.wikipedia.org/wiki/Loi_de_Dirichlet

Librairies

Python

- Gensim https://radimrehurek.com/gensim/
- LDA Python library: https://pypi.python.org/pypi/lda
- Scikit-learn: http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

R

- LDA, LSA
- Topicmodels
- STM package http://structuraltopicmodel.com/

Visualisation

- Gensim: LDAviz
- R STM: stmBrowser

Lab 01 - Est Républicain 1999

35.196.112.120:5000