UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Lab 02:

- Structural Topic Models
- R

Bilel Benbouzid - Alexis Perrier

19 décembre 2017

Est Républicain

Une délégation comtoise de parents d'élèves de l'enseignement libre participera, ce week-end, au 21e congrès de l'APEL qui se déroulera à Lyon. Comme le thème « Parents-école : un enfant à éduquer » l'indique, la réflexion portera sur la relation entre famille et établissement scolaire. Quatre lycéens participeront également à ce grand rendez-vous de l'enseignement libre. Ils participeront dimanche à un spectacle monté par des jeunes et intitulé « Voyage au coeur de l'imaginaire ».

L'équipe B disputera un match amical, le mardi 31 août, à 19h, à Sommerviller. Départ à 18 h. Le dimanche 29, sur le terrain de Damelevières, elle jouera en amical à 17 h 30 contre Blainville B. Départ à 16 h 45. Dimanche 29 août, les espoirs et les moins de 17 ans de l'US Trailor disputeront les éliminatoires du challenge du Chauffage lunévillois.

Une cinquantaine d'enfants de CE2, CM1 et CM2 ont fait leur rentrée mercredi matin au catéchisme. Une célébration a eu lieu à l'église en présence de l'abbé Aubert et de quelques parents ayant répondu à l'invitation. Marie-Thérèse Conraud, parmi la dizaine de catéchistes, a remis des livres aux enfants et leur a donné des informations et des conseils pratiques sur le déroulement des cours de caté. Les enfants de Granges, de Barbey-Seroux et de Champdray se réuniront au foyer paroissial, les autres à la salle polyvalente de Jussarupt.

Recap

1. Nettoyage / Mise en forme du texte
    1. Ponctuation
    2. Tokenization
    3. Stopwords et réduction du volume

2. Processus de numérisation du texte

Corpus brut

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

Matrice Mots - Documents

Shallow

3. Topic Modeling

Latent Semantic Analysis:

déterministe,

=> factorisation de matrice

Latent Dirichlet Allocation:

Hypothèse: les mots dans les topics;

les topics dans les documents suivent une certaine distribution de probabilité

=> le modele LDA va estimer

les paramètres de la distribution

Recap

1. DataFrame Pandas
2. Nettoyage du corpus
    A. Ponctuation
        Characters = !@#$%^&*()
        translator
    B. Tokenization
        nltk.word_tokenize
    C. Stopwords
        nltk.stopwords.words('fr')
        stop_words
        [w for w in tokens if 2 not in stopwords]
    D. Enlever les paragraphes trop long ou trop court
3. LDA / LSA
        num_topics

1. Réduire le bruit pour accroître le signal
2. Ne pas hésiter à *jeter* du contenu
3. Liberté de décision
    1. Ponctuation, chiffres, majuscule,
    2. Mots vs bigrams
    3. Filtrage tokens (stopwords, longueur, …)
4. Variabilité du contenu
    1. Résultats du semi marathon, …

STM en R

Il existe plusieurs package de topic models en R (topicmodels)

Pourquoi ce package STM et pas un autre?

Il permet de:

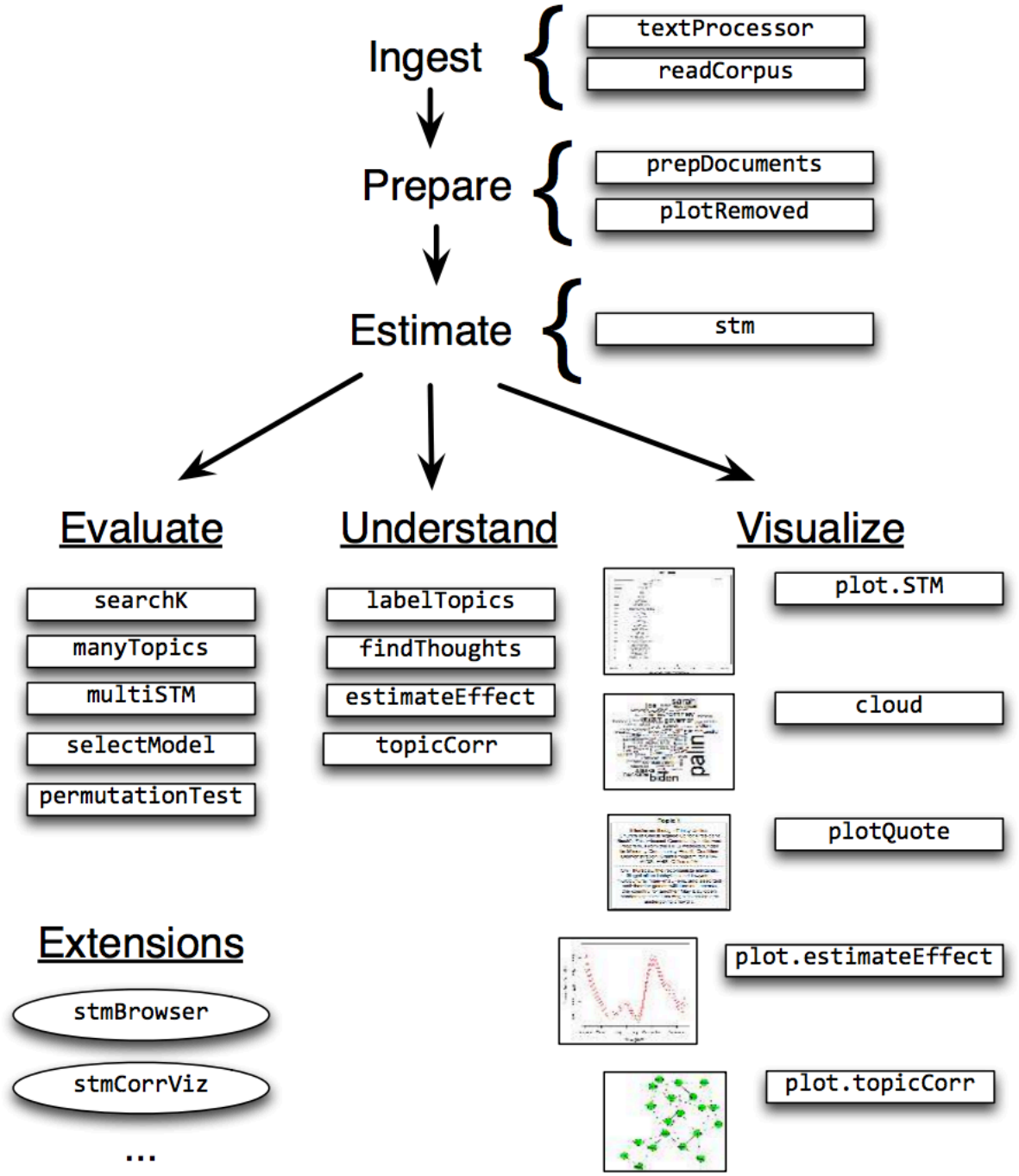
- Déterminer simplement le nombre optimal de topics
- Analyser l'impact de variables externes sur les topics
- Préparer les textes (stem, token, …)

Il possède de nombreuses méthodes d'exploration des résultats

- http://www.margaretroberts.net/ et al
- http://www.structuraltopicmodel.com/
- https://cran.r-project.org/web/packages/stm/vignettes/stmVignette.pdf

Fonctionalités

- Préparation du texte:
  - *textProcessor*
  - *prepDocuments*
- Topic Modeling:
  - *STM*
- Exploration:
  - *plot.STM*
  - *findThoughts*
  - *labelTopics*
  - +++



Qualité des topics

STM définit 2 métriques parlantes:

Semantic Coherence: *maximized when the most probable words in a given topic frequently co-occur together.*

$$C_k = \sum_{i=2}^{M} \sum_{j=1}^{i-1} \log \left(\frac{D(v_i, v_j) + 1}{D(v_j)} \right)$$

FREX exclusivity: *The harmonic mean ensures that chosen terms are both frequent and exclusive, rather than simply an extreme on a single dimension.*

We use the FREX metric (Bischof and Airoldi 2012; Airoldi and Bischof 2016) to measure exclusivity in a way that balances word frequency.[12] FREX is the weighted harmonic mean of the word's rank in terms of exclusivity and frequency.

$$\text{FREX}_{k,v} = \left(\frac{\omega}{\text{ECDF}(\beta_{k,v} / \sum_{j=1}^{K} \beta_{j,v})} + \frac{1 - \omega}{\text{ECDF}(\beta_{k,v})} \right)^{-1} \quad (6)$$

where ECDF is the empirical CDF and ω is the weight which we set to .7 here to favor exclusivity.[13]

Nombre de topics

Choix par défaut

Ou grid search

Le code R

Lire et stocker les donnees dans une dataframe: **read.csv**

- **textProcessor**: preprocessing du texte, transformations
  - removestopwords : filtrer les stopwords
  - removenumbers , removepunctuation : enlever les chiffres et signes de ponctuation
  - striphtml : enlever les tags HTML,
  - stem : ne garder que la racine des mots (non utilisé)
- **prepDocuments**:
  - Création des structures requises pour STM,
  - Filtrage des mots trop ou peu frequents
- Définir les variables externes: *rubrique et journal*
- **stm**: trouver les topics
- **Analyser** les resultats avec:

- labelTopics
- plot.STM
- topicCorr
- topicQuality
- Cloud
- stmBrowser
- findThoughts
- findTopic

Code et experiences

Structurer le code dans une logique d'experimentation et de traçabilité

initialize.R	Charge les packages
config.R	Set les variables de l'experience: • Le numéro de l'expérience • Nom des fichiers input / output • Filtrage des mots • …
Notebook ou script R	Le code

Install R et RStudio

* R disponible par exemple ici: https://cran.univ-paris1.fr/

(voir https://cran.r-project.org/mirrors.html pour la liste des mirroirs)

* R studio, version free disponible sur https://www.rstudio.com/products/rstudio/download/

Et il faudra installer les packages suivants:

```
install.packages( c("stm", "tm", "splines", "stmBrowser"), dependencies = TRUE)
install.packages( c('wordcloud', 'igraph', 'data.table'), dependencies = TRUE)
install.packages( c("stringr", "RColorBrewer", "stringr"), dependencies = TRUE)
install.packages( c("geometry", "Rtsne", "GetoptLong"), dependencies = TRUE )
```