

Exploration des savoirs

Analyser les données

Émilien Schultz - Nicolas Benvegna

Objectifs de la séance

- Passer des données à des résultats
- Choisir sa démarche
- Intégration dans la réflexion

Concrètement, dans 2 semaines : les *avancements intermédiaires*

- délimitation d'un corpus
- démarche pour l'analyse

Données = piliers de l'enquête

Vous en avez fait quoi pour le moment ?

Rappel des différents usages

- Importance (et difficulté) de cibler leur pertinence
- Intervenir à différents moments/modalités
 - Explorer (baliser un domaine)
 - Prouver (nécessite une question bien définie)
 - Renforcer (fiabiliser et représenter)

Analyser des données demande du travail : soyons humbles

Renforcer : l'importance de la triangulation

La triangulation : examiner sous plusieurs angles conceptuels, sources de données ou méthodes

Objectif : Renforcer la validité

- Observations directes + témoignages (entretiens)
- Entretiens + mesures
- Médias traditionnels + médias sociaux
- ...

par exemple, construire une figure

Aller vers l'analyse

Différentes *philosophies* avec leurs avantages/inconvénients

- “**pipelines automatisés dédiés**” : logiciels spécialisés intégrant une philosophie
- “**logiciels métiers**” : boîtes à outils spécialisées
- “**à la main**” : décider de chaque choix proche des données
- “**programmation**” : formaliser les étapes (méta-boîte à outils)

Penser le processus d'ensemble

De nombreuses étapes

- D'abord, un ensemble de données **brutes**
- Ensuite, des **transformations** de données
 - Données/analyses intermédiaires
- Progressivement, **stabilisation** d'analyses
 - Laisser de côté des pistes
- Résultats finaux

Éviter les impasses : la reproductibilité

Etre capable de refaire le plus simplement possible toutes les étapes (souvent très itératives et tortueuses) entre l'idée et les résultats



Notion centrale de la science ouverte

- A minima : **documenter**
 - Carnet de recherche
- A maxima : **formaliser les étapes**
 - Pouvoir tout rejouer
- Entre :
 - Conserver les étapes intermédiaires
 - Ne pas supprimer les étapes précédentes

Aujourd'hui deux focus :

- Pipeline automatisé de données scientométriques
 - VosViewer
- Analyse des données de presse “à la main”
 - Grille de codage + Tableur (Spreadsheet Google)

Retour des vélos

Corpus sur le traitement de la question vélo:

- Scopus : qui sont les chercheurs qui travaillent sur ces questions ?
- Europresse : comment est couverte l'accidentologie ?

Analyse scientométrique

Données scientométriques

- Des données complexes
 - Mots-clés, auteurs, affiliations, contenu, ...
- Des données structurées
 - Scopus permet de générer un fichier propre
- Un domaine à part entière : la *scientométrie*

Conséquence : des outils *calibrés* (métriques, manières de faire, etc.)

Quelles questions poser à partir d'un corpus scientométrique?

Au sens large : production des connaissances et les dynamiques d'expertise.

- Quelles sont les thématiques abordées ?
- Qui sont les auteurs ? Comment sont-ils reliés entre eux ? Est-ce qu'ils forment une communauté ?
- Comment les sujets sont connectés ? Comment ils ont évolués dans le temps ?
- Quels sont les articles les plus importants ? Les plus centraux ?
- ...

Constituer le corpus avec Scopus

- Bien calibrer la question
- Taille raisonnable
- Vérifier en lisant quelques articles
- Faire évoluer si besoin les règles de filtrage
- Garder une trace de la requête
- Choisir un format d'export pertinent (CSV est bien)

Comment la science parle des vélos

- `(bike* OR cyclist*) > 20000`
- `(bike* OR cyclist*) AND france = 352`



Welcome to a more intuitive and efficient search experience. [See what is new](#)

Advanced query ☐

Search within
Article title, Abstract, Keywords

Search documents *
(bike* OR cyclist*) AND france

Save search

Set search alert

+ Add search field

Reset

Search

Beta

Documents

Preprints

Patents

Secondary documents

Research data ↗

330 documents found

Analyze results

Refine search

Search within results

Filters Clear all

Year Clear

Range Individual



☐ All ☐ Export ☐ Download ☐ Citation overview ☐ More

Show all abstracts

Sort by Date (newest)

Grid List

	Document title	Authors	Source	Year	Citations
<input type="checkbox"/> 1	Article • Open access Exposure to particulate matter when commuting in the urban area of Grenoble, France	Aix, M.-L., Claitte, M., Bicout, D.J.	Atmospheric Environment , 339, 120887	2024	0
	Show abstract Try for Full-text Related documents				
<input type="checkbox"/> 2	Article • Open access The Night-Time Sleep and Autonomic Activity of Male and Female Professional Road Cyclists Competing in the Tour de France and Tour de France Femmes	Sargent, C., Jasinski, S., Capodilupo, E.R., Miller, D.L., Beach, G.D.	Sports Medicine - Open , 10(1), 39	2024	0


Rappel : l'analyse commence avec les données

- Sélectionner/filtrer le corpus
- Transformer les données
 - Recoder
 - Compléter
 - Fusionner
 - Supprimer

Par exemple, ne pas hésiter à constituer plusieurs corpus et faire des comparaisons

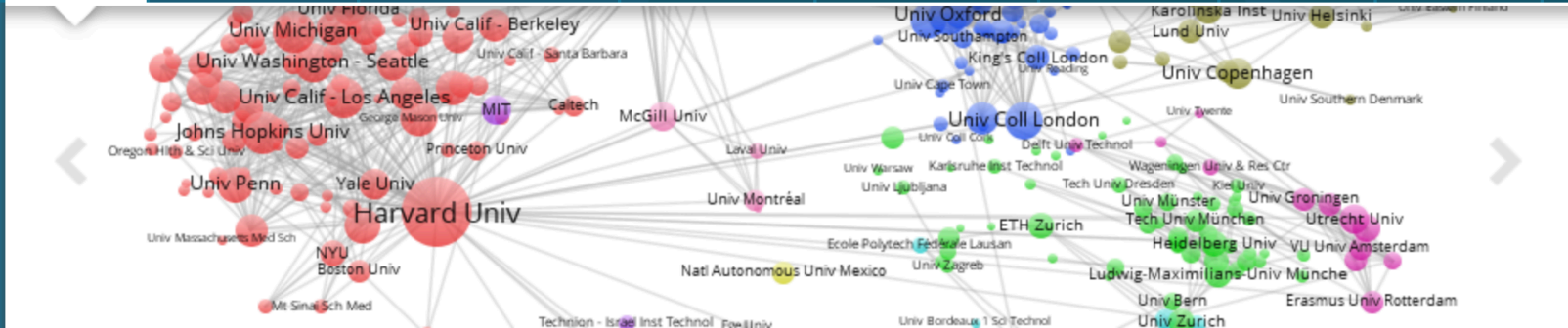
VOSviewer : réseaux & statistiques

Un outil dédié issu de la recherche dont la philosophie est de créer des cartes relationnelles à partir de données scientométriques.

 **VOSviewer**
Visualizing scientific landscapes

Leiden University | CWTs | CWTs B.V. | Other CWTs sites ▾

Home | Features ▾ | Getting Started | Download | Publications | Products | Course | Contact



Welcome to VOSviewer

VOSviewer is a software tool for constructing and visualizing bibliometric networks. These networks may for instance include journals, researchers, or individual publications, and they can be constructed based on citation, bibliographic coupling, co-citation, or co-authorship relations. VOSviewer also offers text mining functionality that can be used to construct and visualize co-occurrence networks of important terms extracted from a body of scientific literature.

De nombreux tutoriaux


Deux mots sur l'analyse de réseaux

- des **entités** qui sont **connectées**
 - des **noeuds** (personnes, mots, etc.)
 - des **liens/rerelations** (proche, contenu, etc.)
- en mettant toutes ces relations ensemble : un **réseau**
- permet de poser la question:
 - quelle forme générale a ce réseau ?
 - comment sont liées les entités ?
 - est-ce que les entités sont proches ou éloignées ?

L'analyse de réseaux en sciences sociales de L. Beauguitte

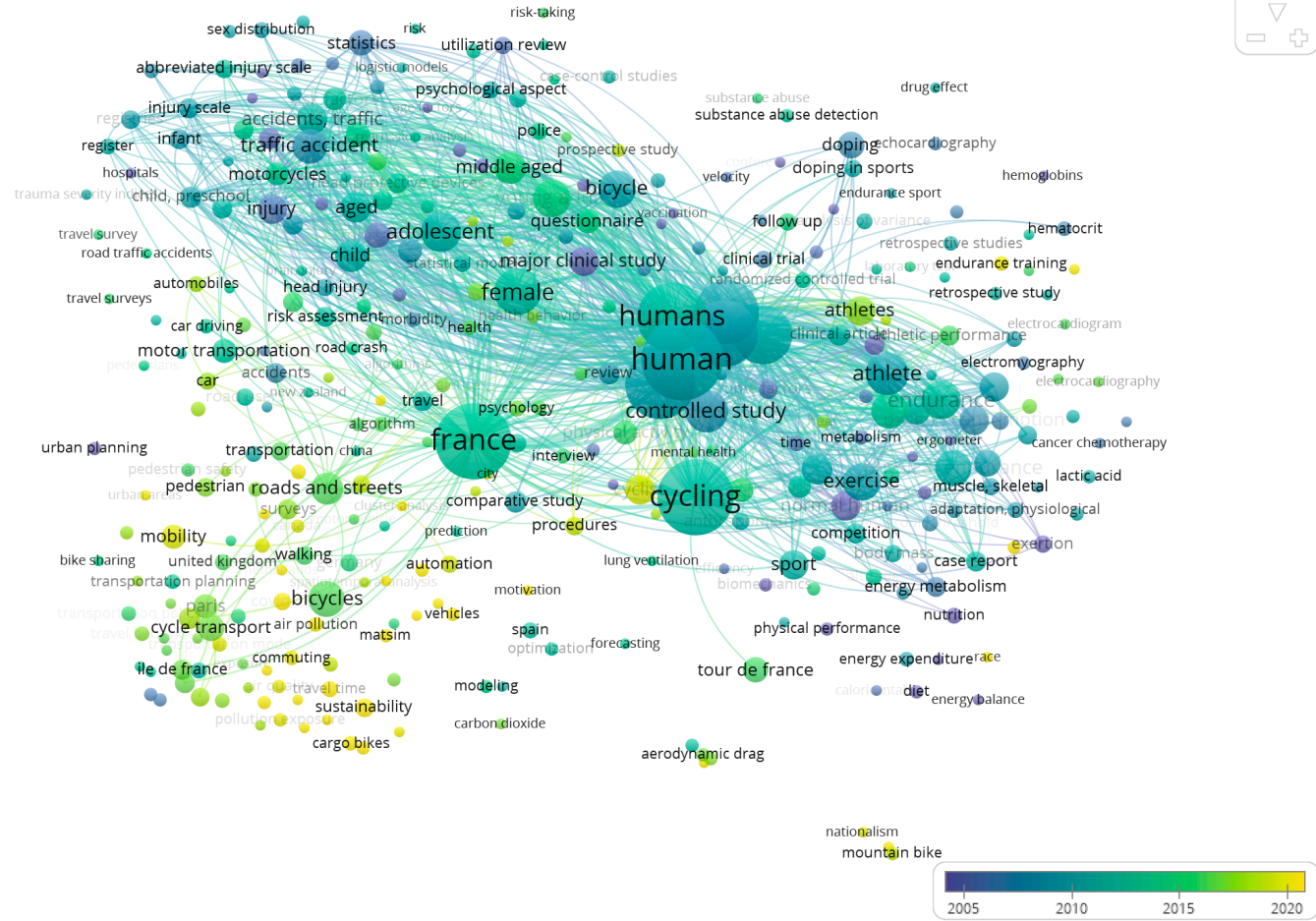
Démo : analyser un corpus

 Est-ce qu'il existe une communauté de chercheurs spécialisée sur le vélo en France ?

 Quelles sont les références les plus mentionnées ?

Lançons VOSVIEWER !

Avoir une vision chronologique



Remarques lors de l'analyse

- Importance de comprendre les métriques :
 - Que représente la couleur ?
 - Que représente la taille ?
 - Qu'est-ce qu'est un lien ?
- Possibilité d'aller plus loin
 - Extraire les éléments non pertinents
 - [Thesaurus pour réunir](#)

Les limites

- Pas mal de possibilités
- Mais sur des données bien calibrées
- Avec une philosophie intégrée spécifique
- Et des marges de manoeuvre limitées

Comment faire quand on veut construire son propre cadre d'analyse ?

Analyse de presse

Presse : des données moins structurées

- Des méta-informations (journal, etc.)
- Du texte

Des questions souvent liées au contenu

- Quels sujets sont traités ?
- Comment ils évoluent ?
- Qui est mentionné ?

Mais pas seulement (Combien d'articles sur un sujet sur une période, etc.)

Un cadre méthodologique à développer

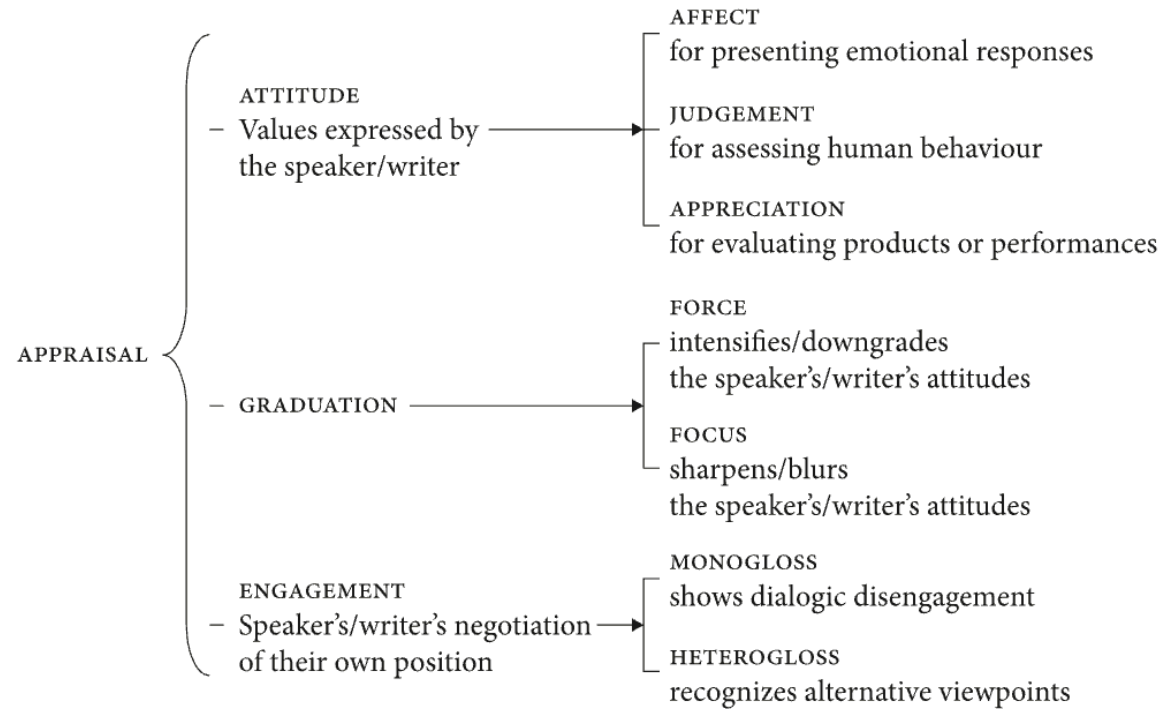
Comment passer du texte non-structuré à des données structurées interprétables ?

- Graphique
- Tableau croisé
- ...

Lipovsky, Caroline. “Cycling the city: Representation in the French media.”
Language, Context and Text 2.2 (2020): 334-367.

D'abord, il faut des concepts

Articulés à une problématique



Ensuite, une opérationnalisation

Le *codebook* : adosser chaque catégorie à une définition

- Explicite
- Opérationnelle
- Simple

Toujours plus facile à dire qu'à faire...

Différentes stratégies pour l'analyse de texte

Et ensuite une implémentation

- Dictionnaire de mots à détecter automatiquement
- Découper le texte en élément et les catégoriser
- Identifier des éléments dans le texte
- LLM ...

(Les différents éléments peuvent être combinés)

Un flux de travail

1. Lire pour connaître ses données
2. Développer une grille de codage :
 - **Cible** : article entier ? paragraphe ? phrase contenant certains mots ?
 - **Codage** : catégories et des règles pour les appliquer
3. Produire un tableau (excel ou autre) adapté
 - Filtrer son corpus
4. Coder une dizaine d'éléments pour vérifier que la grille fonctionne
 - Faire évoluer la grille de codage si besoin
5. Coder l'ensemble du corpus
6. Faire des statistiques (comptage, ou plus si affinité)

Remarques sur le codage

- Il y a toujours des cas ambigus, c'est normal
- Ajouter une colonne pour garder des commentaires
- Toujours garder une possibilité de revenir à l'article initial
 - Surtout si l'unité de travail est plus petite
- Possibilité d'avoir des codages complexes :
 - Plusieurs dimensions
 - Dimension qui code l'intensité/la certitude

Mettons nous à l'oeuvre

Représentations des cyclistes dans la presse

Est-ce que ça a changé ces 4 dernières années ? Comment parle-t-on des cyclistes dans la presse ? Est-ce qu'on parle davantage des violences que subissent les cyclistes ? Est-ce qu'on parle des inégalités liées au vélo ?

Choix d'une question : comment couvre-t-on les accidents ?

D'abord le corpus

- Requête Europresse (portail bibliothèque SciencesPo)
 - accident & vélo* | cycliste*
 - quelle portée ? volume ?
- Récupérer les données
 - Fichier HTML

EUROPRESSE

UNE SOLUTION DE CISION

RECHERCHER

DOSSIERS

PUBLICATIONS PDF

📌 (0)

English

?

Recherche simple

Recherche avancée

Recherche express

Recherche de biographies

TEXT= cycliste* & accident*

Depuis 30 jours

France (FR)

🔍

↶

✕

Presse

Télévision et radio

Médias sociaux

Études et rapports

Répertoires et références

50 sur 207

📄

📁

🖨

✉

Pertinence

Le Monde

Altis, « vélotaffeur » d'utilité publique

2024-11-25 • 1510 mots PDF

Pascale Krémer - Certains des 72 400 abonnés de sa chaîne YouTube (« Altis Play ») le saluent désormais dans les rues de Paris. Ils le reconnaissent à son allure d'échalas, à ses gants noirs ...

Aussi paru dans

Libération

95 %

2024-10-31 • 181 mots PDF

C'est la part des usagers de la route qui redoutent les comportements à risque des autres, selon une étude Ipsos publiée mercredi pour la fondation Vinci Autoroutes. En France, le ...

Les Echos

Midipile Mobilitv invente l'hybride mi-vélo mi-voiture

Tableau de bord

MÉDIAS

234 Presse

100%

Presse

100% - 234

Télévision

0% - 0 doc

Médias so

0% - 0 doc

Études et

0% - 0 doc

Répertoire

0% - 0 doc

ÉVOLUTION

Pic médiatique : 14 documents le 13 no

15

Max éléments : 500 (faire des tranches par période si besoin)

Consolider le corpus

Passer d'un fichier HTML à un tableau manipulable :

<https://dstool.onrender.com/>



Outils disponibles

- Transformer le corpus
- Produire un graphique
- Extraire des phrases autour de mots clés

Puis à la main

- Enlever les entrées non pertinentes
- Enlever les doublons
- Enlever les colonnes inutiles

un fichier clean.xlsx

Décrire le corpus

- Statistiques de colonnes
 - Colonne journal : plein de soucis
 - Dupliquer la colonne
 - Corriger à la main (rechercher/remplacer)
- Et souvent des données sales : il faut transformer les données
 - A la main
 - Long
 - Solution des macros
 - Complexe
 - Des outils adapté : programmation/[OpenRefine](#)

Ajouter une dimension/variable

Principe : transformer le texte en une nouvelle variable (présence d'un acteur, tonalité, nombre de mots, nombre d'occurrence d'un terme, etc.)

- Permettre de compter les occurrences
- De croiser avec d'autres éléments
 - Journal
 - Date
 - Autre variable

Comment faire? Cas d'un codage du traitement des accidents

Définir la grille de codage

- Définir les variables le plus clairement possible
- Avoir des exemples
- Identifier l'échelle

Variables sur les accidents vélo à Paris

- Variable : accident & vélo
 - Cycliste victime
 - Ambigu
 - Cycliste responsable
 - Ne parle pas d'accident
- A l'échelle de la phrase

Du texte à la phrase

Je veux m'intéresser aux phrases spécifiques

Article entier > phrases spécifiques

- présence de mots/combinaisons de mots
- phrase, groupe de phrases, etc.



Extraire des phrases

Outil sur dstool à partir de **regex**

C'est quoi une expression régulière

Regex = expression régulière -> un pattern de texte

Exemples :

- vélo : présence de la chaîne vélo
- vélo|Vélo : l'un ou l'autre
- \b\w{5}\b : mot de 5 lettres
- \b(chat|chien)\b : mots “chat” ou “chien” dans un texte
- \w+@\w+\.\w+ : adresse mail
- \d{2}/\d{2}/\d{4} : date

Présent dans de nombreux logiciels/langages de programmation

Tableau des extraits parlant d'accidents

- Toutes les phrases mentionnant un terme
 - cycliste (une regex simple)
- Garder une phrase avant et une phrase après
 - Avoir du contexte
- Lire et coder chaque élément

Coder (collaborativement)

En pratique :

- Un document partagé Spreadsheet
- Se mettre d'accord sur les règles
- Faire un test sur un petit nombre d'éléments avant de se lancer
 - Ajuster
- Si un doute, prendre des notes dans une colonne dédiée

Faisons un peu de codage

Données codées

	A	B	C	D	E	F	G
		Date_mod	Titre_mod	Contenu_mod	Journal_mod	phrases_3	codage
1							
2	0	#####	Une cycliste percutée par une voiture	Fait divers	Mardi, à 16 h, une voiture a perc	VICTIME	
3	1	#####	Une cycliste percutée par une voiture	Fait divers	Mardi, à 16 h, une voiture a perc	VICTIME	
4	2	#####	Sécurité routièreLes violences routi	Stop violences m	Le Monde	Stop violences motorisées ».C'est le nom de	VICTIME
5	3	#####	Sécurité routièreLes violences routi	Stop violences m	Le Monde	Ce drame a fait émerger la thématique des	HS
6	4	#####	Sécurité routièreLes violences routi	Stop violences m	Le Monde	Les pouvoirs publics ont eux aussi commen	HS
7	5	#####	Sécurité routièreLes violences routi	Stop violences m	Le Monde	Il leur a annoncé une « mission contre les vi	HS
8	6	#####	Sécurité routièreLes violences routi	Stop violences m	Le Monde	Mais au-delà des frustrations des uns et des	VICTIME
9	7	#####	Sécurité routièreLes violences routi	Stop violences m	Le Monde	» Parmi les piétons, on dénombre 438 mort	HS
10	8	#####	Sécurité routièreLes violences routi	Stop violences m	Le Monde	Dès 2020, ses unités ont décliné leur propre	NEUTRE
11	9	#####	Sécurité routièreLes violences routi	Stop violences m	Le Monde	Certaines mesures de sécurité routière réce	NEUTRE
12	10	#####	Un cycliste de 87 ans grièvement ble	Faits divers	Un c	Ils accompagnent le brancard sur les quelq	NEUTRE
13	11	#####	Un cycliste de 87 ans grièvement ble	Faits divers	Un c	Les policiers leur demandent de reculer, un	NEUTRE
14	12	#####	Un cycliste de 87 ans grièvement ble	Faits divers	Un c	« Il y a eu un bruit très impressionnant,dit-i	VICTIME
15	13	#####	« Ce jour-là, renversée par un camio	« Un accident de	Ouest-Fran	Comme chaque matin, elle parcourt 12 km	NEUTRE
16	14	#####	La TalaudièreUn conducteur percut	Une heure après	Le Progrès	Et si l'enquête n'a pas encore établi clairem	RESPONSABLE
17	15	#####	La TalaudièreUn conducteur percut	Une heure après	Le Progrès	Âgé d'une cinquantaine d'années, il a été tr	VICTIME

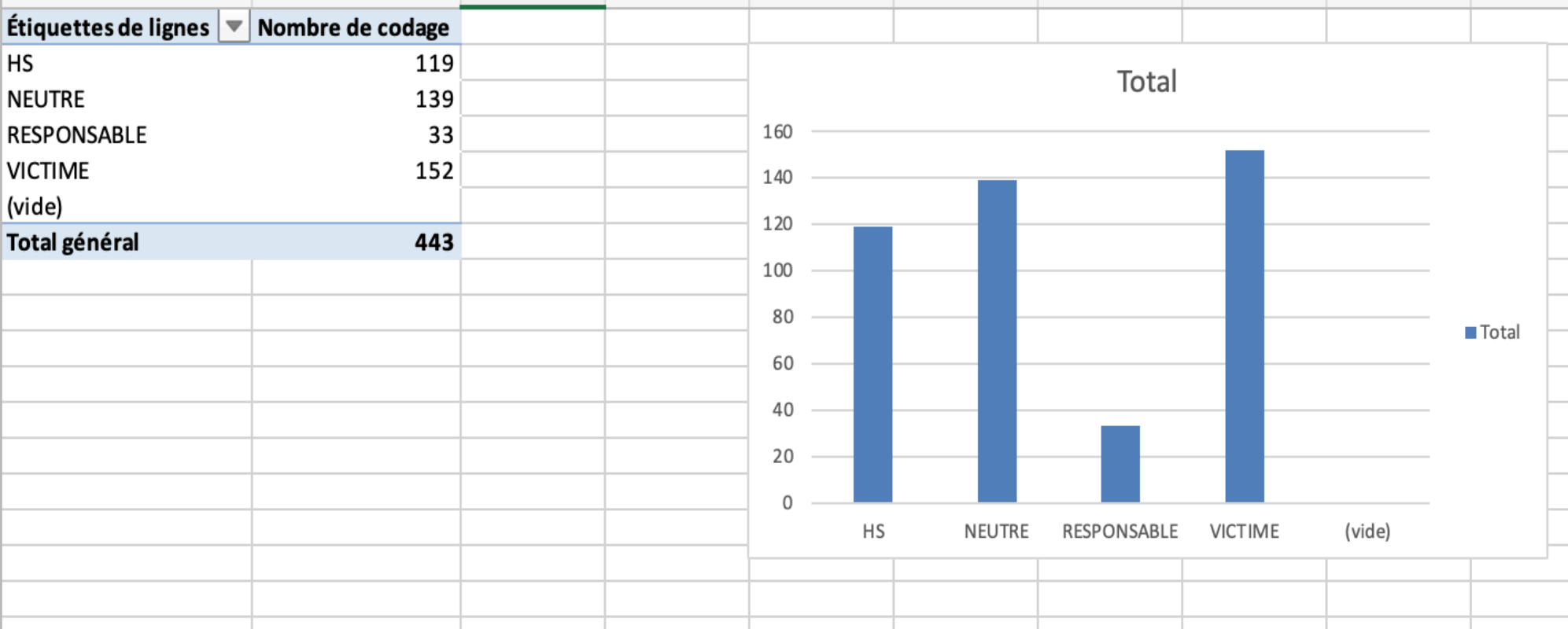
Analyser

Beaucoup d'options disponibles avec les tableurs :

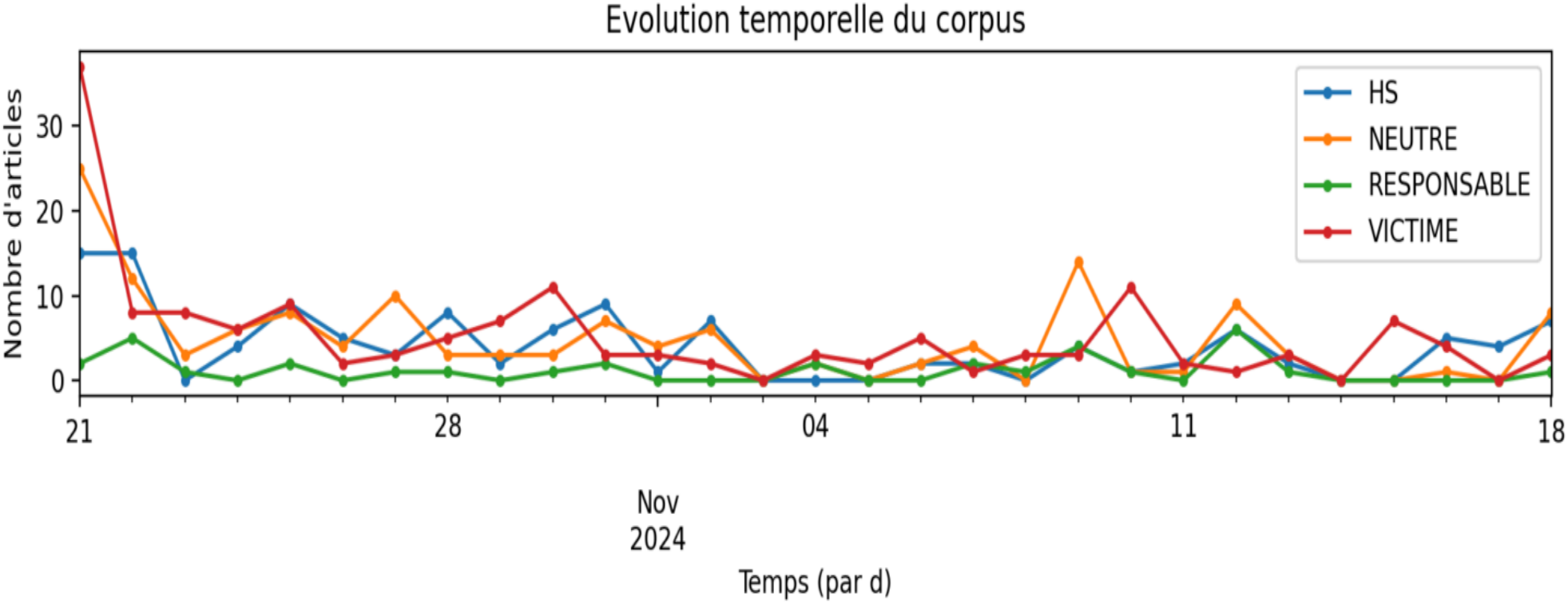
- Distribution (fréquence absolue, %)
 - Statistiques de colonnes
- Tableaux croisés dynamiques
 - Insertion > Tableau dynamique
- Graphiques
 - Insertion > Graphique

Réussir à avoir un graphique cohérent signifie avoir construit les données adaptées

Commencer par des analyses simples



Evolution temporelle avec dstool



D'autres stratégies d'analyse des données de presse

- Évolutions temporelles de plusieurs sous-corpus
 - entre journaux, entre périodes
- Détection de mots/comptage avec les macro excel
- Identifier des entités spécifiques
 - Remplacer par un token (par exemple : Le premier ministre, le PM, XXX par [ministre])

Et pour les entretiens ?

- Entretiens == du texte
- Démarche similaire
 - Mais un corpus pas encore constitué

Donc :

- Construire vous-même un tableau
 - une ligne par entretien, une colonne par info
 - ajouter les variables d'intérêt
 - **bien normaliser**

Encore plus de solutions

De nombreux logiciels

- Nettoyer des données avec [OpenRefine](#)
- Statistiques avec [Jamovi](#)
- Analyse textuelle avec [Iramuteq](#)
- Faire des cartes avec [kchartis](#) ou [Google Maps](#)
- Analyse des entités dans les textes avec [Cortext](#)

Les possibilités de la programmation

Par exemple faire un beau graphique (R ou Python)

- Données bien structurée
- Réflexion sur les objectifs
- Réaliser ensemble



Puissance des LLM

Faire tourner un LLM localement (ollama) ou utiliser une API et des scripts pour faire les requêtes avec des prompts.

La suite ?

Mettre en oeuvre

- Approches vues aujourd'hui sur votre sujet
- Réfléchir éventuellement à des approches plus spécifiques

Mot d'ordre : **intégrer** l'analyse à la problématique dans le dossier d'avancement (quitte à restreindre l'ambition).