Séance 2.2: Analyse données digitales

Les principales méthodes

Visseho Adjiwanou, PhD.

SICSS - Montréal

29 November 2021

Plan de présentation

1 Introduction
2 Méthodes
3 Packages

Méthodes

Introduction

Nous allons décrire ici 4 méthodes d'analyse à partir des corpus de texte:

1. Analyse de sentiment (sentiments analysis)
2. Analyse des sujets (Topic modelling)
3. Analyse structurelle des topics (Structural topic modelling)
4. Analyse des réseaux de texte (Texnet)

Analyse de sentiments (sentiments analysis)

Introduction

- Technique qui permet d'extraire une information d'une document à partir d'une requête précise.
- Se base sur l'utilisation d'un dictionnaire qui décrit les éléments de la requête

Labo

Analyse des termes (Topic modelling)

Définition

- Une procédure automatisée pour coder le contenu des textes (y compris de très grands corpus) dans un ensemble de catégories significatives, ou «sujets».

- Un modèle génératif qui permet d'expliquer des ensembles d'observations (textes) par des groupes non observés (sujets) qui expliquent pourquoi certaines parties (mots) des données sont similaires.

Définition

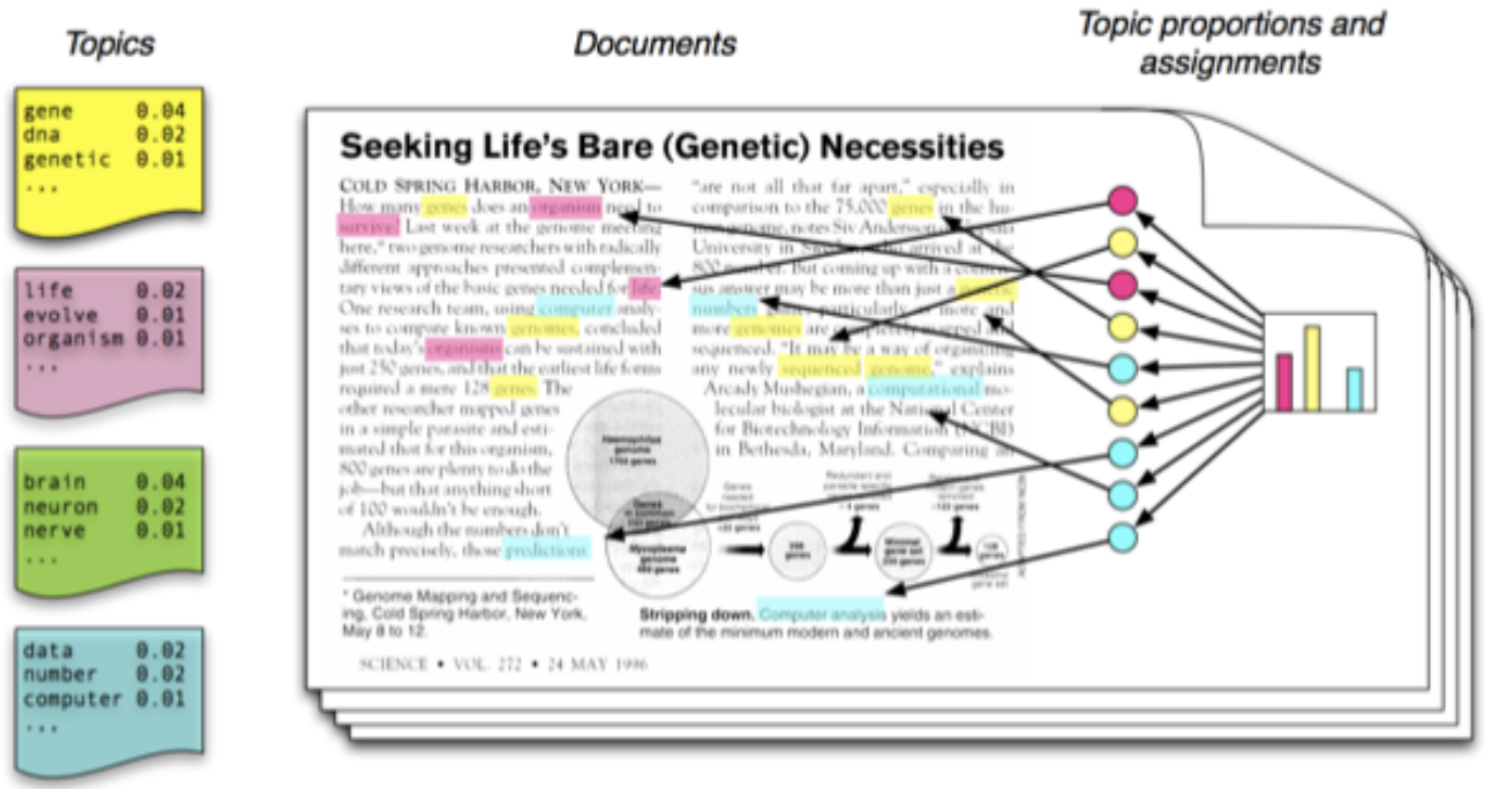

Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Figure 1:

Comment ça fonctionne

- Document (Document): un panier de mots produit selon un mélange de thèmes ou de sujets que l'auteur du texte entendait aborder.
- Sujet (Topic): une distribution sur tous les mots observés dans le corpus.

Comment ça fonctionne

- Document (Document): un panier de mots produit selon un mélange de thèmes ou de sujets que l'auteur du texte entendait aborder.
- Sujet (Topic): une distribution sur tous les mots observés dans le corpus.
- Les mots fortement associés aux sujets dominants du document ont plus de chances d'être sélectionnés et placés dans le sac de documents (c'est-à-dire plus de chances d'apparaître dans le document).

Comment ça fonctionne

- Document (Document): un panier de mots produit selon un mélange de thèmes ou de sujets que l'auteur du texte entendait aborder.
- Sujet (Topic): une distribution sur tous les mots observés dans le corpus.
- Les mots fortement associés aux sujets dominants du document ont plus de chances d'être sélectionnés et placés dans le sac de documents (c'est-à-dire plus de chances d'apparaître dans le document).
- Utilise l'analyse Bayesienne notamment le "Latent Dirichet Allocation"

Comment ça fonctionne

- Document (Document): un panier de mots produit selon un mélange de thèmes ou de sujets que l'auteur du texte entendait aborder.
- Sujet (Topic): une distribution sur tous les mots observés dans le corpus.
- Les mots fortement associés aux sujets dominants du document ont plus de chances d'être sélectionnés et placés dans le sac de documents (c'est-à-dire plus de chances d'apparaître dans le document).
- Utilise l'analyse Bayesienne notamment le "Latent Dirichet Allocation"
- Il s'agit d'un cas d'**apprentissage automatique non**

Apprentissage automatique (machine learning)

-1. Apprentissage supervisé

- Classification

-2. Apprentissage non supervisé

Apprentissage automatique (machine learning)

-1. Apprentissage supervisé

- Classification
- Prédiction

-2. Apprentissage non supervisé

Apprentissage automatique (machine learning)

-1. Apprentissage supervisé

- Classification
- Prédiction
- Régression

-2. Apprentissage non supervisé

Apprentissage automatique (machine learning)

-1. Apprentissage supervisé

- Classification
- Prédiction
- Régression

-2. Apprentissage non supervisé

- Clustering

Apprentissage automatique (machine learning)

-1. Apprentissage supervisé

- Classification
- Prédiction
- Régression

-2. Apprentissage non supervisé

- Clustering
- Analyse en composantes principales

Latent Dirichet Allocation (LDA)

- LDA, comme tous les modèles de sujets, suppose qu'il existe des sujets (termes) qui forment les éléments constitutifs d'un corpus.

- Les sujets sont des distributions sur les mots et sont souvent présentés sous la forme d'une liste de mots classés, avec les mots les plus probables en haut de la liste

Latent Dirichet Allocation (LDA)

- LDA, comme tous les modèles de sujets, suppose qu'il existe des sujets (termes) qui forment les éléments constitutifs d'un corpus.

- Les sujets sont des distributions sur les mots et sont souvent présentés sous la forme d'une liste de mots classés, avec les mots les plus probables en haut de la liste
- Cependant, nous ne savons pas quels sont les sujets a priori; le défi est de découvrir ce qu'ils sont.

Latent Dirichet Allocation (LDA)

Figure 7.1. Topics are distributions over words. Here are three example topics learned by latent Dirichlet allocation from a model with 50 topics discovered from the *New York Times* [324]. Topic 1 seems to be about technology, Topic 2 about business, and Topic 3 about the arts

Figure 2:

Latent Dirichet Allocation (LDA)

- En plus de supposer qu'il existe un certain nombre de sujets qui expliquent un corpus, LDA suppose également que chaque document d'un corpus peut être expliqué par un petit nombre de sujets.

Latent Dirichet Allocation (LDA)

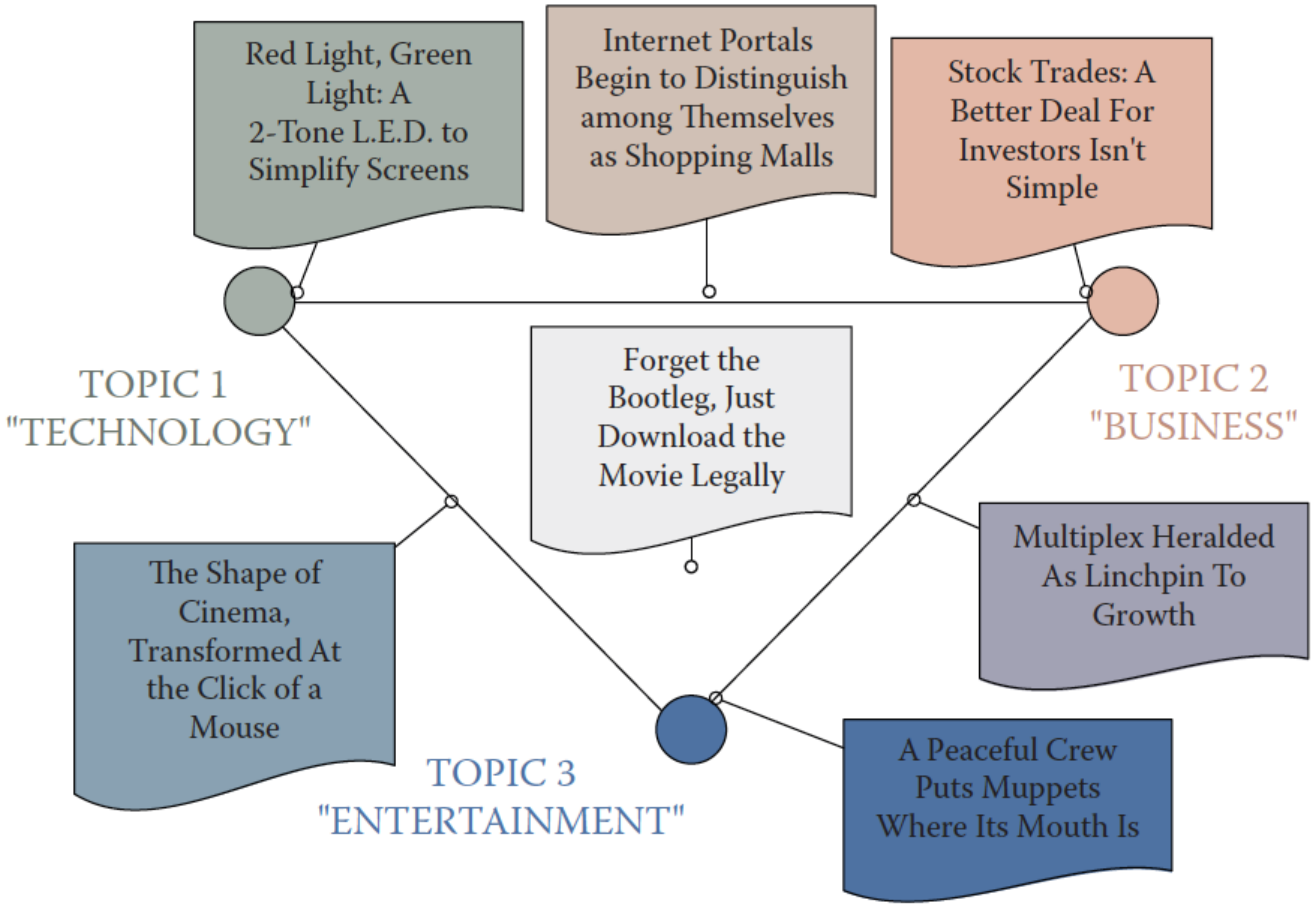

Figure 7.2. Allocations of documents to topics

>-

L'ensemble des sujets utilisées par un document est appelé allocation

Latent Dirichet Allocation (LDA)

- Algorithmiquement, le problème peut être considéré comme une boîte noire (d'où le terme méthode non supervisé).
- Étant donné un corpus et un entier K de termes (en entrée), fournissez les sujets qui décrivent le mieux la collection de documents: un processus appelé inférence postérieure.

Latent Dirichet Allocation (LDA)

- Algorithmiquement, le problème peut être considéré comme une boîte noire (d'où le terme méthode non supervisé).

- Étant donné un corpus et un entier K de termes (en entrée), fournissez les sujets qui décrivent le mieux la collection de documents: un processus appelé inférence postérieure.
- L'algorithme le plus courant pour résoudre ce problème est une technique appelée **échantillonnage de Gibbs**.

Échantillonnage de Gibbs

- Un modèle de sujet veut faire deux choses:
- il ne veut pas utiliser beaucoup de sujets dans un document et

Échantillonnage de Gibbs

- Un modèle de sujet veut faire deux choses:

- il ne veut pas utiliser beaucoup de sujets dans un document et
- il ne veut pas utiliser beaucoup de mots dans un sujet.

Échantillonnage de Gibbs: formule

- Soit :
- $N_{d,k}$ le nombre de fois que le document **d** a utilisé un sujet **k**, et
- $V_{k,w}$ le nombre de fois qu'un sujet **k** a utilisé un mot **w**.
- On a donc:
- $N_{d,.} = \sum_k N_{d,k}$ le nombre de sujets dans le document, et
- $V_{k,.} = \sum_w V_{k,w}$ le nombre de mots associé au sujet k

Échantillonnage de Gibbs: formule

- L'algorithme supprime les comptes pour un mot de $N_{d,k}$ et $V_{k,w}$ puis change le sujet d'un mot (avec un peu de chance en un meilleur sujet que celui qu'il avait auparavant).
- Grâce à plusieurs milliers d'itérations de ce processus, l'algorithme peut trouver des sujets cohérents, utiles et bien caractériser les données.
- Les deux objectifs de la modélisation de sujets - équilibrer les allocations de documents aux sujets et la distribution des sujets sur les mots - se rejoignent dans une équation qui les multiplie ensemble.
- Un bon sujet sera à la fois commun dans un document et expliquera bien l'apparence d'un mot.

Échantillonnage de Gibbs: formule

L'affectation de sujet $z_{d,n}$ du mot n dans le document d au sujet k est proportionnelle à :

$$p(z_{d,n=k}) \propto [\ \underbrace{\frac{N_{d,k}+\alpha}{N_{d,.}+K\alpha}}_{\text{combien le doc aime le sujet}}\ ][\ \underbrace{\frac{V_{k,n}+\beta}{N_{k,.}+V\beta}}_{\text{combien le sujet aime le mot}}\ ]$$

- où α et β sont des facteurs de lissage qui empêchent un sujet d'avoir une probabilité nulle si un sujet n'utilise pas de mot ou si un document n'utilise pas de sujet

En conclusion

Objectif: étant donné un corpus de textes et un nombre précis de sujets, trouver les paramètres qui l'ont probablement généré.

-***Entrée principale***: *texte et nombre de sujets à découvrir.*

-***Processus simplifié***: *choisissez à plusieurs reprises un sujet, puis un mot dans ce sujet, et placez-les dans le sac de mots représentant le document jusqu'à ce qu'un document soit complet .*

-***Sortie***: *distributions de mots pour chaque sujet, distributions de sujets pour le corpus.*

Structural Topic Model

Introduction

- Va pousser plus loin l'analyse des sujets en tenant compte des métadonnées
- Les métadonnées peuvent permettre d'expliquer les sujets (les sujets sont les variables dépendantes)

Introduction

- Va pousser plus loin l'analyse des sujets en tenant compte des métadonnées

- Les métadonnées peuvent permettre d'expliquer les sujets (les sujets sont les variables dépendantes)
- Les sujets peuvent servir de variables explicatives pour expliquer d'autres phénomènes

Labo

Analyse de réseaux et analyse des réseaux de texte

Introduction

- L'analyse de réseau fait référence à une famille de méthodes qui décrivent les relations entre les unités d'analyse.

- Un réseau est composé de nœuds (nodes) ainsi que de liens (edges) ou de connexions entre eux.

Introduction

- L'analyse de réseau fait référence à une famille de méthodes qui décrivent les relations entre les unités d'analyse.

- Un réseau est composé de nœuds (nodes) ainsi que de liens (edges) ou de connexions entre eux.
- Dans un réseau social les nœuds sont souvent des personnes individuelles, et les bords décrivent des amitiés, des affiliations ou d'autres types de relations sociales.

Introduction

- L'analyse de réseau fait référence à une famille de méthodes qui décrivent les relations entre les unités d'analyse.

- Un réseau est composé de nœuds (nodes) ainsi que de liens (edges) ou de connexions entre eux.
- Dans un réseau social les nœuds sont souvent des personnes individuelles, et les bords décrivent des amitiés, des affiliations ou d'autres types de relations sociales.
- Une riche tradition théorique en sciences sociales décrit comment les modèles de regroupement au sein des réseaux sociaux — et la position d'un individu au sein ou entre les clusters — sont associés à un éventail remarquablement large de résultats, notamment la santé, l'emploi, l'éducation et bien d'autres.

Introduction

- Bien que l'analyse de réseau soit le plus souvent utilisée pour décrire les relations entre les personnes, certains des premiers pionniers de l'analyse de réseau se sont rendu compte qu'elle pouvait également être appliquée pour représenter les relations entre les mots.

- Par exemple, on peut représenter un corpus de documents comme un réseau où chaque nœud est un document, et l'épaisseur ou la force des liens entre eux décrit les similitudes entre les mots utilisés dans deux documents.

Introduction

- Bien que l'analyse de réseau soit le plus souvent utilisée pour décrire les relations entre les personnes, certains des premiers pionniers de l'analyse de réseau se sont rendu compte qu'elle pouvait également être appliquée pour représenter les relations entre les mots.

- Par exemple, on peut représenter un corpus de documents comme un réseau où chaque nœud est un document, et l'épaisseur ou la force des liens entre eux décrit les similitudes entre les mots utilisés dans deux documents.
- Ou, on peut créer un réseau de textes où les mots individuels sont les nœuds, et les liens entre eux décrivent la régularité avec laquelle ils coexistent dans les documents.

Introduction

- L'approche réseau de l'analyse de texte automatisée présente de nombreux avantages.

- Tout comme les groupes de connexions sociales peuvent aider à expliquer une gamme de résultats, la compréhension des modèles de connexions entre les mots aide à identifier leur signification de manière plus précise que les approches du «sac de mots» discutées précédemment.

  -Deuxièmement, les réseaux de texte peuvent être construits à partir de documents de n'importe quelle longueur, alors que les modèles thématiques fonctionnent mal sur des textes courts tels que les messages des médias sociaux.

Labo

Packages

Tidytext

- Package
- https://cran.r-project.org/web/packages/tidytext/vignettes/tidytext.html
- Application
- https://www.tidytextmining.com/tidytext.html

tm

- Package
- https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf
- Application
- http://edutechwiki.unige.ch/fr/Tutoriel_tm_text_mining_package

quanteda

- Package
- https://joss.theoj.org/papers/10.21105/joss.00774
- Application
- https://tutorials.quanteda.io/

STM

- Package
- https://cran.r-project.org/web/packages/stm/vignettes/stmVignette.pdf
- Application
- https://warin.ca/shiny/stm/#section-the-structural-topic-model

Références + ressources

- https://www.irit.fr/IRIS-site/images/seminairs/Thonet2016.pdf
- https://www.tidytextmining.com/
- https://www.tidytextmining.com/sentiment.html
- https://www.datacamp.com/community/tutorials/sentiment-analysis-R
- https://www.datacamp.com/community/tutorials/R-nlp-machine-learning