Introduction aux méthodes de traitement du langage naturel

pour les sciences sociales

Sophie Balech et Christophe Benavent

2024-10-15

Table of contents

Préambule

"Au commencement était le verbe", et désormais le verbe est partout, non pas l'esprit de Dieu, mais la parole humaine qui s'éteignait avec le vent, la rumeur mais qui désormais, accumule ses traces imprimées de manière systématique, non plus des petits mots passés de main en main, mais l'enregistrement des transactions informatiques. Le verbe est désormais une copie du monde, moins l'affirmation d'une vision, que la stratification de nos manifestations sociales.

C'est bien sur le fruit d'un développement technique engagés depuis 10 000 ans et dont l'imprimerie est une nouvelle étape. La révolution actuelle réside dans la colonisation de l'espace social, chaque bribes de parole est enregistrée , transcrite, archivée. Le livre n'est plus qu'un îlot dans un océan de texte.

Il y a un défi empirique pour les sciences sociales, la société produit massivement la documentation de son développement. L'exploitation de ces sources est un enjeu majeur pour la psychologique, sociologique, économique, le droit, les sciences de gestion en ouvrant un nouvel accès au données.

A mesure que ces données prolifèrent, des méthodes pour les analyser se sont développées à grande vitesse, moins pour des motivations d'études que pour répondre à des besoins opérationnels. Aujourd'hui, dans le sillage de l'invention des embeddings, nous sommes à l'ère des grands modèles de langages qui ouvrent de nouveaux horizons dans la capacité de produire du texte (par exemple des descriptions d'images), de transformer le texte (résumé ou traduction), et surtout de l'annoter (classification, NER…)

Le but de l'ouvrage

Ce e-book est le syllabus d'un cours que nous dispensons sous différents formats et avec différents degrés d'approfondissement. Il fait l'inventaire des méthodes les plus courantes incluant le développement des embeddings et aboutit à l'usage des LLM pour différentes formes d'annotations. Il a une vocation pratique, étudier des cas, et décrire des codes.

Les contributeurs

On ne peut (encore) en produire la liste exhaustives, mais plusieurs générations d'étudiants ont contribué à ce travail en explorant un certain nombre de jeux de données proposés ici.

Le plan du cours

Il s'organise selon les grandes étapes du processus d'analyse et des types de problèmes posés par le traitement et l'analyse du texte. Ce processus va de l'acquisition des données à leurs représentations en passant par des phases de transformation, mais aussi un ordre de complexité des modèles et des ressources.

Il suit ainsi une sorte de développement historique qui se construit par l'accumulation de différentes philosophies d'analyses et de méthodes et qui peut se présenter en trois grandes périodes:

- compter les mots et jouer avec leur co-occurences
- annoter des mots en s'appuyant sur leurs régularités syntaxiques.
- encoder les mots dans un espace vectoriel, cette perspective a été ouverte par word2vec , étendue avec les transformers, systématisée par les grands modèles de langages tels que gpt4 ou Bloom.

Voici les raisons qui organisent le cours en 20 chapitres (c'est à ajuster au cours de la rédaction)

- Préambule (ce que vous êtes en train de lire !)
- Chapitre 1 : une introduction générale à des éléments linguistique et technique et sociaux de l'analyse du langage naturel et de ses fulgurantes évolutions au cours de la dernière décennie.
- Chapitre 2 : Constituer des corpus. Pour les sciences sociales, le texte est un document qu'on étudie par collection. Il faut aussi penser à la notion de corpus.
- Chapitre 4 : corpus - techniques avancées D'un point de vue matériel le corpus est aussi
- Chapitre 5 : Explorer et naviguer dans le corpus
- Chapitre 6 : Analyse quantitative du corpus
- Chapitre 7 : Tokenisation et dtm
- Chapitre 8 : Analyse des co-occurences
- Chapitre 9 : au début il y avait l'analyse factorielle
- Chapitre 10 : SVD et LSA :
- Chapitre 11 : Topic model
- Chapitre 12 : word2vec to doc2vec
- Chapitre 13 : Transformers

- Chapitre 14 : Supervised modeling
- Chapitre 15 : LLM
- Chapitre 16 : NER
- Chapitre 17 : abstracting
- Chapitre 18 : generative models , un monde à inventer.

Data sets

On s'appuiera sur des data test "réel" et publics

- **Trump Tweeter Archive** : c'est un site qui rassemble tous les tweets émis par Donald Trump depuis 2010, et propose un outil de navigation.
- **Tripadvisor en Polynésie** constitués par Gwevy et Chabrier de l'Université de Polynésie.
- **Airbnb Paris 2023** à partir de Inside Airbnb
- **PMP40ans** : ce set de données comprend tous les résumés d'articles publiés par la revue Politique et Management Public sur une période de 40 années de sa naissance en 1983 jusqu'en 2023. C'est un observatoire intéressant d'une discipline naissance et de l'évolution des normes professionnelles de la recherche en gestion.
- **RapLyrics** : Une collection de texte de rap français des années 80 à 2020 constituées par un groupe d'étudiants.

Packages

Ce cours utilise les ressources de r, de rstudio et de Quarto, de l'analyse des données à la publication de ce book. On utilise Zotero pour gérer la bibliographie.Il n'y a jamais une solution unique, nous avons fait des choix de méthodes qui se reflètent dans celui des packages dont voici la liste commentée. Pour l'usage des modèle LLM on basculera progressivement en python qui propose immensément plus de ressources.

```
#environnement de travail
library(tidyverse) # the perfect tool box. Comprend dplyr pour la gestion des données, ggplo
library(flextable) # pour produire des tableaux élégants

#ses compléments
library(ggwordcloud) ##complete ggplot pour les nuages de mots
```

```
library(ggrepel) ## pour éviter la superposition de label

#glosaire
library(glossary)
glossary_path("glossary.yml")
# un exemple pour ajouter des termes au cours de la rédaction
#glossary_add(term = "loi de Dirichlet",
#             def = "la loi de Dirichlet, souvent notée Dir( ), est une famille de lois de pr

#ressource nlp
library(quanteda)
library(udpipe)

#analyse de données
library(FactoMineR)
library(factoextra)

#python
library(reticulate)
```

Ressources complémentaires

- r en français
- On encourage vivement le lecteur à jouer avec les modèles de Hugging Face
- Le Stanford Natural Language Processing Group est sans doute la principale institution scientifique dans ce domaine.
- En France : Acss-PSL, Ollion, science po ...

Pour aller plus loin, il est possible de suivre le NLPWorkshop de Acss ou le cours " introduction au NLP pour les sciences sociales " dispensé chaque année en novembre au sein de la PSL week.

1 Introduction

Le texte connaît une double révolution. La première concerne son système de production : il se produit désormais tant de textes que personne ne peut plus tous les lire, cela même en réduisant son effort à sa propre sphère d'intérêt et de compétence, la lecture a besoin d'une béquille, celle d'être conditionnée par les moteurs de recherche et de recommandation. La seconde est qu'au-delà de sa production, le texte est amené à être retraité, à être absorbé par les machines, réduit, condensé, pour se lire moins dans la singularité du texte, que dans sa synthèse algorithmique. Le texte s'enfouit dans des corpus.

La production primaire de textes voit aujourd'hui son volume croître exponentiellement. Elle se compare à la transition moderne, où le texte était d'abord copié de manière manuscrite, avant de connaître une recopie plus standardisée, associée à l'invention et l'essor de l'imprimerie. Ce qui était alors avant destiné à être reproduit, était le résultat d'un processus long et exigeant, qui permettait à un petit groupe de lettrés et d'imprimeurs de contrôler l'essentiel de ce qui pouvait être lu. En ce sens, la "révolution digitale" permet à un ensemble encore plus grand d'individus, via des interfaces numériques simples, de confier sous forme d'écrits, leurs états d'âme, réflexions et autres opinions. Cette production se soumet donc à ceux qui en contrôlent les flux et en exploitent les contenus, les mettant en avant ou les écartant, définissant ainsi la composition de ce que chacun va pouvoir lire. La diffusion de cette production suit des lois puissances, c'est ainsi que la révolution de la lecture est venue avec les moteurs de recherche, et les pratiques de curation (ref) : une lecture sélectionnée et digérée par les divers moteurs et algorithmes de recommandations. (ref).

S'il ne fallait citer qu'un exemple, on pourrait évoquer la transformation radicale de la littérature scientifique, dont le niveau de production croit de 4% par an environ, doublant en moins de 20 ans. aujourd'hui presque tous les dix ans. (Bornman 2021) .A cette production exponentiellement croissante, s'ajoute un effort d'inventaire : . Des standards sont proposés, l'indexation a donné naissance à l'immatriculation systématique de la moindre note, l'interopérabilité est de mise, le réseau des co-citations se maintient en temps réel. Différents scores qualifient autant les articles que leurs auteurs, comme les revues qui les accueillent. Le monde de la recherche, processus par nature plus exploratoire, allant vers l'inconnu, est désormais totalement balisé et quantifié. Le volume généré est si grand, que la production automatique de résumés, revues bibliographiques et autres synthèses, se généralise.

Le Natural Langage Processing (en français Traitement Automatique du Langage) est au cœur de ces technologies, et se nourrit massivement de l'intelligence artificielle. Chatgpt en est le dernier artéfact spectaculaire. Nous en verrons de nombreux exemples à tous les stades du

traitement : identifier la langue, mesurer le sentiment, isoler des sujets, calculer une relation syntaxique, évaluer une intention, détecter une tonalité…produire du texte.

Le NLP est aussi une ressource pour les chercheurs en sciences sociales, tant par les matériaux empiriques nouvellement mis à disposition, que par les méthodes d'analyse proposées. C'est un mouvement qui affecte toutes les sciences humaines. L'emballement de la production de texte génère une nouvelle matière d'étude pour les sociologues, gestionnaires, économistes, juristes et psychologues.

Si dans ce manuel, on choisit de présenter les différentes facettes de ce qui s'appelle le "TAL", le "NLP", le "Text Mining", dans une approche procédurale qui suit les principales étapes du traitement des données, nous rendrons compte à chaque étape des techniques disponibles, que l'on illustrera d'exemples. Nous suivrons ici une approche plus fidèle au processus de traitement des données, lequel peut connaître tant une stratégie inférentielle et exploratoire (Quelles informations sont utiles au sein d'un corpus de texte ?), qu'une stratégie plus hypothético-déductive. Sur ces questions portant sur son usage, nous choisissons d'être délibérément agnostique, préférant présentement de rester au niveau plus technique et procédural des outils de recherche.

Il a eu une ère Gutenberg, nous sommes à celles des embeddings. l'imprimerie a transformé le monde en répliquant le texte, les embeddings, en le compressant.

1.1 De la linguistique à l'approche computationnelle

On ne doit pas se faire aveugler par l'éclat d'une apparente nouveauté de ces méthodes. Les techniques d'aujourd'hui dépendent d'idées semées depuis longtemps dans champs de recherche : la linguistique et l'informatique.

On peut alors en synthétiser l'idée avec cette figure annotée. elle en exprime deux veines principales. La première, est une tension du champs entre la langue comme structure, quand la seconde considère le langage en tant que capacité et usage.

![Les domaines de la linguistique](./images/linguistique.jpg)

1.1.1 d'abord persuader

Penser la langue est un effort constant qui a commencé il y a de nombreuses années, certainement avec les sophistes, et l'idée qu'en maniant le langage, il est possible de convaincre, en construisant une logique propre (protagoras, Diogène…). Pour les sophistes : plier le langage à ses intérêt est une première sciences du langage qui témoigne d'une connaissance des dispositifs les plus efficaces. Pas sûr que cette discipline ait trouvé une "episteme" reconnue, mais elle n'en reste pas moins commune et contemporaine : c'est l'art de la publicité. La rhétorique

n'est pas une discipline morte, elle se développe de manière concrète dans toute les agences publicitaires.

Donnons quelques points de repère en commençant par quelques définitions préliminaires, avant de se concentrer sur trois idées essentielles qui vont prospérer avec le développement de la linguistique computationnelle et de l'intelligence artificielle. Ces trois idées sont relatives aux principales branches de la linguistique : à savoir la syntaxe, la sémantique et la pragmatique. Nous resterons ici silencieux sur la phonologie (étude de la formation des sons et de la phonétique) dont l'importance est considérable quand il s'agit de traiter la production et les interactions orales. Pour ne donner qu'un exemple, la prosodie (le rythme données aux phrases) est un objet d'étude essentiel dans les courants de recherches en informatique affective.

1.1.2 Langue, langage, texte et parole

La langue se définit comme un ensemble de règles plus ou moins formelles que constitue une parole : ce qui se dit de l'un à l'autre ou de l'un aux autres. Le langage est la capacité à produire cette parole. La constitution de cette parole par l'écriture constitue le texte. Le miracle du passage de la parole au signe est celui du symbole. Parmi les distinctions terminologiques proposées par Ferdinand de Saussure au début de siècle dernier, autour de la langue, du langage et de la parole se sont révélées particulièrement pertinentes et restent toujours utilisées de nos jours.

Le *Langage* : faculté inhérente et universelle de l'humain de construire des langues (des codes) pour communiquer. (Leclerc 1989:15)

Le langage réfère à des facultés psychologiques permettant de communiquer à l'aide d'un système de communication quelconque. Le langage est inné.

La *Langue* : système de communication conventionnel particulier. Par « système », il faut comprendre que ce n'est pas seulement une collection d'éléments mais bien un ensemble structuré composé d'éléments et de règles permettant de décrire un comportement régulier (Pensez à la conjugaison de verbes en français par exemple). La langue est acquise.

Le langage et la langue s'opposent donc par le fait que la langue est la manifestation d'une faculté propre à l'humain, qui n'est autre que le langage.

La *Parole*: une des deux composantes du langage qui consiste en l'utilisation de la langue. La parole est en fait le résultat de l'utilisation de la langue et du langage, et constitue ce qui est produit lorsque l'on communique avec nos pairs.

Le *texte* : Il est la transcription de la parole, même si le plus souvent, sa production est directe sans étape intermédiaire de traduction du langage oral.

1.1.3 Syntaxe et grammaire générative

Nous nous référerons ici à Chomsky et sa grammaire générative. En dépit de leur très grande diversité, le projet s'appuie sur l'idée qu'un nombre fini de règles doit produire une infinité d'énoncés. Une grammaire est générative dans la mesure où elle possède cette propriété.

L'analyse est ainsi tournée vers la compétence, et le linguiste s'intéresse à l'idéal qu'un locuteur qui, en connaissant ces règles, serait en mesure de produire une pluralité de discours.

Observant que les enfants apprennent, enracinant le phénomène linguistique dans la cortalisation du langage, il apporte une idée forte et structuraliste d'une équivalence entre les langues. Sous la lumière de Tesnière et de ses arbres syntaxiques, les treebanks contemporains s'inscrivent dans cette perspective et nourrissent les analyseurs (parser) syntaxiques du langage naturel qui constituent désormais la première couche d'un traitement de données textuelles.La grammaire générative a conduit la linguistique dans un tournant

formel où la langue est étudiée indépendamment de ses locuteurs. On pourra méditer sur le "pourquoi" des algorithmes génératifs contemporains de deep learning (le fameux GPT3) qui réussissent à former des phrases syntaxiquement correctes mais absurdes.

1.1.4 Sémantique : Une conception distributionnelle

La tradition lexicologique file le lexique comme une affaire ancienne. Le français est aidé par des institutions fondamentales : le Littré, l'Académie Française et les premiers dictionnaires des éditeurs. Pour étudier un langage il faut se rapporter à des formes stables, les dictionnaires les documentent et renseignent ces normes pour les coder . Un moment clé a été de penser le signe, Saussure apporte alors cette idée fondamentale que dans le symbole, le signe et le signifiant sont les deux faces d'une même monnaie, qu'il existe une relation entre l'artefact et l'idée. En d'autres termes; il est possible qu'un signe particulier puisse signifier une idée : c'est un penseur de la correspondance.

Selon Saussure, la langue est le résultat d'une convention sociale transmise par la société à l'individu et sur laquelle ce dernier n'a qu'un rôle accessoire. Par opposition, la parole est l'utilisation personnelle de la langue (toutes les variantes personnelles possibles: style, rythme, syntaxe, prononciation, etc.). Le changement de la langue relève d'un individu mais son acceptation relève de la communauté et des institutions. ex.: le verbe « jouer » conjugué «jousent » est pour l'instant à considérer comme une variante individuelle (parole), une exception, et il le demeurera tant qu'il ne sera pas accepté dans la communauté (les locuteurs du français dans ce cas-ci). Sa conception du signe répond à cette approche conventionnelle : la dualité du signe comme signifiant et signifié est opérée.

Dans le traitement des données textuelles le "signifié" est le terme cible de l'analyse, pour en découvrir son signifié on se tourne vers son contexte : l'ensemble des signifiés. C'est une idée ancienne qu'a proposé Firth dans les années 30. Firth (1957) construisant ainsi la genèse du

paradigme distributionnel. Un mot trouve son sens dans ceux qui lui sont le plus associés. C'est, dans cette veine, le contexte qui donne alors le sens. Cette idée va être retrouvée avec la notion fondamentale des embeddings.

L'idée de quantifier le langage n'est pas particulièrement innovante, et ce, moins encore s'il s'agît de compter les mots et leurs co-occurences. Le premier a adopté cette méthodes est certainement Zipf, le père ce cette loi fameuse qui réduit le produit de la fréquence des mots et de leur rang à une constance, étudiant empiriquement cette distribution d'homère à James Joyce de 1934 à 1949 (On lira avec bonjeur la synthèse de Bully). Derrière la régularité statistique il avance un argument celui du locuteur qui tend à utilisé le mois de mots et les plus simple, et celui de l'interlocuteur qui en a besoin de plus précis pour décoder le message.

Un vaste mouvement s'est formé dans les années soixante autour de la lexicologue, stimulé par l'école française de l'analyse de données (benzecri). Le descendant de ce mouvement se retrouve dans l'excellent iramuteq de l'équipe de Toulouse, précédé par le fameux Alceste, et maintenant durablement intégré dans le package R Rainette.

Nous y consacrerons un chapitre complet sur le plan technique. Il reste important de souligner que cette école française de l'analyse textuelle ne se limite pas au comptage d'entités. Un logiciel comme trope qui d'ailleurs ne connait aucun équivalent dans l'écosystème que nous allons explorer, manifeste aussi cette inventivité, où s'exprime pleinement la logique distributionnelle.

1.1.5 L'approche pragmatique : les fonctions et acte du langage

Si la grammaire générative se tourne délibérément plutôt vers la compétence et ignore la performance, c'est à dire la production d'énoncés par les humains en situation d'interaction plus que sur les effets de l'énoncé lui-même, un autre courant de la linguistique s'est emparé de la question, le courant dît de la pragmatique du langage.

Le grand classique de ce courant est la théorie des fonctions du langage, qui sous-tendent la production d'un message : l'acte de parole, proposée par Jakobson. Inspiré par la cybernétique, la structure de son modèle est celle d'un acte de communication. Jackobson identifie les éléments de l'évènement discursif (speech event) et les fonctions qui lui sont associées. Pour le paraphraser, un LOCUTEUR envoie un MESSAGE à un ou plusieurs INTERLOCUTEUR(S) qui, afin d'être compris, requiert un CONTEXTE dont les acteurs de l'évènement discursif sont capables de saisir et de verbaliser, ce qui suppose l'existence d'un CODE au moins partiellement commun et d'un CONTACT, canal physique ainsi que d'une connection psychologique. On listera alors, au sens du théoricien (**jakobson_linguistics_1981?**) à lire [ici](https://pure.mpg.de/rest/items/item_2350615/component/file_2350614/content)

- La fonction référentielle ou représentative (aussi dénommée sémiotique ou symbolique), où l'énoncé donne l'état des choses , où
- le message dénote un contexte. Jakobson emploie aussi les termes de dénotatif ou cognitif.

- La fonction expressive (émotive), où le sujet exprime son attitude propre à l'égard de ce dont il parle.
- La fonction conative, lorsque l'énoncé vise à agir sur le destinataire : elle s'exprime grammaticalement par l'impératif ou le vocatif.
- La fonction phatique, empruntée à Malinoswki (ref) où l'énoncé révèle les liens ou maintient les contacts entre le locuteur et l'interlocuteur.
- La fonction métalinguistique ou métacommunicative, qui fait référence au code linguistique lui-même, qu'il soit théorisé ou internalisé par le locuteur, comme la prose de Monsieur Jourdain.
- La fonction poétique, lorsque l'énoncé est doté d'une valeur en tant que tel, valeur apportant un pouvoir créateur et dont Jakobson illustre avec l'exemple de la jeune fille qui a l'habitude de désigner Harry par "Horrible Harry" sans pouvoir expliquer pourquoi

il ne serait pas l'odieux, le dégoûtant, ou le terrible Harry, alors que sans s'en rendre compte, elle emploie une paronomasie/alliteration : la ressemblance phonologique et prosodique des mots produit un puissant effet poétique. John Langshaw Austin s'intéressant à la fonction conative développe le concept d'acte de langage, introduisant l'idée fondamentale que les actes de langage (la production d'un énoncé) ne sont pas uniquement destinés à décrire le monde tel qu'il est, mais bien à agir sur le monde par le biais du locuteur et du destinataire. Parler devient alors également, faire. La théorie des actes de langage est d'abord une catégorisation des différents actes. Il distingue trois types de réalisations s'opérant au travers du langage :

Le locutoire : cette dimension du langage est réalisée à partir du moment où un énoncé, est juste grammaticalement, dans les règles de la langue dans laquelle il est émis. Prononcer à table, la phrase :

"Est-ce-qu'il y a du sel ?", est une construction langagière correcte, et se réalise dans sa première dimension. Cependant, dans la théorie linguistique de John Langshaw Austin, le message convoyé par un énoncé va au-delà de son sens immédiat, et s'intègre dans une seconde fonction.

L'illocutoire : L'exemple précédent n'a pas, du seul fait de sa formulation, uniquement pour fonction de s'informer sur la présence de sel dans la maison (ou dans le plat, contenu locutoire de l'énoncé). Il exprime plutôt que l'on voudrait saler son plat (fonction illocutoire) et se traduit généralement par le fait que l'un des convives réagisse, par exemple en passant la salière au locuteur. Ce faisant, le langage performe un troisième niveau de discours, qu'Austin nomme la dimension perlocutoire..

Le perlocutoire : Cette dimension finale se conjugue donc avec les deux précédentes, mais son produit n'est pas commutatif, dans le sens où les actions et interprétations sujettes de notre énoncé 1 dépendent desfonctions plus basses de ces derniers, et s'enracinent également dans le contexte de ceux-ci et d'éléments plus ou moins extra langagiers. Cette idée est au cœur des sous-bassements de la théorie des actes de langage d'Austin, qui se détache donc fortement

d'un langage uniquement communicatif, au détriment d'une vision de ce dernier comme un outil plus performatif que descriptif. Dire devient alors faire, car le langage agit et transforme l'univers des interlocuteurs.

1.1.6 le texte

La lanque étudiée en-soi, ne prête guère l'attention à ses support, elle est abstraite et sa matérialisation langagière se traduit par une parole. L'énoncé que le linguistique dissèque est une parole, ce qui sors de la bouche du locuteur.

Mais depuis quelques livres fondateurs et la récolution de l'imprimerie, l'expression de la langue et du langage s'est logé dans le texte. Des convention de caractères, de lettres, de ponctuations et des règles de grammaire et de typographie. La langue se socialise, s'instituitionnalise.

Les textes ne sont pas isolés, ils se répondent l'un à l'autre. La note de bas de page, la référence bibliographique,

Genette et l'intertextualité, le palimpseste. c'est une question de sens, le sens d'un texte vient de ses prédécesseurs et de ceux à qui ils se réfèrent. Les auteurs au travers des textes se répondent l'un l'autre, et ce n'est pas dans leur contenu qu'on trouvera une vérité mais dans le rapport qu'ils établissent avec leurs prédécesseurs, par l'appareil des notes et des bibliographies, des références et mises en perspectives. Cette approche vient questionner l'apparente et rassurante téléologie naturelle que chacun est tenté de voir, dans la remise en continuité d'éléments qui sont alors détachés de leurs contextes de production.

1.1.7 La narrativité .

L'acte de parole se réalise dans un lieu à un moment, avec des protagonistes, dans une atmosphère, avec une histoire, les mots qui s'en échappent ne sont que des traces, autant que des photographies. Ces données se sédimentent dans les grands bassins du cloud hybride et dans les corpus constitués historiquement et méthodiquement.

1.2 Linguistique computationnelle

Les points de contact entre linguistique et informatique se produisent en rapport à diverses questions pratiques, portées sur le traitement et la computation d'éléments langagiers recensés selon des sources tant orales qu'écrites, pour diverses finalités opérationnelles.(traductions, analyses, transcriptions, synthèses…)

Les apports de la fouille de données les nomenclatures

une convergence nécessaire

Le monde des bibliothèques et celui de la GED.

1.2.1 Les facteurs de développement de l'usage en sciences sociales

Ces développements sont favorisés par un environnement fertile où trois facteurs se renforcent mutuellement. Ils conduisent à l'élaboration de nouvelles méthodes.

- La naissance et généralisation de langues informatiques universelles
- L'émergence de vastes ensembles de données textuelles
- La naissance d'une communauté épistémique, de pratique et de

1.2.1.1 Une lingua franca

Le premier facteur de développement est l'expansion de la programmation orientée objet (POO). Plus spécifiquement dans le cas de la manipulation des données, deux langages de programmation se distinguent particulièrement, dans un usage proprement statistique pour R et plus généraliste en ce qui concerne Python. Le propre de ces langages est, prenons le cas de R, de permettre d'accéder à des interfaces et fonctions mathématiques, dont un ensemble cohérent pour réaliser certaines tâches peut être rassemblé dans une bibliothèque appelée "package" ou "librairie". Ces bibliothèques de fonctions se chargent en mémoire facilement via la commande R suivante : library(nomdupackage). On dispose désormais de milliers de packages (17 788 sur le CRAN) destinés à résoudre un nombre incalculable de tâches. Une petite représentation ci-dessous témoigne de l'évolution exponentielle des outils mis à dispositions de la communauté R :

![hornik](./Images/number_CRAN_packages.png)

Développer et concevoir le code d'une analyse revient ainsi à jouer avec un immense jeu de briques, similaires aux Lego de notre enfance, dont de nombreuses pièces bas niveau sont déjà pré-moulées. D'un point de vue pratique, les lignes d'écritures sont fortement simplifiées, permettant à un chercheur non spécialisé en programmation d'effectuer simplement des opérations complexes. En retour, cette facilitation de l'analyse abonde le stock de solutions.

1.2.1.2 La multiplication des sources de données.

Le second facteur d'évolution est la multiplication des sources de données et leur facilité d'accès.

- Le contenu écrit des réseaux sociaux,

L'acte de parole se réalise dans un lieu à un moment, avec des protagonistes, dans une atmosphère, avec une histoire, les mots qui s'en échappent ne sont que des traces, autant que des photographies. Ces données se sédimentent dans les grands bassins du cloud hybride et dans les corpus constitués historiquement et méthodiquement. Les rapports d'activités des entreprises,

- Les compte-rendus archivés de réunion,
- Les avis des consommateurs sur les catalogues de produits,
- Les articles et les revues scientifiques,
- Les livres numériques…

Les sources les plus évidentes sont proposées par les bases d'articles de presse telles que presseurop ou factiva. Les bases de données bibliographiques sont dans la même veine particulièrement intéressantes et pensées pour ces usages.

Les données privées, et en particulier celles des réseaux sociaux, même si un péage doit être payé pour accéder aux plateformes via différentes APIs, popularisent le traitement de données massives. Les forums et sites d'avis de consommateurs sont pour les sociologues de la consommation et les spécialistes du comportement de consommation une ressource directe et précieuse.

Le mouvement des données ouvertes (open data) proposent et facilitent l'accès à des milliers de corpus de données : grand débat, EuropeanSurvey…

1.2.1.3 Une communauté

Le troisième facteur de développement , intimement lié au premier, est la constitution d'une large communauté de développeurs et d'utilisateurs qui se retrouvent aujourd'hui dans des plateformes diverses. Le savoir, autrement dit des codes commentés se trouvent dans une variété importante

de lieux :

- Des plateformes de dépôts telles que Github, qui rassemblent une trentaine de millions de développeurs et data scientists.
- Des plateformes de Q&A (question et réponses) telles que Stalk OverFlow,
- Des tutoriaux de toute sortes : cours, vidéos et autres Mooc
- Des blogs ou des fédérations de blogs (BloggeR),
- Des revues (Journal of Statistical Software) et de bookdown.

Des ressources abondantes sont ainsi disponibles et facilitent l'auto-formation des chercheurs et des data scientists, en proposant des ressources pour la résolution de leurs problèmes pratiques. Quiconque n'arrive pas à résoudre un problème a une bonne chance de trouver la solution d'un autre, à un degré de circonstances près. Elles sont d'autant plus utiles que certaines règles ou conventions s'imposent progressivement pour fluidifier l'échange et les projets individuels :La principale démarche est alors celle de l'exemple reproductible.

La seconde est le maintien d'une éthique du partage qui encourage à partager le code, et dont une littérature importante étudie l'effet positif sur les performances économiques et la durabilité [rauter]. Les

externalités de réseaux y sont fortes.

Toutes les conditions sont réunies pour engendrer une effervescence créative. Python ou R, sont dans cet univers en rapide expansion, les langues véhiculaires qui favorise une innovation constante. Les

statistiques de Github en témoigne : près de 50 millions d'utilisateurs, 128 millions de "repositories" et 23 millions de propriétaires. ![source](./Images/github.png)

voir aussi <https://towardsdatascience.com/githubs-path-to-128m-public-repositories-f6f656ab56b1>

1.2.2 De nouvelles méthodologies pour les sciences sociales

Pour les chercheurs en sciences sociales, et donc nécessairement, pour les chercheurs en sciences de gestion, lieu de rencontre entre toutes les sciences sociales, cette révolution textuelle offre de nouvelles opportunités d'obtenir et d'analyser des données solides pour vérifier leur hypothèses et mener leurs enquêtes. Ce sont de nouveaux terrains, de nouvelles méthodes et un nouvel objet de recherche qui se dessine dans le développement du champ scientfique contemporain.

1.2.2.1 Nouveaux terrains :

La multiplication des sources de données précitées, associées à leur progressive normalisation, permet une prolifération de techniques provenant de multiples courants disciplinaires, convergeant toutes vers un langage commun. En ce sens, la production abondante d'avis de consommateurs, de discours de dirigeants, de compte-rendus de conseils et colloques, d'articles techniques, de travaux en linguistique computationnelle, de diverses fouilles de données, des moteurs de recommandation, de la traduction automatique, offre des ressources nouvelles et précieuses pour traiter l'abondance des données générées.

1.2.2.2 Nouvelles méthodes :

Un nouveau paradigme méthodologique se construit à la croisée de données abondantes et de techniques intelligentes de traitement . Il permet d'aller plus loin que l'analyse lexicale traditionnelle en incorporant des éléments syntaxiques, sémantiques, et pragmatiques, proposés par l'ensemble des outils de traitement du langage naturel. Il se dessine surtout une nouvelle approche méthodologique qui prend place entre l'analyse qualitative, et les traditionnelles enquêtes par questionnaires capables de traiter des corpus d'une taille inédite. Le travail de Humphreys and Wang (2018) en donne une première synthèse dans le cadre d'un processus qui s'articule autour de 6 différentes

phases d'une recherche :

- La formulation de la question de recherche
- La définition des construits,
- La récolte des données
- L'opérationnalisation des construits
- L'interprétation et l'analyse,
- La validation des résultats obtenus.

1.3 Un nouvel objet :

On pourrait croire qu'avec des données massives et des techniques "intelligentes" nous assistons à un retour du positivisme qui bénéficierait enfin des instruments de mesures et de calculs ayant permis à certains chercheurs au plus proche de la matière des succès majeurs. Sans nul doute, l'administration de la preuve va être faciliter par ces techniques et va encourager l'evidence based policy (REF) afin de résoudre en partie la crise de la réplication et de la reproductibilité des travaux de recherche.

Cependant, à mesure que se développe l'appareillage de méthode et de données, moins l'on peut supposer que l'observateur reste neutre. En effet, ni les télescopes géants, ni les synchrotrons, n'affectent les galaxies lointaines ou les atomes proches. Le propre des données que l'on est amené à étudier est de résulter de la confrontation d'un système d'observation (certains préfèrent alors parler de surveillance), à un agent, doué de buts, d'une connaissance, de biais, et de ressources. Le dispositif de mesure est en lui-même performatif.

L'exemple le plus évident est celui des systèmes de notation, qui sous

prétexte de transparence donne la distribution des répondants

précédents. L'agent qui va noter choisit la valeur en fonction d'une

norme apparente - la note majoritaire- et de sa propre intention - se manifester ou se confondre à la foule. Pour se donner une idée plus précise de ce mouvement, examinons quelques

publications récentes dans les champs qui nous concernent.

1.3.1 Sociologie et histoire

classes sociales avec word to vec en sociologie

Kozlowski, Taddy, and Evans (2019)

L'article révolution française

On citera cependant jean-baptiste Coulmont et son obstination à étudier les entités nommées, prénoms et autres marqueurs culturels de l'identité et des classes.et au luxembourg

1.3.2 Psychologie

Très tôt la psychogie s'est intéressée au langage, pas seulement comme produit des processus psychologiques, mais comme expression de ceux-ci.Dès les années 1960 dans le champ de la psychologie de l'éducation, douéd'une forte motivation positiviste, s'est posée la question de la mesurede la difficulté d'un texte pour un niveau d'éducation donné. La mesure de la lisibilité des textes s'est alors développée, profitant à d'autresecteurs tels que ceux de la propagande. Dans cette même perspective,l'approche scientifique de la richesse lexicographique comme concept représentant les compétences a à son tour développé de nouvelles instrumentations.

James W. Pennebaker a développé son approche à partir de l'étude des traumas; donnant une grande importance à la production discursive des patients. Sa contribution majeure est l'établissement d'un ensemble de dictionnaires destinés à mesurer des caractéristiques du discours. Un instrument qu'on présentera dans le chapitre 7 (à vérifier) Tausczik and Pennebaker (2010). Son approche se poursuit en psychiatrie avec l'analyse des troubles du langage, et a connu un coup d'éclat avec la démonstration que l'analyse des messages sur les réseaux sociaux comme facebook permet de détecter des risques de dépression.@eichstaedt_facebook_2018.

1.3.3 Management

La finance et l'analyse du sentiment

Dans le champ du management, on trouvera des synthèses pour la recherche

en éthique Lock and Seele (2015), en comportement du consommateur

Humphreys and Wang (2018) en management public

Anastasopoulos, Moldogaziev, and Scott (2017) ou en organisation

Kobayashi et al. (2018)

1.3.4 Economie

économie des brevets intervention des institutions mesure de l'innovation

2 Des comptables à l'industrie de la langue

La situation nouvelle qui semble être la notre consiste dans le fait que la parole qui disparaissait avant comme emportée par le vent, laisse désormais des traces et s'enregistre. L'ironie est qu'au titre de la protection de la vie privée, cet enregistrement systématique doit être mis à notre disposition. On a le choix : rien n'en faire, les détruire, ou bien encore les donner, afin de bénéficier de son potentiel de connaissance. Nous sommes passé de la parole au texte. Si seule la parole de Dieu et celle des chants étaient transcrites, c'est désormais aussi celle du vulgum. Si sa précision reste incertaine, son volume quant à lui a gagné de nombreuses échelles.

Cette matière ne s'organise plus dans les papyrus, tablettes d'argiles et autres manuscrits, ni même dans les livres sués par les callygraphistes, elle s'incruste dans un édifice de plus en plus complexe d'interfaces textuelles et vocales. La parole est comme absorbée par les machines. Elle ne s'envolent plus avec le vent, elle se sédimente dans des data center, nouveaux monolithes modernes. Le langage a acquis une dimension matérielle qu'il n'a presque jamais connu. Il gagne de l'autonomie avec les systèmes génératifs : chat bots, transcriptions, traductions, résumés.

L'histoire se définit selon son écriture. [Daniel Gaxie, La raison Graphique] L'écriture est le produit d'une société de procès-verbal, de comptabilité et cela se poursuit. Voilà qui facilite le travail de l'historien, du sociologue et de l'économiste.

Dans les années 90 s'est dessinée une société de l'information, sauvage jusqu' à Napters, et le rêve du peer to peer, elle s'est socialisée dans les années 2000, plateformisée dans les années 2010, généralisée pour la décennie qui nous concerne. Toute cette architecture s'appuie sur les données qu'on y injecte, et au premier rang, y siège le texte, la transcription automatique de la parole, dans une recodification

constante et des traitements de plus en plus hyperconcentrés.

3 Conclusion

Dans ce manuel on va privilégier le “comment faire”.Cependant on se donnera des espaces de réflexion et d’interrogation, que ce soit sur une certaine philosophie du langage, ou sur les paradigmes techniques ou des questions plus anthropologiques.

Une première parenthèse est expérientielle, c’est en faisant que nous avons découvert une autre écriture. L’expérience de ce livre, qu’on partage avec de nombreux utilisateurs de ces nouveaux outils, est celle d’une écriture programmatique, performative. Ecrire c’est faire, les meta-langage transforment la transcription de la parole en une nouvelle connaissance. On peut agir sur la parole, sur le texte, le tordre, le presser, le décoder.

L’étude du texte en littérature peut se concentrer sur une oeuvre, sur un texte ou un fragment, parfois on compare deux ou trois auteurs. Le texte vernaculaire est produit par des foules . On peut lire les foules. le texte a un auteurs, les méthodes étudient le textes de nombreux auteurs, et des mots qu’ils ont en comment.

Les langages tels que la linguistique classique étudie sont verbaux, d’autres sont iconiques, architecturaux, graphiques, chorégraphiques, musicaux. Elles se rencontre dans le flux d’une parole qui associe le texte à l’image dans des rapports d’illustrations et de commentaires,jouant du contrepoint à travers les médias. Par le texte, le sociologue, l’économiste ou le gestionnaire veulent co-prendre, ou comprendre, la génèse et la détermination des choix. Etudions donc le texte.

4 La notion de corpus

```
#les librairies du chapître
library(tidyverse)
library(readr)
library(quanteda)
library(flextable)

theme_set(theme_minimal())

set_flextable_defaults(
  font.size = 10, theme_fun = theme_vanilla,
  padding = 6,
  background.color = "#EFEFEF")
```

Objectifs du chapitre :

- *Introduire la notion de corpus, la développer et conclure sur des premiers éléments de code qui vont lui donner une forme matérielle.*

La notion de corpus est très intuitive, c'est un ensemble de document qui peuvent prendre des formes matérielles très différentes : des cahiers de doléances comme au moment des gilets jaunes, un ensemble de pdf, un tableur où une des colonnes contient une variable de type chaîne de caractères. De manière plus classique c'est un recueil réunissant ou se proposant de réunir, en vue de leur étude scientifique, la totalité des documents disponibles d'un genre donné.

Dans la perspective d'une analyse automatisée ce chapitre, on ne tiendra pas compte des aspects matériel du corpus, qui peuvent être important par ailleurs comme le montre Marie-Anne Chabin.

Dans le cadre d'une analyse automatisée, le document se réduit à une chaîne de caractères. Sa collection peut se présenter dans une série de fichiers distincts ou être déjà structurée sous la forme d'un tableur, ou mieux d'une base de données relationnelles.

4.1 Les éléments du corpus

Ce qui fait un corpus est composé de trois éléments : une collection, des textes qui se manifestent comme une suite de caractère répondant à des règles plus ou moins connue, et à des informations associées, les méta-donnée. Examinons chacun de ces éléments.

4.1.1 Le document

L'ensemble des documents constitue la collection , elle peut être systématique et exhaustives, par exemple s'il s'agit de toute la correspondance d'un écrivain. Elle peut aussi se constituer comme un échantillon et répondre au règles de la théorie de l'échantillonnage.

4.1.1.1 Le document comme chaine de caractère

Une unité de texte se définit comme une chaîne de caractères qui constitue un document. Celui ci peut être de forme quelconque : un livre, un article, une note, une transcription, il reste une séquence plus ou moins longue de caractères qui répondent à des règles de composition dont on ignore la nature a priori. Dans le cas de texte courts, cette chaîne de caractères est simple.

Dans celui d'un article, elle peut être plus complexe et comporter des combinaison de caractères qui signalent un effet de composition. par exemple la chaîne `\n` définit un saut de page, et si le document est codé en xml, certaines séquences identifient des balises. Par exemple la chaîne `<h1>` indique que les caractères qui vont suivre définissent le contenu d'un titre de niveau 1 et que cette définition s'achève avec la balise `<|>`. L'exemple adresse une convention qui est celle des langage xml qui séparent le contenu du contenant, c'est à dire des actions qui seront opérées sur le contenu.

L'analyse de ces structure est un préalable indispensable pour les lire correctement. Et il sera souvent néçéssaire de nettoyer le texte avant son exploitation.

4.1.1.2 la structuration

Différents types de corpus Le degré de structuration : séquentielle, plan spatiale (ex = le curriculum viate, la fiche de brevet,…)

4.1.1.3 Metadonnées

Un document est souvent associé à des méta-données qui permettent de le situer ce qui permet des comparaisons dans le temps l'espace ou les locuteurs. Un bon corpus est celui qui associe aux documents, l'ensemble le plus pertinent et le plus large de méta-données.

- Un ou des auteurs. A chaque document peut être attaché un ou plusieurs auteurs. Et c'est auteurs eux-même peuvent être caractérisés.
- Chaque document peut être associé à une date de production ou de publication.
- Chaque document peut être associé à une origine géographique, générique comme le pays d'origine ou géolocalisé dans le cas d'un message émis dans un réseau social.
- Il peut être documenté par des éléments de contexte : les unités précédentes, et subséquentes - c'est le cas des chats et des forums qui peuvent être identifié par les fils dans lesquels ils s'inscrivent et la position qu'ils y occupent.
- une ou des autorités : par exemple le média où il a été publié ce qui pose la question de la source et celle de sa crédibilité

La liste n'est pas limitative et peut comprendre des éléments de diplomatique, cette discipline, ancienne, qui vise à établir une compréhension critique des documents écrits, pour en particulier établir leur authenticité .voir

4.1.2 Une collection

Il y a une très forte diversité de corpus. Celle-ci est cependant peut se décrire au travers d'un certains nombre de critères.

4.1.2.1 L'échelle et l'étendue

Il peut y avoir de très petits corpus, par exemples la transcription de quelques dizaines d'entretiens. d'autres qui rassemblent des millions de textes courts. L'échelles des corpus va de quelques unités à plusieurs millions. Les milliards sont rares. Concrètement il y a des corpus de quelques dizaines ou centaines d'éléments, d'autres qui se comptent en dizaines de milliers, d'autres en centaines et millions d'unités. Méthodes et capacité de calculs ne sont pas forcément les mêmes.

Courts à l'image des tweets, ou de résumés d'articles, ou longs s'il s'agit d'articles de recherche complet, ou très long, le cas des livres. les textes court sont ce que les méthodes avancée traitent bien . Les textes longs pose la question de co-références. On peut cependant les décomposer, le niveau de phrase, celui du paragraphe, le chapitre, le livre.

A ce stade, 4 types de corpus

	textes Peu	Nombreux
court	Cas	Social data
Long	Comparative	Deep data

4.1.2.2 les auteurs

Les corpus peuvent aussi se distinguer par la diversité de leurs auteurs

- la monographie quand le corpus compile tous les documents d'une même source
- la polygraphie quand les auteurs sont presque aussi nombreux que le nombre de textes. Quand les auteurs sont multiples, il peuvent constituer le texte de deux manières
- par addition : c'est l'exemple du message publicitaire qui superpose une scène, des incrustations, de la musique.
- par interaction : c'est le mode tu texte de théâtre ou la structure du texte distribuent entre les personnage des fragments de paroles qui interagissent.

encore une autre typologie

Auteurs	Addition	Interaction
Unique	monographie	
Double	contrepoint	Dialogue
Multiple	Polygraphie	Théatre

4.1.2.3 les niveaux de langues

Ils varient aussi selon le niveaux de langues

- texte savant, texte standard, texte vernaculaire
-

4.1.3 des corpus bien préparés

Dans ce chapitre nous avons appris comment faire la cueillette dans les sources de textes et constituer matériellement un corpus. Il reste à traiter la question de la représentativité. La collecte doit rester raisonnée.

Un corpus se construit. D'abord en fonction d'un objectif de recherche? Cherche-t-on à décrire, à comparer, à explique ?

Ensuite en fonction d'une perspective d'écoute. par exemple prendre le corpus de ceux qui produisent de la publicité, ou le corpus de ce à quoi les consommateurs sont exposés ?

4.1.3.1 Production et réception

Unités de production et de réception, Un texte est produit et puis, peut-être, lu. Analyser le texte peut se faire dans deux perspectives, celle de la production et celle de la réception. Les corpus doivent être construits en fonction de ce critère.

4.1.3.2 Les conditions de production des documents

Examiner la question de l'engagement dans ce cadre est essentiel, certains acteurs sur un sujet donné sont amenés à produire plus que les autres, et participent donc de surcroît à une sur représentation statistique. La question du biais de sélection

Prenons le cas des sites d'avis de consommateurs.

la question des textes absents

4.2 Le corpus comme objet de traitement

Sur le plan pratique nos discussions dépendent largement des modalités opérationnelles que les langages de traitement de données nous proposent.

Dans l'univers r le package `tm` est fondateur, mais d'autres solutions sont offertes. On signalera d'abord `stringr` qui fait partie de la suite `tidyverse`.

Parmi elles on fait le choix des méthode de la suite `Quanteda`.

4.2.1 Un exemple

Il s'agit du corpus "TrumpArchive".

```
df <- read_csv("data/TrumpTwitterArchive01-08-2021.csv")%>%
  select(id, text, favorites, retweets, date)
head(df)
```

```
# A tibble: 6 x 5
       id text                                favorites retweets date
    <dbl> <chr>                                   <dbl>    <dbl> <dttm>
1 9.85e16 Republicans and Democrats have~            49      255 2011-08-02 18:07:48
2 1.23e18 I was thrilled to be back in t~         73748    17404 2020-03-03 01:34:50
3 1.22e18 RT @CBS_Herridge: READ: Letter~             0     7396 2020-01-17 03:22:47
4 1.30e18 The Unsolicited Mail In Ballot~         80527    23502 2020-09-12 20:10:58
5 1.22e18 RT @MZHemingway: Very friendly~             0     9081 2020-01-17 13:13:59
6 1.22e18 RT @WhiteHouse: President @rea~             0    25048 2020-01-17 00:11:56
```

C'est le champs texte qui définit notre corpus, les autres variables / colonnes du tableau représentent les méta données.

```
corp <- corpus(df$text, docvars = df)
summary(corp[1:5,])
```

```
Corpus consisting of 5 documents, showing 5 documents:

  Text Types Tokens Sentences           id
 text1    10     10         1 9.845497e+16
 text2    42     50         3 1.234653e+18
 text3    23     24         1 1.218011e+18
 text4    49     61         3 1.304875e+18
 text5    25     25         2 1.218160e+18


            I was thrilled to be back in the Great city of Charlotte, North Carolina with th

 The Unsolicited Mail In Ballot Scam is a major threat to our Democracy, &amp;amp; the Democrats

 favorites retweets                date
        49      255 2011-08-02 18:07:48
     73748    17404 2020-03-03 01:34:50
         0     7396 2020-01-17 03:22:47
     80527    23502 2020-09-12 20:10:58
         0     9081 2020-01-17 13:13:59
```

Une fois le corpus constitué , on peut en sélectionner des sous-ensemble. Dans l'exemple suivant on isole les tweets qui ont été retwittés plus de 20 000 fois.

Il va être temps de jouer avec les données !

```r
# on peut choisir un sous-corpus
corp_recent <- corpus_subset(corp, retweets >20000)
ndoc(corp_recent)
```

```
[1] 7277
```

```r
head(corp_recent, 10)
```

```
Corpus consisting of 10 documents and 5 docvars.
text4 :
"The Unsolicited Mail In Ballot Scam is a major threat to our..."

text6 :
"RT @WhiteHouse: President @realDonaldTrump announced histori..."

text7 :
"Getting a little exercise this morning! https://t.co/fyAAcbh..."

text9 :
"https://t.co/VlEu8yyovv"

text11 :
"https://t.co/TQCQiDrVOB"

text16 :
"As per your request, Joe... https://t.co/78mzcfLEsF https://..."

[ reached max_ndoc ... 4 more documents ]
```

4.3 Conclusion

Techniquement c'est aussi simple que cela. D'un point de vue méthodologique la question de la constitution d'un corpus reste complexe, elle demande de l'intelligence dans l'exploitation des sources, une maîtrise technique des interfaces (les APIs).

Dans le chapitre suivant, nous explorons des aspects plus techniques : comment saisir la forme matérielles des documents, comment interroger les bases de données, et exploiter des Apis.

5 Corpus : techniques avancées

```
#les librairies du chapître
library(tidyverse)
library(readr)
library(pdftools)
library(flextable)


theme_set(theme_minimal())

set_flextable_defaults(
  font.size = 10, theme_fun = theme_vanilla,
  padding = 6,
  background.color = "#EFEFEF")
```

Objectifs du chapitre :

- – Explorer différentes techniques de collectes de données : exploitation de bases textuelles, méthodes de scrapping, APIs, extraction de document pdf, extraction de textes dans des images, et une perspective oral avec les techniques de speech2tex.

La constitution d'un corpus est la première étape d'un projet NLP. Il se définit d'abord par la constitution d'une collection de textes dont la provenance est la nature peut être diverse.

Dans ce chapitre on va examiner plusieurs techniques de collecte, puis conclure avec quelques réflexions sur la question de la constitution de l'échantillon.

- L'exploitation de bases textuelles
- Les méthodes de scrapping
- Le recours aux APIs
- La collection de documents pas uniquement textuelle
- Les sources orales

5.1 La gestion des documents numériques

Dans certains cas le matériau se présentera sous forme de documents numériques tels qu'un pdf, ou même de simples images.

voir aussi

https://cran.r-project.org/web/packages/fulltext/fulltext.pdf

5.1.1 Extraire du texte des pdf

Le package pdftools est parfaitement adapté à la tâche. Des fonctions simples extraient différents éléments du pdf :

- Les information relatives au document pdf lui-même
- La liste des polices employées
- Les attachements
- La table des matières (si elle a été encodée)
- Les chaînes de caractères constituant le texte dans un ordre de droite à gauche et ligne à ligne, reconnaissant cependant les retours chariot, et autres sauts de lignes séparant les paragraphes

Chaque page est contenue dans une ligne.

```
info <- pdf_info("./data/2021neoliberalismegouverner_Meunier_Esprit.pdf")
info
```

```
$version
[1] "1.4"

$pages
[1] 12

$encrypted
[1] FALSE

$linearized
[1] TRUE

$keys
$keys$Author
```

```
[1] ""

$keys$Creator
[1] ""

$keys$Keywords
[1] ""

$keys$Producer
[1] "TCPDF 6.2.12 (http://www.tcpdf.org)"

$keys$Subject
[1] ""

$keys$Title
[1] "Le néolibéralisme et l\u0090art de gouverner"

$keys$Trapped
[1] ""


$created
[1] "2021-05-04 17:33:26 CEST"

$modified
[1] "2021-07-02 14:10:59 CEST"

$metadata
[1] "<?xpacket begin=\"ï»¿\" id=\"W5M0MpCehiHzreSzNTczkc9d\"?>\n<x:xmpmeta xmlns:x=\"adobe:ns
```

```
fonts <- pdf_fonts("./data/2021neoliberalismegouverner_Meunier_Esprit.pdf")

files <- pdf_attachments("./data/2021neoliberalismegouverner_Meunier_Esprit.pdf")
```

```r
toc <- pdf_toc("./data/2021neoliberalismegouverner_Meunier_Esprit.pdf") #il n'y a pas de tabl

text <- pdf_text("./data/2021neoliberalismegouverner_Meunier_Esprit.pdf")
cat(text[[1]]) # pour afficher le texte de la page 1
```

```
Le néolibéralisme
et l'art de gouverner
À propos de Naissance de la biopolitique
de Michel Foucault
François Meunier



O          n dit parfois du métier de l'historien qu'il consiste avant tout
           à découper en périodes, à indiquer les ruptures dans le temps
           historique, à montrer les changements d'environnement et
de paradigme. C'est à ce travail que se consacre Michel Foucault dans
son célèbre cours de 1978-1979 au Collège de France connu sous le
nom de Naissance de la biopolitique 1. Il devait porter initialement sur la
« biopolitique », un mot chatoyant recouvrant les pratiques politiques
contemporaines autour du vivant (santé, démographie, sexualité, etc.).
Mais Foucault voulait montrer d'abord à quel point la venue du libé-
ralisme avait modifié en profondeur les pratiques gouvernementales.
Première rupture, celle advenue à la fin du xviiie siècle avec le libéralisme
économique classique, selon lequel le marché devient l'instance clé dans
l'art de gouverner, donnant à l'action publique un lieu de légitimation en
même temps que des limites. Seconde rupture, celle qui sépare libéralisme
et néolibéralisme, que Foucault situe dans l'après-guerre en Allemagne,
avec ce qu'on appelle « l'ordolibéralisme ».


Équivocité du néolibéralisme
Reprenant, quelque quarante ans après, le fil de ce cours, nous remettons
ici en cause le découpage historique. D'abord, il nous semble que ce

1 - Michel Foucault, Naissance de la biopolitique. Cours au Collège de France (1978-1979), Pa
EHESS/Seuil/Gallimard, 2004.
```

Il va falloir traiter ce texte en analysant précisément sa composition. Pour ce faire, il s'agîra de définir une séquence d'opérations logiques qui permette un premier nettoyage du texte. Dans l'exemple nous allons de plus essayer de conserver la structure des paragraphes du texte.

- Suprimer haut et bas de pages
- Supprimer les sauts de ligne
- Identifier les sauts de paragraphes
- Enlever les notes de bas de page
- Corriger l'hyphénation ()
- regrouper les document en un seul bloc de texte
- le splitter en autant de paragraphes.

On va utiliser des fonctions de traitement de chaines de caractère avec Stringret le recours à l'art (ici simple) des regex auxquels on consacre un développement dans le chapitres X.

```
tex<- as.data.frame(text)
tex[1,]
```

```
[1] "Le néolibéralisme\net l'art de gouverner\nÀ propos de Naissance de la biopolitique\nde M
```

```
t_reg<-str_replace(tex$text,"[\\s+].*Meunier[\n]+", " ") # entete droite
## on selectionne tout bloc de texte qui commence par un nombre indéterminée de blanc qui s'
t_reg<-str_replace(t_reg,"[\\s+].*gouverner[\n]+", " ") # entete gauche
t_reg<-str_replace_all(t_reg,"[\\s+].*2021[\n]", " ") # bas de page  gauche
t_reg<-str_replace_all(t_reg,"ESPRIT.*[\n]", " ") # bas de page droit

#on marque les paragraphes avec la chaine XXX pour les splitter dans un second temps


t_reg<-str_replace_all(t_reg,"\n\n\n", "XXX")

# On supprime les saut de ligne en les remplaçant par un espace

t_reg<-str_replace_all(t_reg,"[\n]", " ")

#on enlève les notes de bas de page
t_reg<-str_replace_all(t_reg,"\\d\\s[\\-].*XXX", "XXX")

#on regroupe les pages
```