Lab 03:

- Facebook: God Emperor Trump

Bilel Benbouzid - Alexis Perrier

20 décembre 2017

Corpus Facebook God Emperor Trump

Milestones

2017	250,000+ PATRIOTS!
2016	200,000+ Patriots
	100,000+ Beautiful Supporters
	90,000+ Likes
	80,000+ Likes
	70,000+ Likes
	60,000+ Likes
	50,000+ Supporters
	40,000+ Likes on Memorial Day
	30,000+ Likes
	20,000+ Likes
	15,000+ Likes
	10,000+ Likes
	Over 9,000 Likes
	8,000+ Likes
	7,000+ Likes
	6,000+ Likes
	5,000+ Likes
	4,000+ Likes
	3,000+ likes
	2,000 Likes
	100 Likes

315k	319k
Total Likes	Total follows

https://www.facebook.com/GodEmperorTrump/
418790 posts et commentaires,
100+ Mb de donnees en csv

Corpus Facebook God Emperor Trump

In 1945 Russia included all the various countries that now exist individually (Ukraine, Georgia etc. Their majority wants to be back with Russia after a failed separate system. And to say ""Cold War"" (Joseph Stalin) in reference to Putin is like saying Trump thinks the same way Obama does.

She's not allergic to god emperor Donald J. Trump she is allergic to her own bullshit

I was all like "I hate politics, everybody always sucks." Now I be like "Bring the Trumpenreich!!!"

A stoned populace is a more compliant populace. Why do you think the democrats have disseminated bullshit 'studies' on the 'benefits' of dope and are trying to legalize it?

From brexit Britain with love this is just more msm bullshit & labour has went the same way as the Democrats they have went so far to the left that they are now irrelevant & I love it I love it I fucking love it God, guns, putin & trump bitches haha

Difficultés

Séparer le bruit de l'info

Noms de personnes:

- Info: Trump, Clinton, Bernie, Zuckerberg
- Bruit: tags de copains (cindy smith, john doe, bruce Johnson, …)

Argot

- Info: murica, dindu nuffin, guac bowl, cuck, (urbandictionnary.com très utile)
- Bruit: lmfaoooooo, lolololololol, xdd, …

Abreviations:

- Info: fb (facebook), swj (social justice warrior) , blm (black lives matter)
- Bruit: thatll (that will), shed (she would), I'm (I am)…

Stopwords: trouver le juste équilibre: enlever les mots trop fréquents sans toucher aux mots signifiants

Les noms importants

Voici toutes les variantes des noms pour Trump, Clinton et Zuckerberg

Donald Trump, Trump, God King, God Trump, God-Emperor, **EMPEROR TRUMP**, Emperor Palpatine, Emperor Trump, Donald J. Trump, Donald J Trump, Drumpf, Baron Trump, The Emperor, Der Donald, Lord Emperor, **Fuhrer Trump**, #Trump, Adolf Trump, **Adolf Trumper**, Donaldus Magnus, **Trumpachu**, Trumpaloompas, Trumpamaniacs, Trumpanzee, **Trumpasaurus Rex**, Trumpboner, Trumpborne, Trumpbot, Trumpcloaks, Trumpen Reich, **Trumpenreich**, Trumpenverse, Trumpepe, Trumper, Trumpeteer, Trumpeter, **Trumpettes**, Trumpf, Trumpfucker, Trumph, **Trumphalla**, Trumpidari, Trumpinator, Trumpinreich, Trumpire, Trumpis, Trumpivia, Trumpkin, Trumpkins, **Trumplon**, Trumpman, Trumpmanía, Trumpmas, **Trumpmeister**, Trumpmoji, Trumpo, Trumpophiles, Trumpover, Trumpp, Trumps, Trumpsherd, Trumpster, Trumpstika, Trumptrain, **Trumpulus** , Caesar, Trumpumvirate, Trumpus, DJT, Drumft, Drumpf

Shillary, Hillary Clinton, Crooked Hillary, Killary Clinton, Hillary, Clinton, Hilary, Killary, HRC, Killiary

Mark Zuckerberg, Zuckerberg, zucc, Mark Cuckerberg, Mark Fuckerberg, Cuckerberg, Cuckerbrg, Suckerberg, Fuckerberg, Mark Zuck, Zuck

STM sur corpus Brut

- Fichier data: fb_get_sample_01.csv
- Script R: lab03/stm_01.R

Python to the rescue!

Lol omfg. Did I just get called a sjw? Vote Drumpf! Not Killary

```
[lol, expletive, call,  vote, donald_trump, hillary_clinton, social_justice_warrior]
```

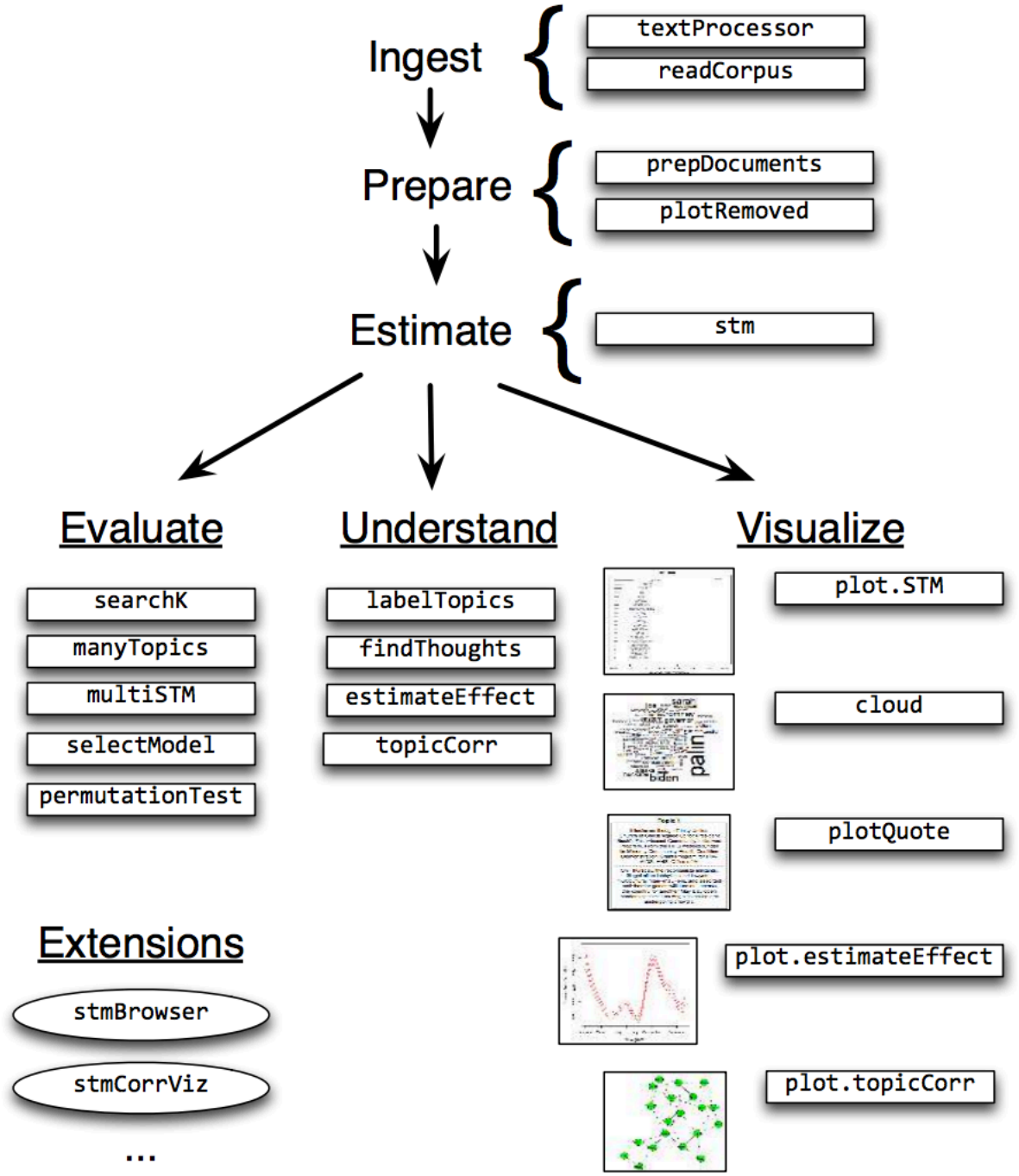
lol expletive call social_justice_warrior vote donald_trump hillary_clinton

Booster l'arsenal:

- La base: distribution anaconda, jupyter notebooks, pandas
- NLTK: tokenization et bien d'autres choses
- Spacy.io: Part-of-speech tagging, Named entity recognition, tokenization, …
- Inflect: pluriels, singuliers, string transformations
- Enchant: spellchecking library for Python
- Regex: pour réduire certaines strings *lmaooooo => lmao*
- textBlob, vader pour sentiment analysis
- Wordnet et Sentiwordnet: synonyme, proximité, signification

Fonctionalités

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



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 )
```