**Mise en place de l’algorithme LDA**

## LDA (Allocation de Dirichlet Latente) ?

Supposons que nous avons les phrases suivant :

* J’aime manger des brocolis et des bananes.
* Je déteste les bananes et les épinards au petit déjeuner.
* Les chinchillas et les chatons sont mignons
* Ma sœur a adopté des chatons très mignons
* Regarde ce beau hamster près de ce brocoli.

Et maintenant quel est le rôle de Latent Dirichlet Allocation ?

Cet algorithme permet de manière automatique de découvrir les thèmes associés aux phrases précédentes. Par exemple, étant donné ces phrases données pour 2 sujets, LDA produit quelque chose comme ça :

* Phrases 1 and 2 : 100% Topic A
* Phrases 3 and 4 : 100% Topic B
* Phrase 5 : 60% Topic A, 40% Topic B
* Topic A : 30% brocoli, 15% bananes, 10% petit déjeuner, 10% épinards, … (au point où on peut interpréter que le Topic A est associé à la nourriture)
* Topic B : 20% chinchillas, 20% chatons, 20% mignons, 15% hamster, … (a au point où on peut interpréter que le Topic B est associé aux animaux)

## Mon utilisation

Mon souhait est d’utiliser les descriptions des différents événements qui ont eu lieu sur Nantes en 2016. Je souhaite étudier les mots contenus dans une description afin de pouvoir identifier les différents Topic associés. Le but final est que mon modèle soit suffisamment intelligent et autonome pour déterminer le topic lié à une unique description d’événement.

## Récolte des données

Je vais prendre toutes les données des événements de Nantes créé en 2016.

Ceux-ci sont disponible en donnée ouverte sur **opendatasoft** :

<https://public.opendatasoft.com/explore/dataset/evenements-publics-cibul/table/?disjunctive.tags&disjunctive.placename&disjunctive.city&refine.department=Loire-Atlantique&refine.region=Pays+de+la+Loire&q=&refine.date_start=2016&refine.city=Nantes&location=13,47.21263,-1.55946&basemap=mapquest>

## Source

<http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/>

<https://fr.wikipedia.org/wiki/Allocation_de_Dirichlet_latente>

<https://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf>