NLP avec r et en français - un Manuel synthétique

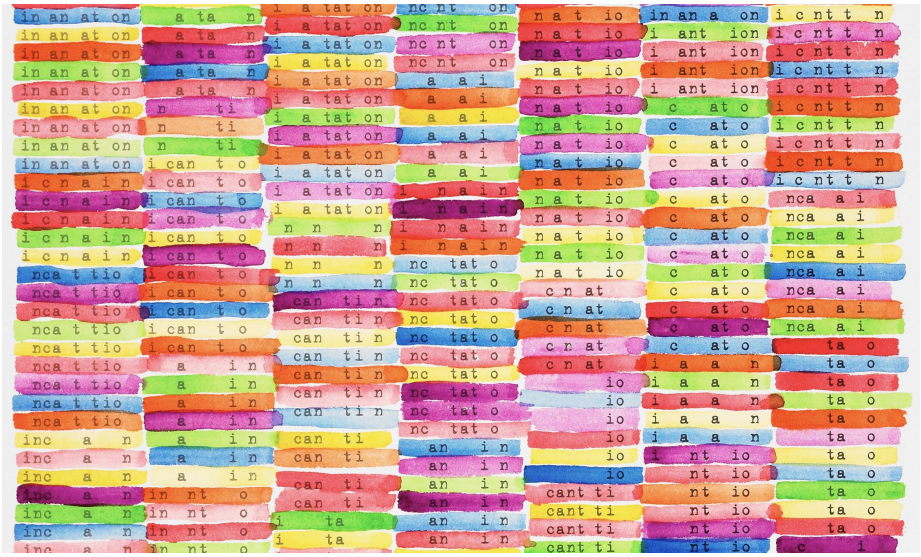
Sophie Balech et Christophe Benavent et al

2021-07-22

Contents

Chapter 1

Préface



[1]

L'eco système r s'est enrichi ces dernière années à grande vitesse dans le domaine du traitement du langage naturel, l'objet de ce manuel a pour but d'en donner une synthèse. Sa vocation est pratique même si on y laissera germer quelques considérations plus méthodologiques, voire épistémologiques. On ouvrira cependant chaque fois que c'est possibles aux questions théoriques et éthiques de ces méthodes. Leur réalisation computationnelle est le fruit souvent d'une longue histoire, au cours de laquelle les linguistes ont semé des idées essentielles qu'ont systématisé les informaticiens.

[1] Incantation for 6 voices Scott helmes, 2001. Museum of Minessota

On soignera la bibliographie de manière synthétique pour en faire un état de l'art essentielet actualisée.

La rédaction de l'ouvrage est mené avec une règle de reproductibilité et de transparence, c'est le pourquoi le choix de ce support et des jeux de données associés.

Il sera dynamique, modifié à mesure de nos cours, séminaires, ateliers et observations des lecteurs.

1.1 Cours et séminaires

La liste des cours et séminaires où il sera présenté et utilisé.

- Colloque Marketing digital 1-2 septembre 2021
- AFM decembre 2021
- Ed Sorbonne - février 2022
- Dauphine master 204 - octobre 2021
- Master Siren - Dauphine - mai 2022
- Toulouse
- Lille master Data Science

1.2 La structure du livre

L analyse NLP peut être analysée comme un processus qui va de la production jusqu' à la diffusion des analyses. Elle est aussi traversée par des évolutions profondes de méthodes qui ont complexifié au sens formel les modèles initiaux. L'apprentissage submerge le comptage,et les catégorisations….

Rappelons nous que les modèles de langages désormais distribués par les grands acteurs, comprennent des dizaines, voir des centaines de milliards de paramètres.

Le plan suit une logique qui va du simple au très compliqué, et de l'acquisition des données, de leur traitement et leur modélisation, jusqu'à la propagation….

- acquisition des données : directe, api et scrapping
- corpus dtm et cooccurence
- AFC et typologie
- l'annotation syntaxique et lexicale
- analyse du sentiment et sa généralisation
- word embedness
- factorial models
- Topic analysis
- ML

- deep learning
- translation : parsceque qu'il faut traiter des corpus multi lingual et que la communication peut aussi etre multilinguales.
- génératives : parce que la prochaine étape c'est quand on appliquera ces méthode sur la productions textuelles des bots.

1.3 Les jeux de données

Au cours du développement, plusieurs cas pratiques - souvent réduit en volume pour rester exemplaires, seront employés. Les donées seront partagées.

En voici la présentation systématique.

- Trump Twitter Archive : L'intégralité des tweets de Trump jusqu'à son banissement le 8 Janvier 2021.
- Confinement Jour J
- Citations : un recueil de citations littéraires pour de petits exemples et ponctuer le texte aride d'un peu de littérature et de poésie.
- Trip advisor Polynésie, un extrait d'un corpus établi par Pierre Ghewy et Sebastien de l'UPF
- Airbnb
- Covid

disponibles dans le repositery avec le code du book. Les amendements et améliorations sont souhaitées et attendues.

1.4 Les ressources

Ce *livre* est écrit en **Markdown** (**?**) et avec le package **Bookdown** (**?**)

Le code s'appuie très largement sur **tidyverse** et emploie largement les ressources de **ggplot** et **dplyr** . On recommande au lecteur de consulter donc les ouvrages suivants quand il s'interroge sur la construction des graphiques. On part du parti-pris que les lecteurs ont une connaissance satisfaisantes de ces outils génériques. Une mention particulère doit être faite sur la question du traitement du texte, **stringr** est aussi un des outils fondamentaux,

- rmardown
- ggplot
- dplyr
- stringer

###les packages

Les packages seront introduits au fur et à mesure. En voici la liste complète.

```
knitr::opts_chunk$set(echo = TRUE, message=FALSE,warning=FALSE)

#boite à outils et viz
library(tidyverse) # inclut ggplot pour la viz, readr et
library(cowplot) #pour créer des graphiques composés
library(ggridges) # le joy division touch
library(citr)

#networks
library(igraph)
library(ggraph)

# Accéder aux données
library(rtweet)  # une interface efficace pour interroger l'api de Twitter

# NLP
library(tokenizers)
library(quanteda)
library(quanteda.textstats)
library(quanteda.textplots)
library(udpipe) #annotation syntaxique
library(tidytext)
library(cleanNLP) #annotation syntaxique

#sentiment
library(syuzhet)              #analyse du sentimeent
#mise en page des tableaux
library(flextable)


#statistiques et modèles
library(lme4)
library(jtools)
library(interactions)

#ML
library(caret)

#graphismes
theme_set(theme_bw())
```

```
#palettes
library(colorspace) #pour les couleurs

# chapitre II
library(revtools)
library(rvest)
```

1.5 Disponibilité

L'ensemble du code est disponible sur github. A ce stade c'est encore très embryonnaire. Les proches pourrons cependant y voir l'évolution du projet et de la progression

1.6 conventions

Quelques conventions d'écriture du code r

- On appele les dataframes de manière générale `df`, les tableaux intermédiaires sont appelé systématiquement `foo`
- Gestion des palettes de couleurs ** une couleur :" royalblue" ** deux couleurs ** 3 à 7 couleurs
- On emploie autant que possible le dialecte tidy.
- Les chunks sont notés X, le chapitre, 01 à n, les jeux. 502 est le second chunk du chapitre 4.
- On commente au maximum les lignes de code pour épargner le corps du texte et le rendre lisible

1.7 A faire

todo list :

- insérer un compteur google analytics (voir https://stackoverflow.com/questions/41376989/how-to-include-google-analytics-in-an-rmarkdown-generated-github-page)
- modifier le titre en haut à gauche
- vérifier le système de références voir (https://doc.isara.fr/tuto-zothero-5-bibtex-rmarkdown-zotero/)
- Vérifier la publication en pdf
- restructurer le plan

Chapter 2

Introduction

Le texte connaît une double révolution. la première est celle de son système de production, il se produit désormais tant de textes que personne ne peut plus tous les lire, même en réduisant son effort à sa propre sphère d'intérêt et de compétence, la seconde est celle de sa lecture, c'est une lecture conditionnée et recommandée..

La production primaire de textes voit son volume croître exponentiellement. Prenons quelques exemples :

La production se soumet ensuite à ceux qui en contrôlent les flux et en exploitent les contenus, qui les mettent en avant ou les écartent, définissant la composition de ce que chacun va lire. La diffusion de cette production suit des loi puissances, c'est ainsi que la révolution de la lecture est venue avec les moteurs de recherche, et les pratiques de curations (ref), c'est une lecture sélectionnée et digérée par les moteurs de recommandation. (ref).

S'il ne fallait qu'un exemple on pourrait évoquer la transformation radicale de la littérature dite scientifique sur le plan technique. La recherche par mots clés est complétée de plus en plus par des outils de veille, l'indexation a donné naissance à l'immatriculation de la moindre note, les fichiers ont adopté des standards, l'interopérabilité est de mise, le réseau des co-citations est maintenu en temps réel. Les scores qualifient autant les articles que leurs auteurs et les revues qui les accueillent.

Elle risque de ce poursuivre par la production de résumés, la transcription automatique (speech2tex) etc.

Le NLP est au coeur de ces technologies, il se nourrit de plus en plus d'intelligence artificielle. Nous en verrons de nombreux exemple à tout les stades du traitement : identifier la langue, mesurer le sentiment, isoler des sujets.

Le NLP est aussi une nouvelle ressource pour les chercheurs en sciences sociales à la fois par les matériaux empiriques et les méthodes d'analyse. C'est une mouvemenjt qui affecte toutes les shs. L'emballement de la production de texte génère une nouvelle matière d'étude pour le sociologues, le gestionnaire, l'économiste, le psychologue pour n'évoquer que quelques disciplines.

2.1 Une réflexion ancienne et un nouveau champ méthodologique

On se doit pas se faire aveugler par l'éclat de la nouveauté, les techniques d'aujourd'hui dépendent d'idées semées depuis longtemps dans au moins deux champs disciplinaire la linguistique et l'informatique

Les pratiques et techniques que nous allons étudier ne tombent pas de mars mais résultent de plusieurs flux de pensées qui se croisent se confortent et amène l'energie pour créer un nouveau bras dans le champs immensement étendu de l'étude de la langue et du langage. Et c'ets sans doute par celà qu'il faut commencer. La langue c'est l'ensemble des règle formelles et moins formelle qui constitue une parole, ce qu'on se dit de l'un à l'autre ou de l'un à aux autres. Le langage est la production de cette parole. L'inscription de cette parole par l'écriture constitue le texte. Le miracle du passage de la parole au signe est celui du symbole.

Si dans ce manuel, on choisit de présenter les différentes facettes de ce qui s'appelle TAl, NPL, Text Mining, dans une approche procédurale qui suit les principales étapes du traitement des données. On rendra compte à chaque étape des techniques disponibles, et on illustre d'exemples. Nous suivrons ici une approche plus fidèle au processus de traitement des données, lequel peut connaître une stratégie inférentielle et exploratoire - quelles informations sont utiles au sein d'un corpus de texte -, tout aussi bien qu'une stratégie hypothético-déductive. Nous resterons agnostique sur cette question, restant délibérément à un niveau technique et procédural.

2.1.1 L'héritage linguistique

la convergence de deux grands mouvements.

Sans en faire l'histoire minutieuse que ce domaine réclame, nous pouvons au moins rappeler un certains nombre d'étapes et de contributeurs clés.

La langue et le langage sont l'objet d'une interrogation millénaire mais quelques auteurs clés ont mis à jour les idées essentielles qui justifient l'usage des méthodes actuelles. Donnons en un aperçu rapide de manière historique, en bornant un champs de connaissance que nous ne pouvons qu'affleurer.

- les sophistes : plier le langage à ses intérêy est une première sciences du langege qui produit une connaissances des dispositifs les plus efficaces. Par sur que cette displines aient trouvé un chemine de vérité,; mais elle ereste commune, c'est l'oeuvre de la publicité.L
- Saussure : il apporte une idée fondamentale que dans le symbole, le signe et le signifiant sont les deux face d'une même monaiie, qu'il existe une relation entre l'artefact et l'idée. Qu'un signe particulier puisse signifier une idée. c'est un penseur de la correspondance.
- Frith et l'idée distributionnelle. un mot trouve son sens dans ceux qui lui sont le plus associés
- Zipf
- Tesnière et les arbres syntaxiques.
- Chomsky et sa grammaire génératitive. enracinant le phénomène linguistique dans la cortalisation du langage, il apporte une idée forte et structuraliste d'un équivalence des langues.
- Genette et l'intertextualité, le palimpseste. c'est une question de sens, le sens d'un texte vient de ses prédecesssurs de ceux à qui ils se réfèrent. Les textes se parlent l'un l'autre, et ce n'est pas dans leur contenu qu'on trouvera une vérité dans dans le rapport qu'ils établissent avec leur prédecesseur par l'appareil des notes et des bibliographies.
- Austin et l'idée que le langage n'est ^pas que communication mais performation . ce qu'on dirt agit sur le monde
- La narrativité

2.1.2 la tradition lexicologique

le lexique est affaire ancienne, le français est aidé par des expériences les fondamentales : le littré, l'académie française et les dictionnaires des éditeurs. pour étudier un lanage il faut se rapporter à des formes stables, les dictionnaires les fournissent et fournisse les normes pour les coder.

L'idée de quantifier le langage n'est pas nouvelle. Encore moins s'il faut compter les occurences et les cooccurences des mots.Un vaste mouvement s'est formé dans les années soixante autour de la lexicologiue stimulée par l'école française d'analyse de données. Le descendant de ce mouvement se retrouve dans l'excellent iramutek de l'équipe de toulouse, il a été précedé par le fameur Alceste.

Nous y consacrerons un chapitre plein sur le plan technique. Mais il est important de souligner que cette école française de l'analyse textuelle ne se limite pas au comptage. Un logiciel comme trope qui d'ailleurs ne connait aucun équivalent dans l'écosystème que nous allons explorer manifeste aussi cette inventivité.

S'y exprime pleinement la logique distributionnelle.