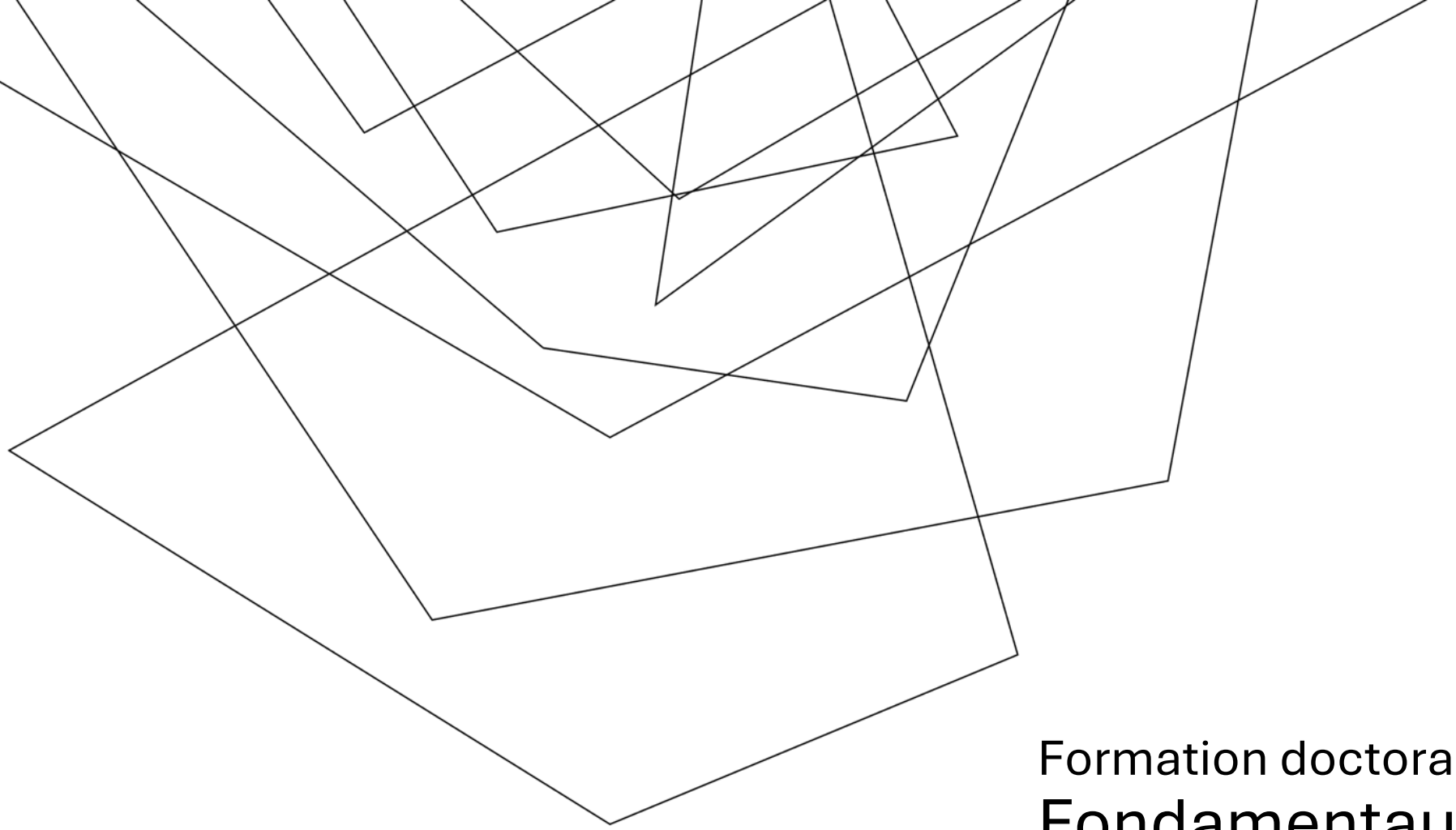


Formation doctorale

Fondamentaux de la recherche en sciences sociales

Victor Potier – LISIS
Victor.potier@univ-eiffel.fr

16 octobre 2026. 9h30 – 17h

PROGRAMME DE LA SÉANCE 5

Séance 5. Méthodes numériques pour les sciences sociales

Matin (9h30 – 12h30)

- 9h30 - 10h : tour de table = vos sujets de recherche
- 10h - 11h30 : présentation générale / Histoire d'Interne et création / Ce que le numérique fait aux sciences sociales
- 11h30 - 12h : présentation des outils sur le site du Medialab : Gephi, Minivan, Hyphe, le DMI

Après-midi : 13h30 - 17h

- Présentation de Cortext - Atelier exploratoire dans le cadre de vos sujets de recherche

Petit tour de table

Présentation

Discipline, parcours, sujet de recherche

Votre approche des méthodes numériques en sciences sociales

Attentes, préjugées, connaissances, critiques…

Votre connaissance du numérique et de l’enquête en ligne

Expériences passées ou en cours, attentes, limites ?

Partie 1. Présentation générale



Petite histoire d’Internet et des médias numériques

L’innovation portée par l’usage

Que savez-vous de la naissance…

- D’Internet ?
- Du Web ?



Numérique : une définition ?

- Peu de définition : mais on peut parler « d'informatique » : un **langage**, du code, des machines qui calculent…
- Alan Turing, 1937 : une machine à tout calculer grâce à l'invention du code binaire, première machine à appliquer les règles langagières booléenne (et / ou / non) à 1945 : premier ordinateur (US Army, balistique)
- ***Distinction*** Signal analogique : forme continue qui oscille entre deux valeurs // Signal numérique : deux valeurs, 0 et 1. Numérisation : tous les *dispositifs qui transforment* des sons, des mots, des vibrations… en 0 et 1

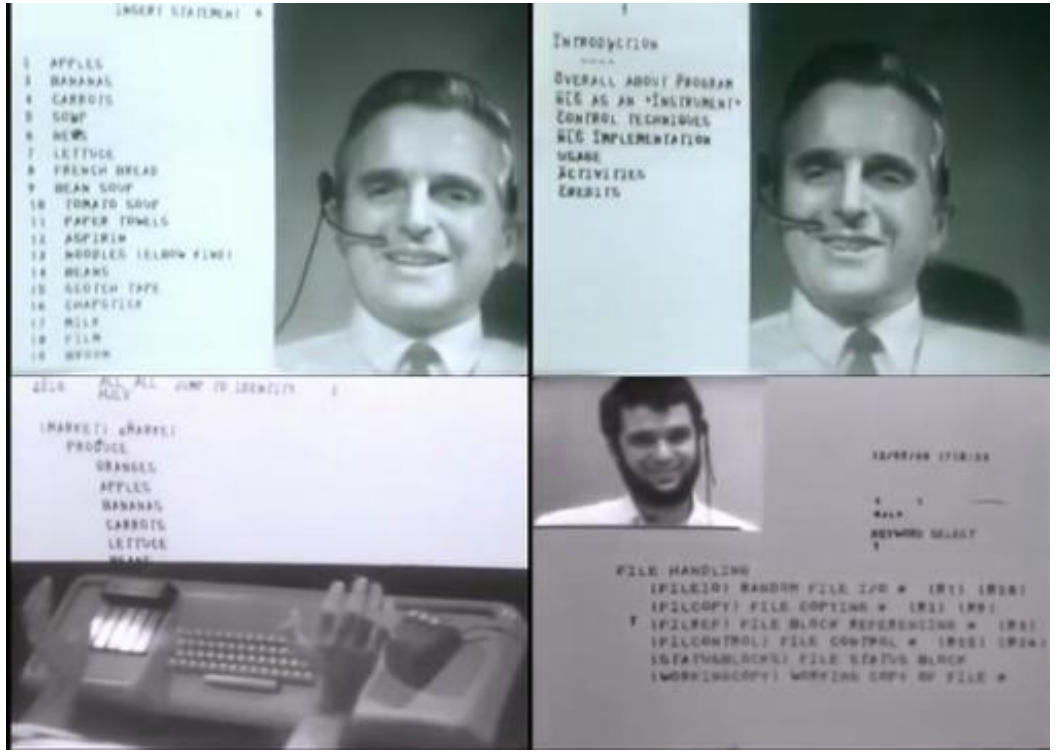


Mother of All Demos (1968)

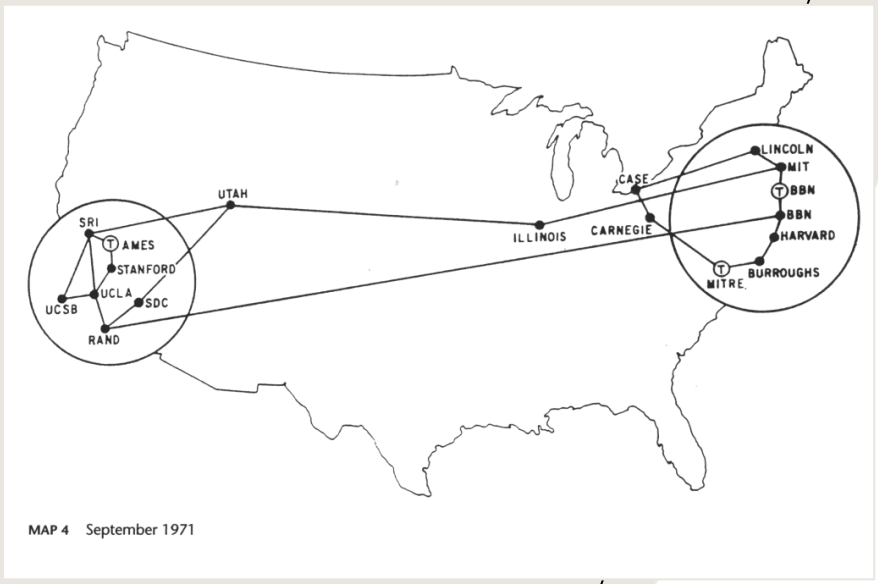
Interface graphique manipulable, sourie, changement de taille de fenêtre, travail coopératif sur un document, visioconférence, liens hypertextes…

→ Ne relie pas simplement des machines, mais des personnes qui interagissent par leur intermédiaire







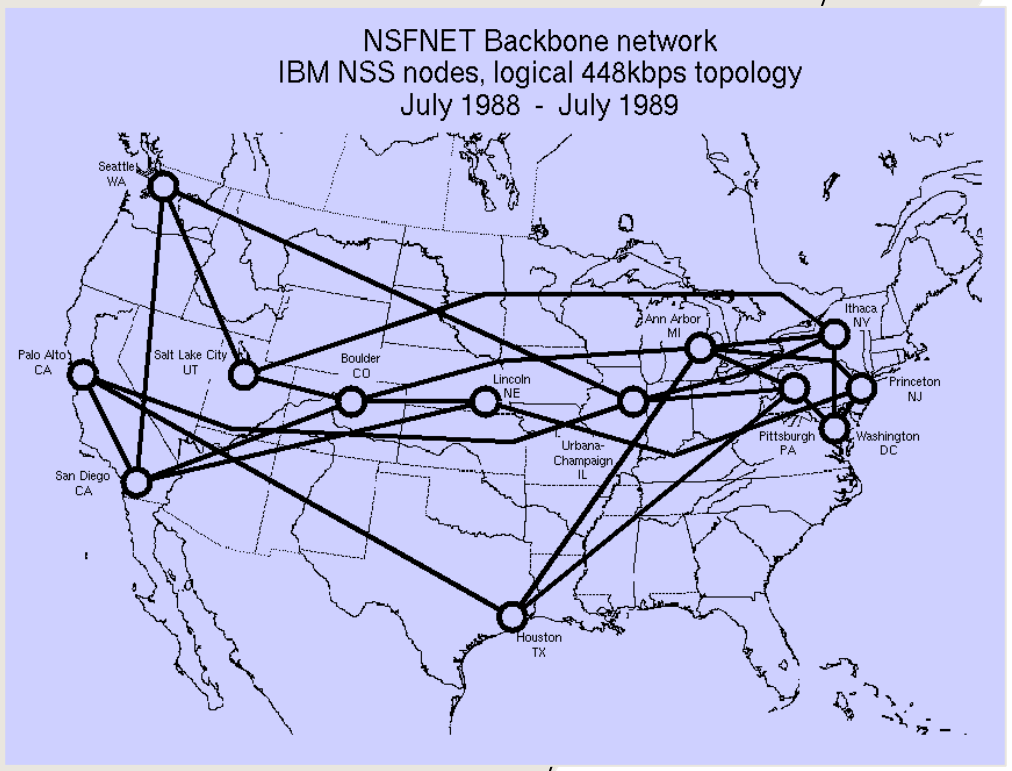


Trois étapes

1. Invention du protocole TCP/IP par l'ARPA = ARPANET (1969)



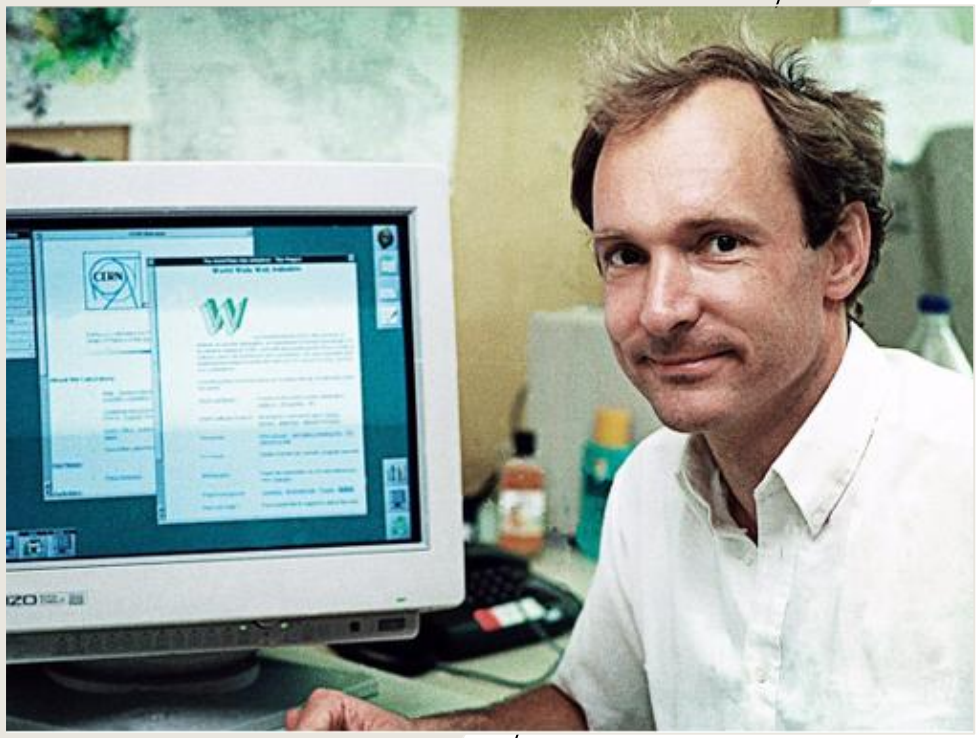
Le premier routeur : l' Interface Message Processor



Trois étapes

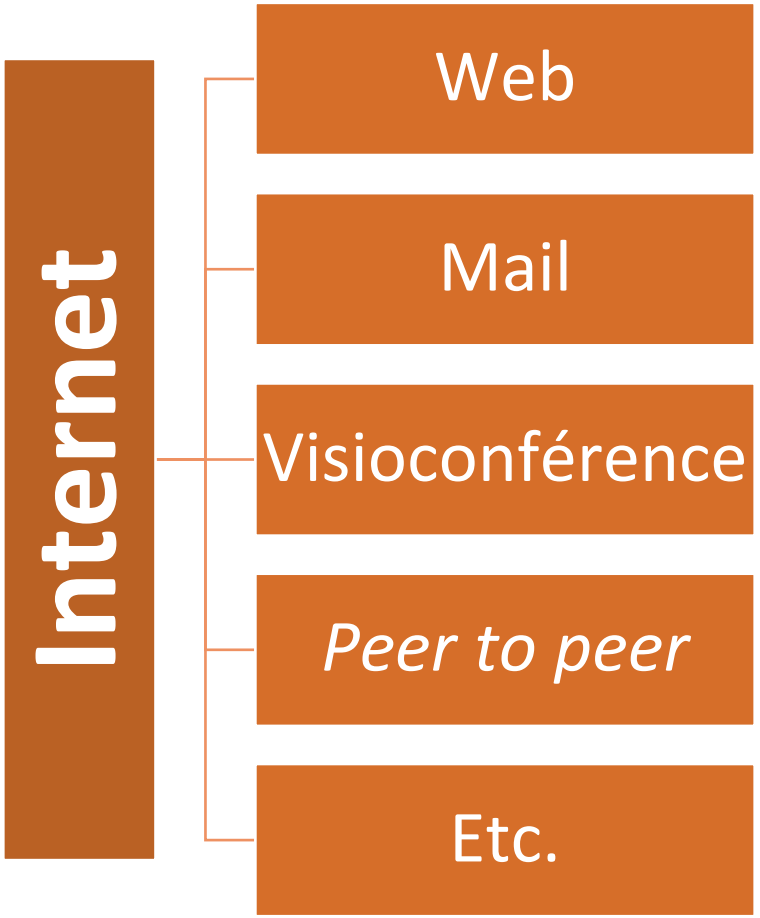
2. Le NSFnet, réseau de recherche académique (1985), puis ouverture commerciale (1991)

Trois étapes

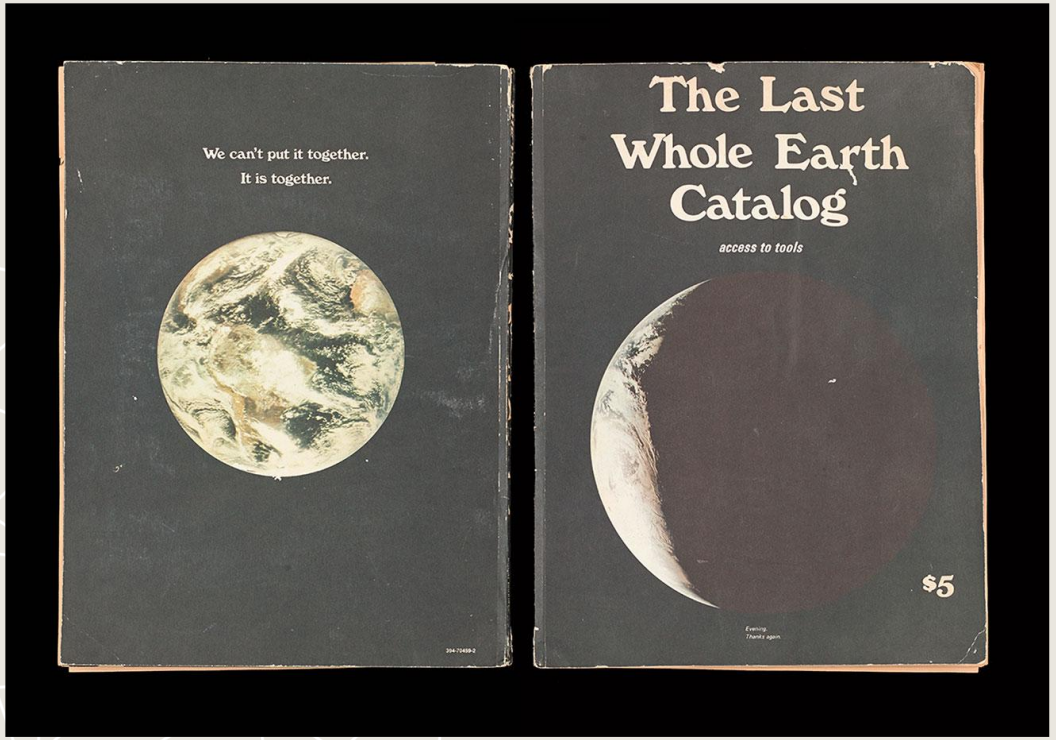


3. Invention de l’URL et du format World Wide Web par Tim Berners Lee au CERN (1989)

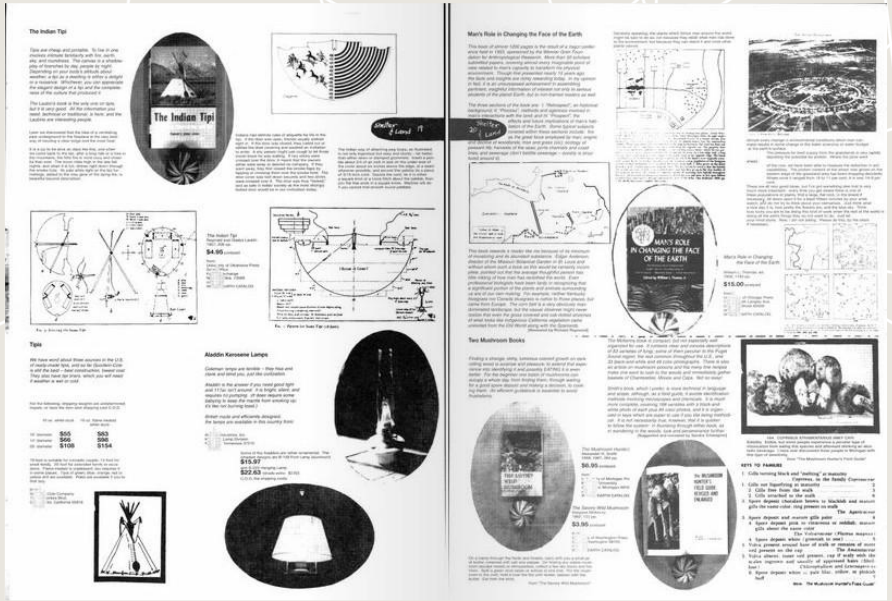
Internet vs. Web



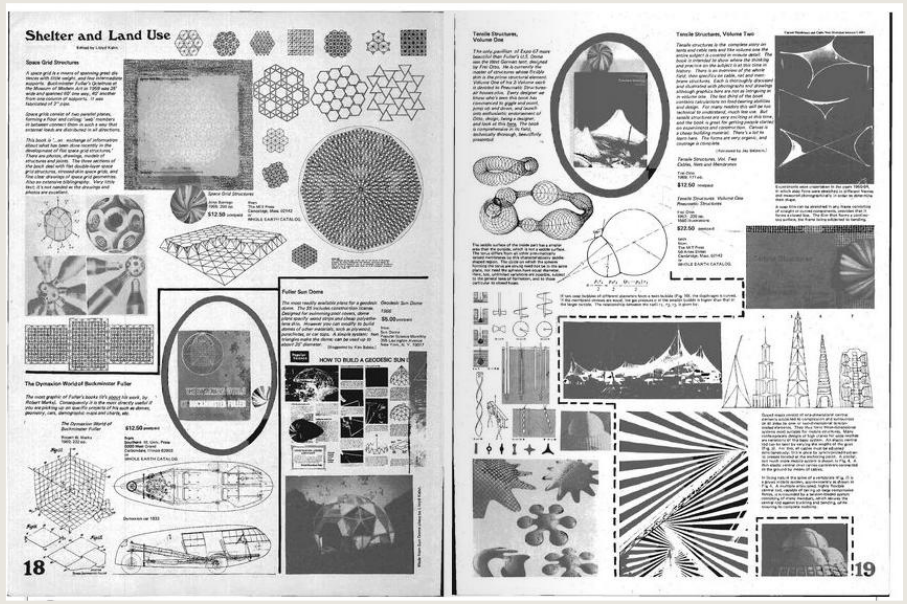
Whole Earth Catalogue, 1968 – 1972, Steward Brand

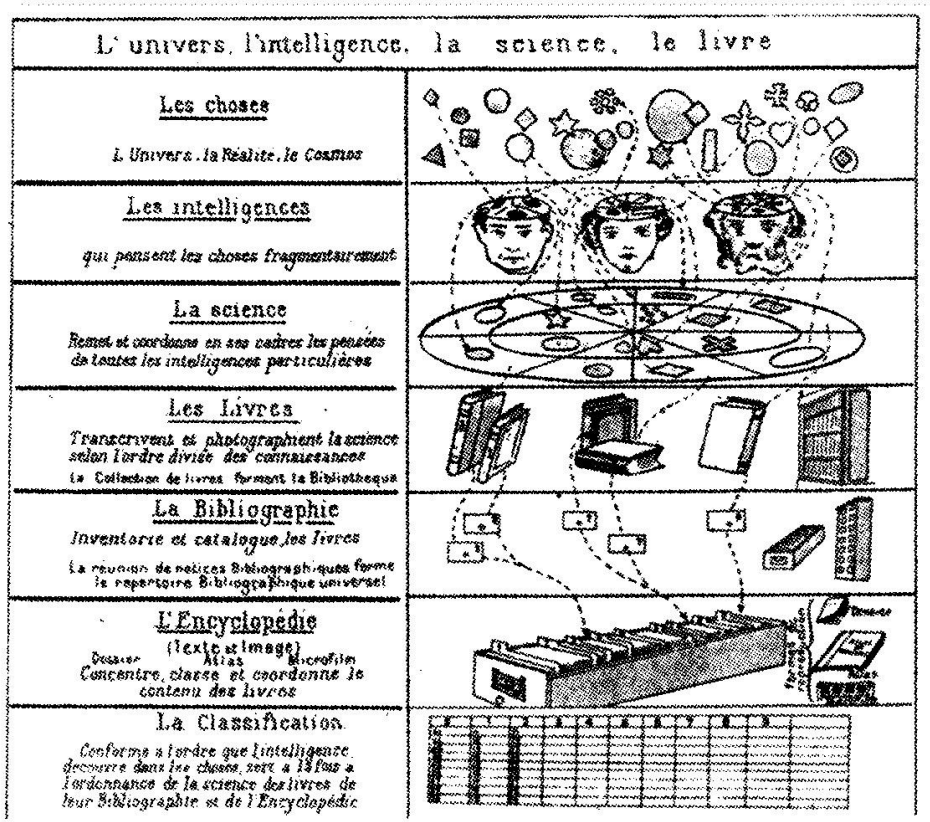


Résumés de livres, guide de vie, recettes végétariennes, almanachs mystiques bouddhistes ou New Age, conseils écolos, sciences, technologie et bricolage, etc.

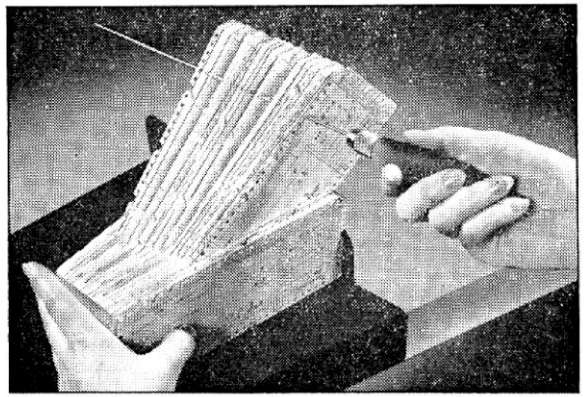


1 million d'exemplaires par an











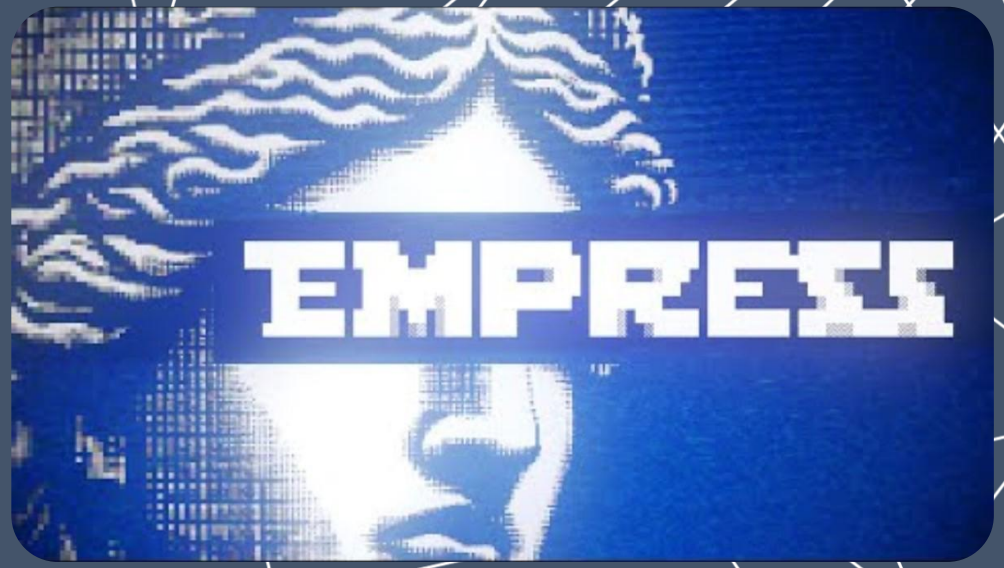
Le Mondanuem de Paul Otlet et Henri La Fontaine : ancêtre du lien hypertexte

En premier lieux = Investissements de l’armée : trajectoire balistique des bombes

- Logique de contrôle et de commande de l’armée qui prévaut, financement par le budget de la défense des ordinateurs et des premiers réseaux qui les relient… mais d’autres motivations interviennent aussi.
- Les pionniers enferment dans la technologie des valeurs, un esprit, une politique qui fixent durablement l’identité d’Internet : ce sont d’abord des choix techniques qui découlent d’une histoire, d’une culture
- Débats encore en cours : liberté d’expression, auto organisation, propriété intellectuelle, neutralité du net, économie du partage ou des plateformes, etc.

La coopération : l’ingrédient humain dans la création d’Internet

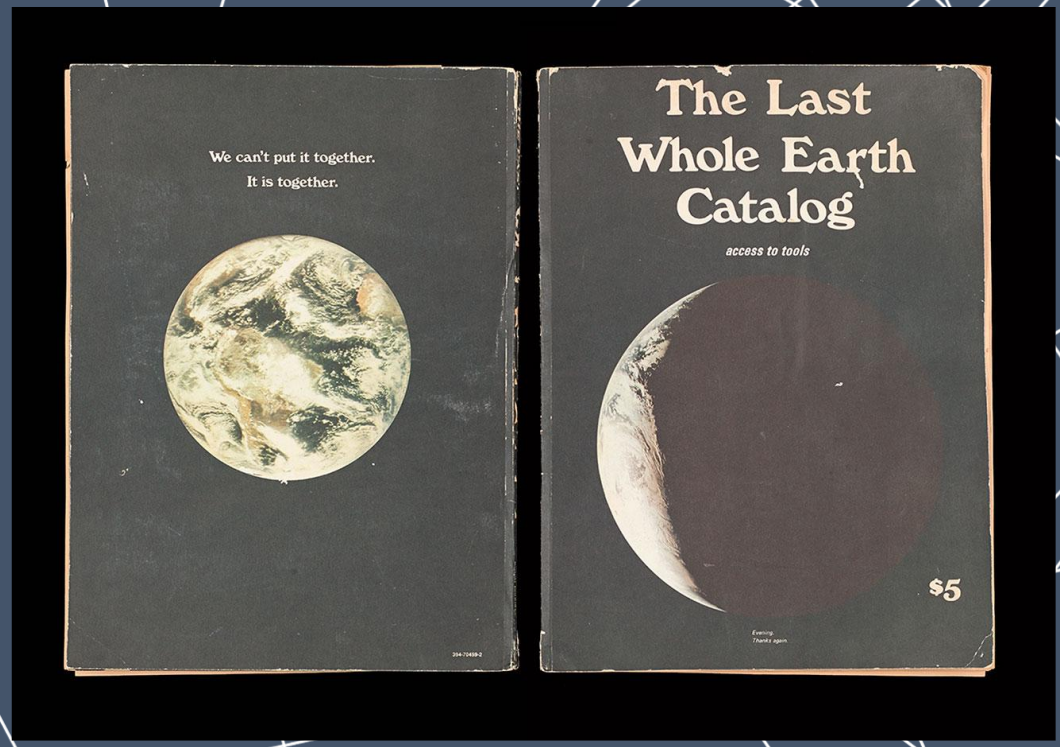
- Internet : un assemblage de technologies + une invention dans le processus d’innovation (usages)
- Premiers usagers : chercheurs, militaires et hippies. Leur point commun ? Coopérer ; une communauté de pairs, une « république des informaticiens ».
- Opposition entre les « Communs » et l’appropriation (patents) des logiciels. Mais à l’origine : logiciels libres, partagés, ouverts.
- Culture « hackers » : une pratique virtuose et curieuse du code, une forme technique et politique qui donne à la culture numérique son avant-garde, mais qui forme aussi une aristocratie des compétents



https://youtu.be/TKRpvgtyYNY

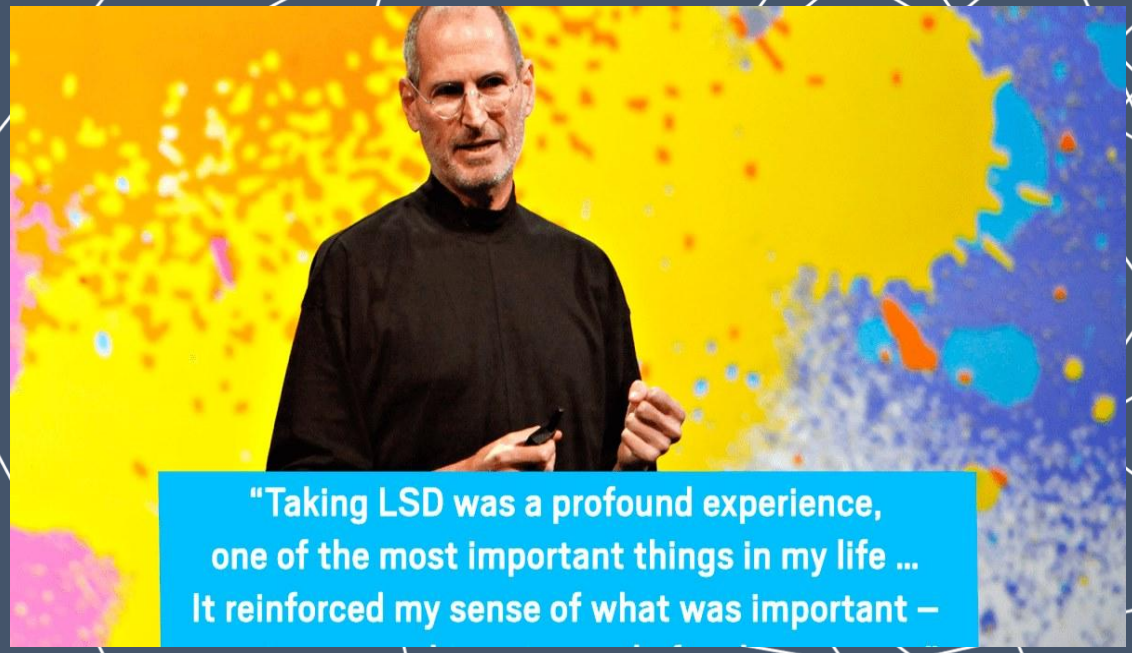
De la culture numérique au développement de l’informatique personnel

- Enjeux culturel et politique des pionniers (Doug Engerlbart) : Libérer les Hommes de leurs contraintes physiques, spatiales, temporelles, les sortir de leurs carcans culturels
- Proximité avec la contre-culture hippie : transformation de l’individu, libération de la subjectivité et changement des modes d’existences quotidien… sans prendre le pouvoir. Refonder des communautés (petits groupes en microsociétés). Cf. Le WEC de Steward Brand
- Inflexion dans la trajectoire de l’informatique : les industriels ne voient aucun intérêt à rendre l’ordinateur individuel VS les hippies réclament qu’on le donne aux gens, qu’on puisse le fabriquer, le bidouiller, le programmer… s’approprier la science et la technique, au service de l’individu à DIY (punks, amateurs – voir P. Flichy – et Makers en Fablab, voir aussi la critique de l'Ecole de Francfort, ou celle de Jacques Ellul sur la technique)



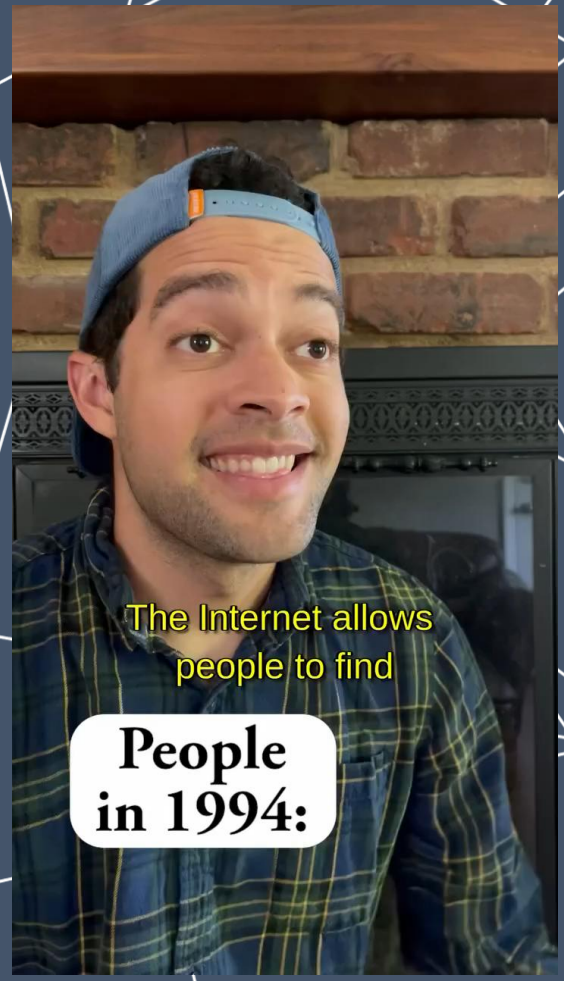
Cinq caractéristiques du projet politique des pionniers

- Un projet d'émancipation individuel
- Un projet de construction de nouvelles communautés, choisies et non contraintes, sur d'autres bases que celles de nos assignations sociales
- Changer la société sans prendre le pouvoir
- Pas d'incompatibilité entre la société des individus et le marché (… vers l'économie des plateformes ?) La culture numérique est traversée d'une tension entre valeurs libertaires et libertariennes
- La technologie est un instrument d'action politique qui peut changer la société (… projet politique de changement et de solutionnisme aux problèmes du monde)



Les pionniers lèguent une série d’ambivalences : créatif mais marchand, communautaire mais globalisé, librement expressif mais aristocratique, ouvert mais inégalitaire, etc.





@shadeofgreaves

Partie 2. Les défis que le numérique lance aux sciences sociales

Dominique Cardon

CULTURE NUMÉRIQUE

LES PETITES HUMANITÉS

L'entrée du numérique dans nos sociétés est souvent comparée aux grandes ruptures technologiques des révolutions industrielles. En réalité, c'est avec l'invention de l'imprimerie que la comparaison s'impose, car la révolution digitale est avant tout d'ordre cognitif. Elle est venue insérer des connaissances et des informations dans tous les aspects de nos vies. Jusqu'aux machines, qu'elle est en train de rendre intelligentes. Si nous fabriquons le numérique, il nous fabrique aussi. Voilà pourquoi il est indispensable que nous nous forgions une culture numérique.

SciencesPo LES PRESSES

SCIENCES HUMAINES ET SOCIALES

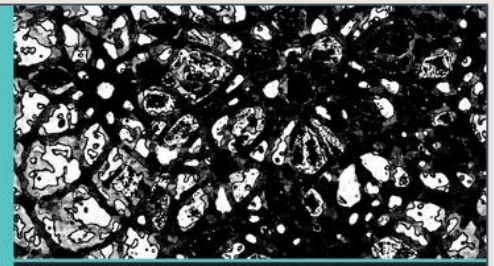
collection U

Sous la direction de Christine BARATS

MANUEL D'ANALYSE DU WEB



ARMAND COLIN



Sciences sociales

Dominique Pasquier

L'Internet des familles modestes

Enquête dans la France rurale

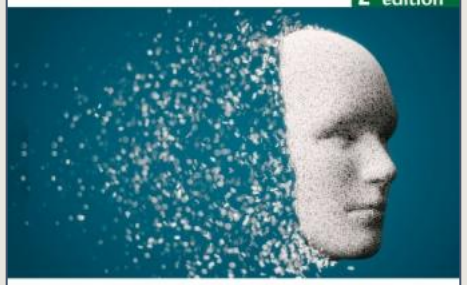
PSL

Presses des Mines

Dominique BOULLIER

SOCIOLOGIE DU NUMÉRIQUE

2e édition



ARMAND COLIN

JEAN-SAMUEL BEUSCART
ÉRIC DAGIRAL
SYLVAIN PARASIE

Sociologie d'internet

CURSUS



2e édition

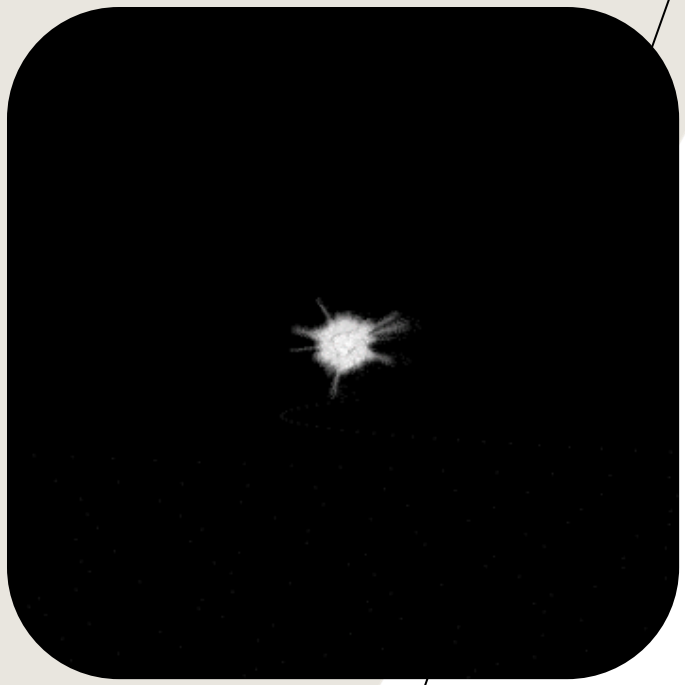
ARMAND COLIN

Baptiste Kotras

La voix du web

Nouveaux régimes de l'opinion sur Internet

LA RÉPUBLIQUE DES IDÉES | Seuil



Un changement dans l'économie de la production de la connaissance

- Bataille des **plateformes de publication** : ouvertes (Persée, Plan Science Ouverte en 2018 en France, Initiative de l'Union Européenne) vs. fermées (Elsevier par ex., système premium et accaparement de la plus-value)
- Hausse de la **scientométrie** (impact factor) = devient normative = modèle de management libéral classique dans la production et la valorisation de la science (*publish or perish*)
- Mais : meilleure circulation des **ressources détaillées** (rapports, sources, méthodologies, *preprints*, etc.) = des chercheurs en **réseau** !
- Reconnexion des sciences à la société ? Nouveaux « **forums** » de production des données. Par ex : Algopol / Activistes, politistes du CEE de Sciences Po et data scientists du Medialab qui créent La Fabrique de la Loi

2 - <img alt="Photo by Hector Suarez in Islas Afortunadas - Canarias. **L'image contient peut-être : 1 personne, lunettes_soleil, ciel et plein air.**" class="FFVAD" decoding="auto" style="object-fit: cover;" sizes="598px" srcset="https://instagram.fcdg1-1.fna.fbcdn.net/v/t51.2885-15/sh0.08/e35/s640x640/118673185_334022218010542_525557579010736289_n.jpg1.fna.fbcdn.net&_nc_cat=110&_nc_ohc=lvwCVcm_JGYAX9QnRCx&oh=2f40cf2c6044778e7ee9f510c8f6f0b8&oe=5F818F1D 1080w"



Un renouveau ? Un idéal ?

- La place des **images** ? Question d'autant plus importante que la pratique de la photographie se massifie en ligne (72 milliards de photos sur Facebook en 2017) : nouveaux champs d'application de la sociologie visuelle. Ex : les mèmes, les photos de vacance, la recherche d'images en ligne, etc.

- **Réseaux et data visualisation** ? **Opportunité** : exploration par les graphes & continuité entre les échelles (Qualiquantitative, Latour). **Mais** : la narrativité scientifique risque de pencher vers un story telling (méconnaissances) & grand ***malentendu*** entre réseau social et réseau social

- **Réseaux du web** = l'analyse des traces et la façon dont la socialisation produit des attachements préférentiels formant une topologie du web / **analyse sociologique de réseaux** = propriété sociale des individus, forces des liens, régularité, distance, effets de la structure sociale sur les comportements individuels, etc.



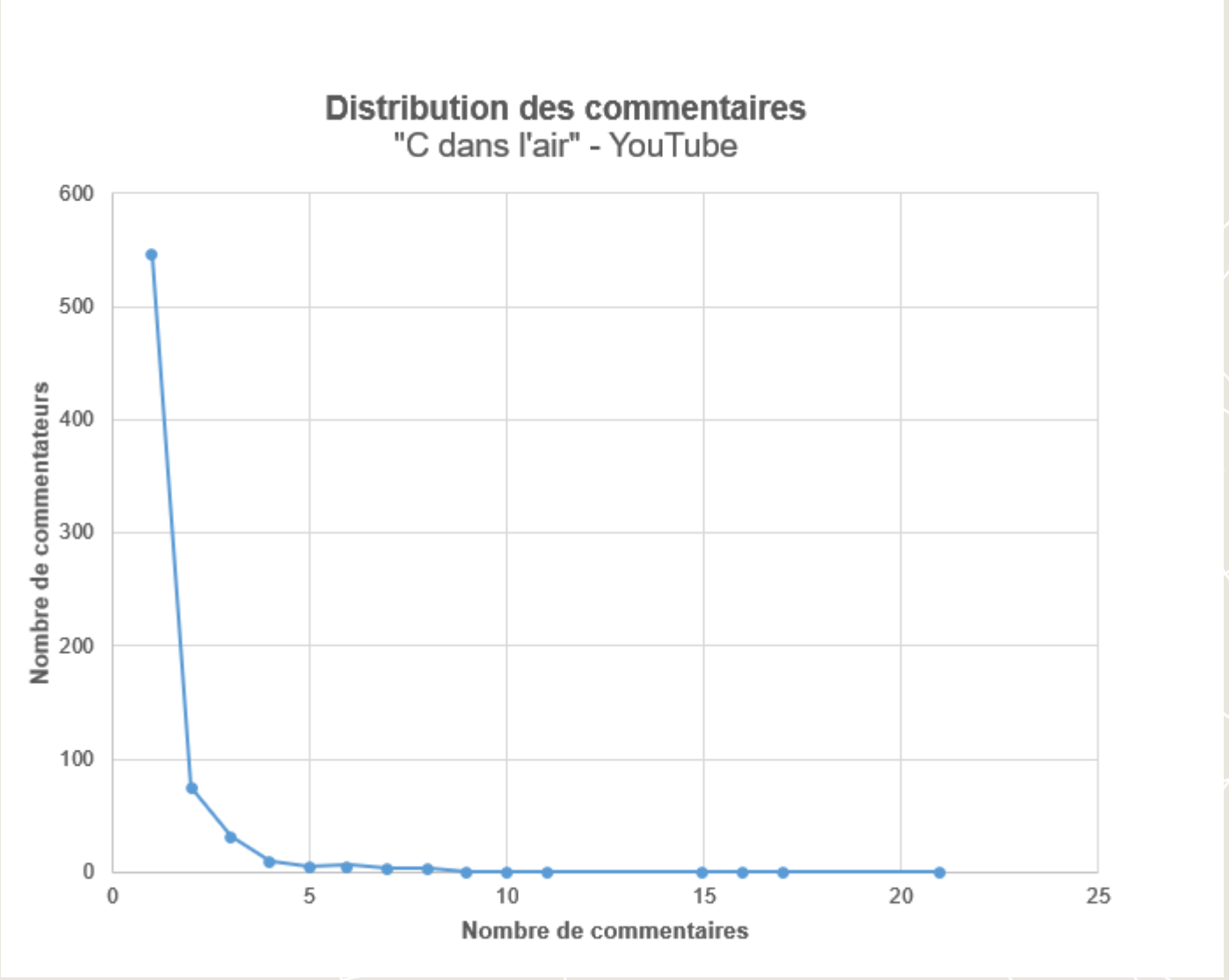
Quelles méthodes pour quels enjeux ?

- **Deux traditions face à face** : Big Data (quanti) contre Media Sciences (épistémologie du web, analyse sémantique de contenus, etc.) … voir des approches très ethnographiques (ex. : entretien semi-directif avec une donnée)
- **La question méthodo** : faut-il se déprendre du « tout » (exhaustivité, représentativité) pour remplacer par d'autres valeurs des données (vélocité, traçabilité, analyse des réplications, etc.)
- **Traces et Big Data (= corrélations**) : « La fin de la théorie » (C. Anderson dans *Wired* en 2008) ? En creux se dessine un nouvel enjeu gigantesque : les sciences sociales doivent attester de leur valeur ajoutée face à la mosaïque des corrélations permanentes calculées statistiquement

Les inégalités structurelles du web

Loi de puissance : distribution statistique très inégale dans laquelle un tout petit nombre d'individus de la population considérée concentre l'essentiel de la valeur considérée.

Ici : sur une population de 688 commentateurs différents, les 68 commentateurs les plus actifs (le top 10%) cumulent à eux seuls 441 commentaires, soit 42% de l'ensemble des commentaires produits.



Étudier le web en sciences sociales

Quels terrains pour les sciences sociales sur le web ?

« Ce sont *des **pratiques médiatiques**, qui renouvellent celles qui sont associées aux médias de masse tels que la presse, la radio de la télévision. Internet, ce sont en outre des **pratiques culturelles** ou de **loisirs**, dès lors que les individus y trouvent le moyen d'accéder à des œuvres, de les produire ou de les détourner. Les pratiques associées à internet renvoient également à **l'économie**, dès lors qu'il met en relation des consommateurs et des producteurs autour d'un grand nombre de biens et de services. Mais ce sont aussi des **pratiques de travail**, pour un grand nombre d'individus qui travaillent à distance, interagissent avec des clients en ligne ou dont la productivité est évaluée par le biais d'internet ; tout comme des **pratiques politiques** et **militantes**, également constitutives du réseau* »

Beuscart et al. (2019), *Sociologie d'internet*. Paris: Armand Colin.



Quelles méthodes pour quels enjeux ?

- Comment saisir / sauvegarder la temporalité des traces en ligne qui se sédimentent et disparaissent ?
- Quels enjeux éthiques / légaux (ex. le scraping ?)
- Approche textométrique : comment constituer les corpus ? Faut-il privilégier une approche sémiologique ou lexicométrique ?
- Approche ethnographique : gérer l'ambivalence entre les persona / les individus (enquêter l'anonymat ?)
- Quels outils de quantification des flux, des réseaux, des pratiques en ligne ? Quelles métriques pour saisir l'audience ou l'opinion ?

Panorama des approches en sciences sociales numériques



Les méthodes classiques fonctionnent toujours

Entretiens, observations, enquête par questionnaire, etc.

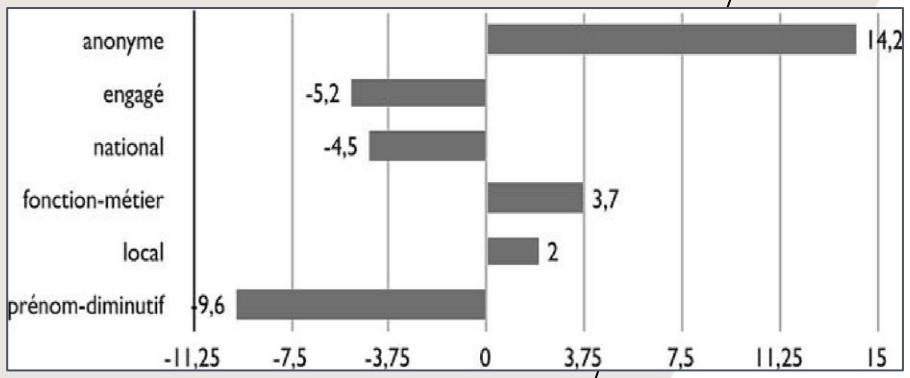
« Approches computationnelles » ?

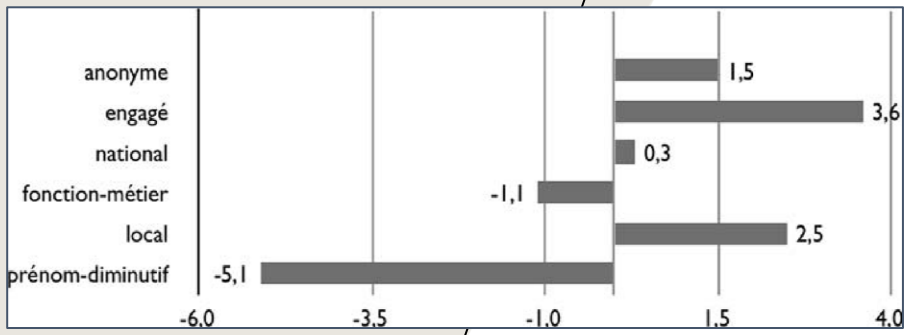
→ Collecte et l'analyse, avec un programme informatique ou non, des traces numériques = une archéologie d'un phénomène social

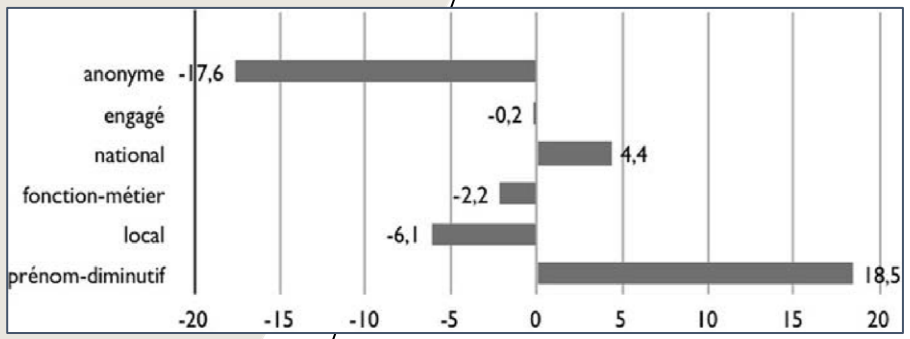
« Traces numériques » ?

- Publications diverses : messages et textes, images, photos, sites sur les avis marchands, gifs, émojis, etc.
- Liens hypertextes
- Partages, retweets, likes, etc.
- Notes, étoiles, évaluations… le « sentiment » des usagers
- Playlists, usages, dynamiques communautaires dans un JV, etc.

Cf. PARASIE S. et J.-P. COINTET, « La presse en ligne au service de la démocratie locale » Une analyse morphologique de forums politiques. *Revue française de science politique*, 2012/1 Vol. 62, p. 45-70.



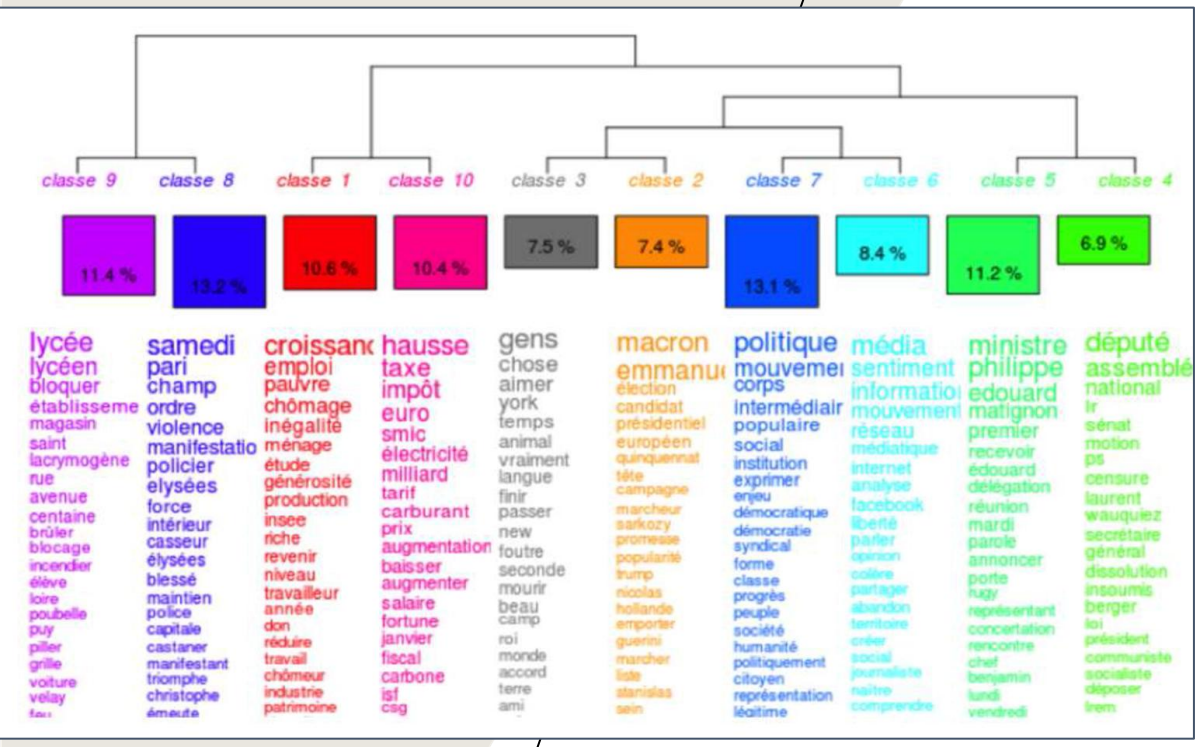




(1) Analyse qualitative & netnographie

- Analyse (qualitative) de corpus : collecte de données a posteriori (archives d'un forum, d'une discussion, de publications diverses, etc.).
- Ethnographie en ligne : observer les interactions au moment même où elles se déroulent.
  - Ci-contre : analyse des types de pseudonymes adoptés par les contributeurs des forums de La Voix du Nord en fonction de la taille de leur commune de résidence (petite, moyenne, grande).

Smyrnaios et al. (2020), "Les Gilets Jaunes, étude d'un mouvement social au prisme de ses arènes médiatiques". Terminal [en ligne], 127 | 2020



(2) Traitement automatique du langage

Grande variété des traitements algorithmiques possibles du langage (à condition de constituer des corpus massifs)

- *Sentiment analysis* (analyse de la tonalité positive ou négative)
- Lexicométrie (fréquence des termes)
- Modélisation de catégories sémantiques (cooccurrence de termes, registres d'argumentation, etc.), par exemple avec CorText ou Iramuteq.



Sur l’analyse textuelle

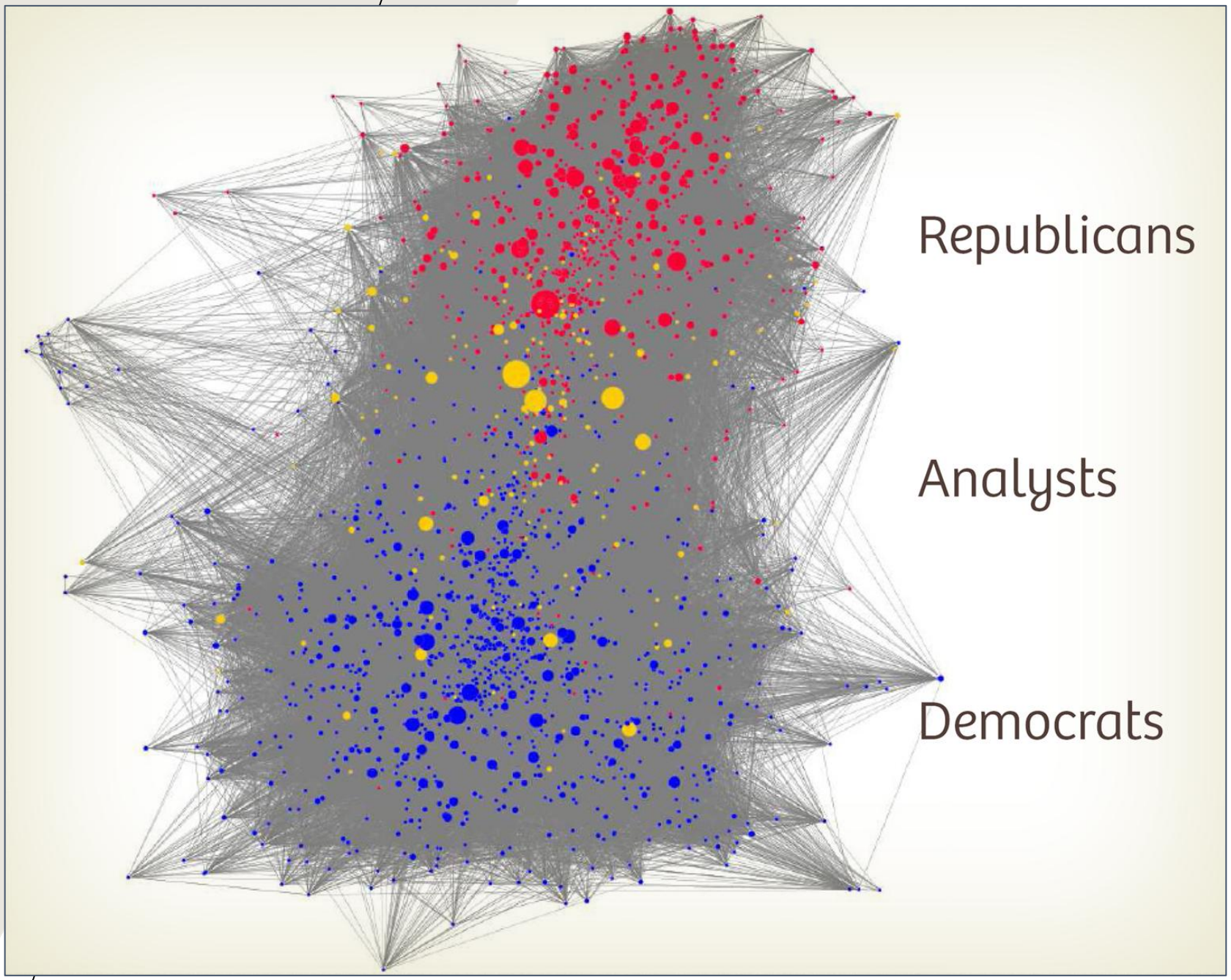
- L’analyse textuelle est pleine de promesses : un·e chercheur·se seul·e peut traiter des milliers de documents et pas seulement les 4 entretiens qu’il aura eu le temps de faire et de retranscrire.
- Mais recourir à un logiciel ne dispense pas de connaître les documents qu’on veut analyser, leur contexte de production, et de maîtriser les outils.



(3) Cartographie du web & network analysis

Études pionnières dans les années 2000 : considérer les liens hypertextes comme des objets sociaux reliés entre eux pour une raison

- Utilisation d'un « crawler » qui parcourt et indexe le web : les liens ne se distribuent pas de façon aléatoire = « clusters », grappes connectées
- Analyses quali : analyse des contenus (sites Web, communautés) / analyse topologique (centres, périphéries, rapprochements, éloignements…)
- Analyse quanti : métriques spécifiques de l'analyse de réseaux (densité, centralité, intermédiarité, etc.) : mesurer l'autorité de certains sites, les échanges entre communautés, etc.


Republicans
Analysts
Democrats

Matrice des sources de données et méthodes d'analyse en sociologie

Sources / Méthodes	**Qualitative** (interprétation, compréhension du sens)	**Quantitative** (mesure, généralisation)	**Computationnelle** (analyse assistée par ordinateur, grands corpus)
Entretiens	Analyse thématique, codage manuel, interprétation du discours	Comptage de variables sociodémographiques issues des questionnaires d'entretien	Analyse automatique du texte (text mining, détection de mots-clés, sentiment analysis)
Observations (directe, indirecte, participante, non-participante, etc.)	Notes ethnographiques, analyse de situations, récits détaillés	Codification des comportements observés (ex. fréquence des interactions, chronométrage des postures, des gestes, etc.)	Vidéo-analyse avec détection automatique (tracking, reconnaissance de patterns)
Bases de données (enquêtes statistiques, registres, etc.)	Études de cas qualitatives avec extractions de trajectoires individuelles ou de portraits / ethnographie des conditions de production de la base de données	Analyses statistiques descriptives et inférentielles (régressions, corrélations, tests)	Usage du machine learning pour classification, clustering, prédiction sur de grands échantillons
Ressources en ligne (forums, réseaux sociaux, blogs, etc.)	Analyse discursive ou narrative de postes, de discussions, etc.	Comptages, sondages en ligne, métriques de mesure d'engagement (likes, partages)	Scraping pour analyse de réseaux sociaux (SNA), constitution de larges corpus textuels et Traitement Automatisé du Langage : reconnaissance d'entités de lieux, de personnes, etc. ; topic modeling, etc. ; détection d'opinions ou dynamiques virales, etc.
Archives et documentations (textes, dossiers, presse, etc.)	Analyse de contenu, contextualisation historique, analyse de discours	Séries temporelles (par exemple fréquence d'un terme dans la presse au fil du temps), quantification de l'usage de termes, de la taille des documents, de la fréquence de leur production, etc.	OCR + traitement automatique de corpus massifs (topic modeling, stylométrie, IA pour catégorisation par entraînement LLM)

Partie 3. Présentation de quelques banques d’outils en ligne

Rendez-vous sur le site du Medialab

- Le site du Medialab : https://medialab.sciencespo.fr/outils
- Le site du Digital Methods Initiative : https://wiki.digitalmethods.net/Dmi/ToolDatabase
- Le site du Démonstrateur (LISIS) : https://demonstra.hypotheses.org/

Une chaîne méthodologique axée sur l'exploration

1. Sourcing
   Define your field a priori
   and gather starting points

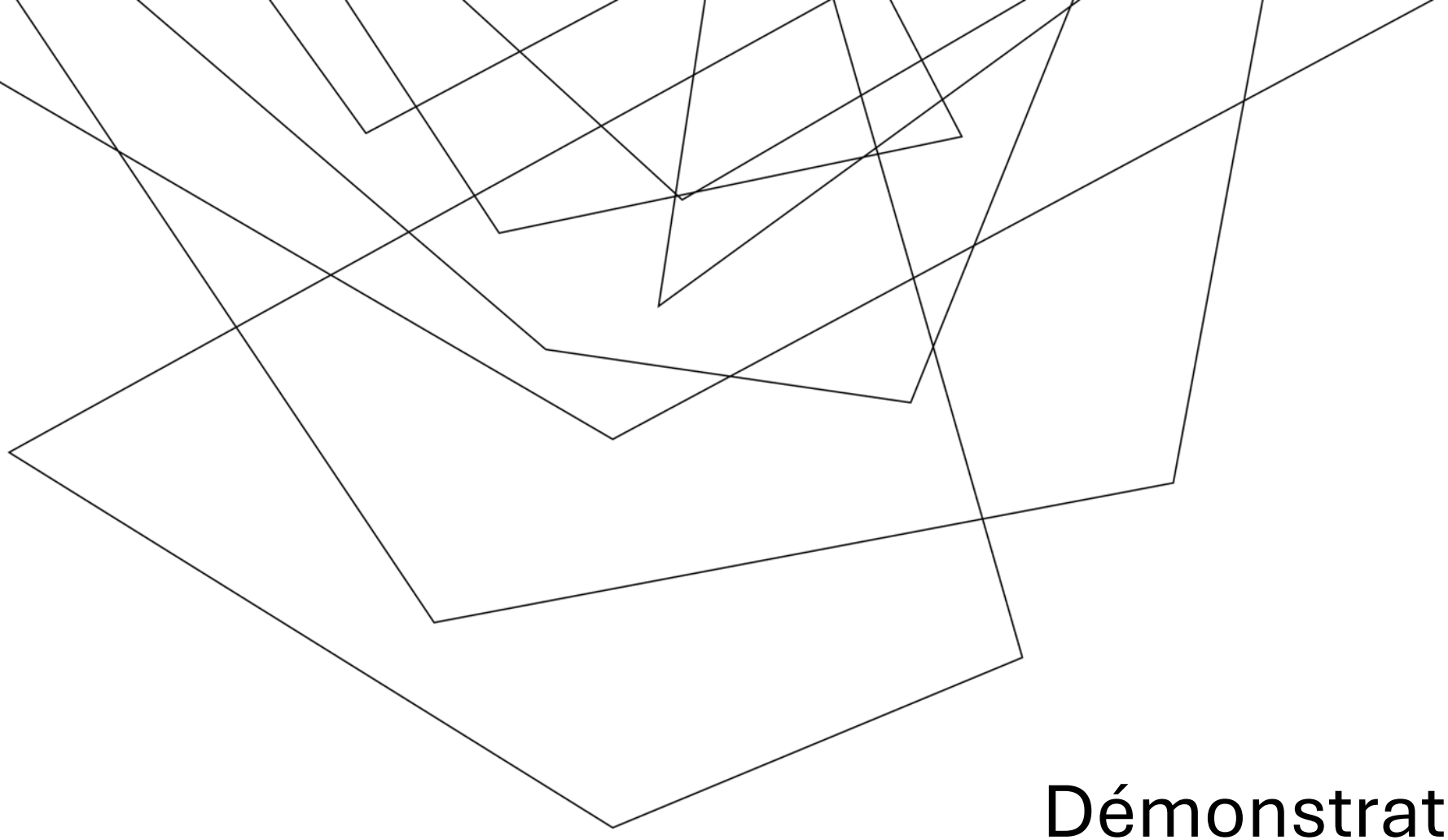
2. Harvesting (crawl)
   Download the data
   with a crawler



3. Monitoring
   Visualize corpus
   and monitor its properties

4. Curation
   Select documents to limit
   topic drifting and adjust
   corpus boundaries

5. Finalization
   Validate general quality
   and export corpus



Démonstration et mise en usage

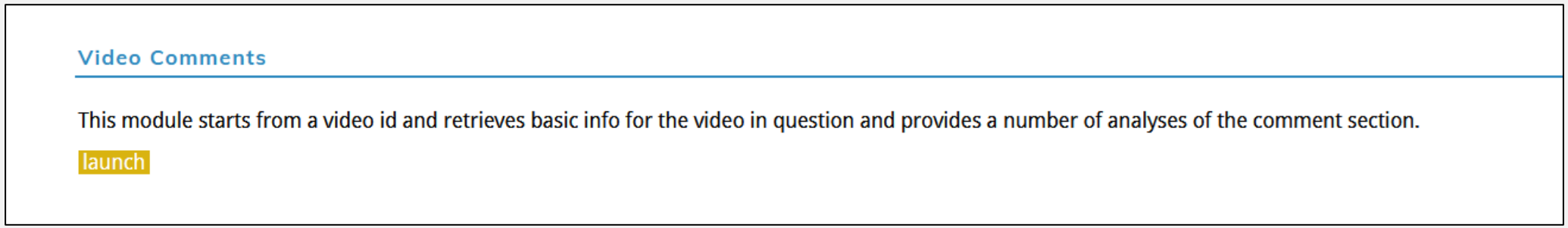
YouTube Data Tool - CorText

Principe de l’exercice

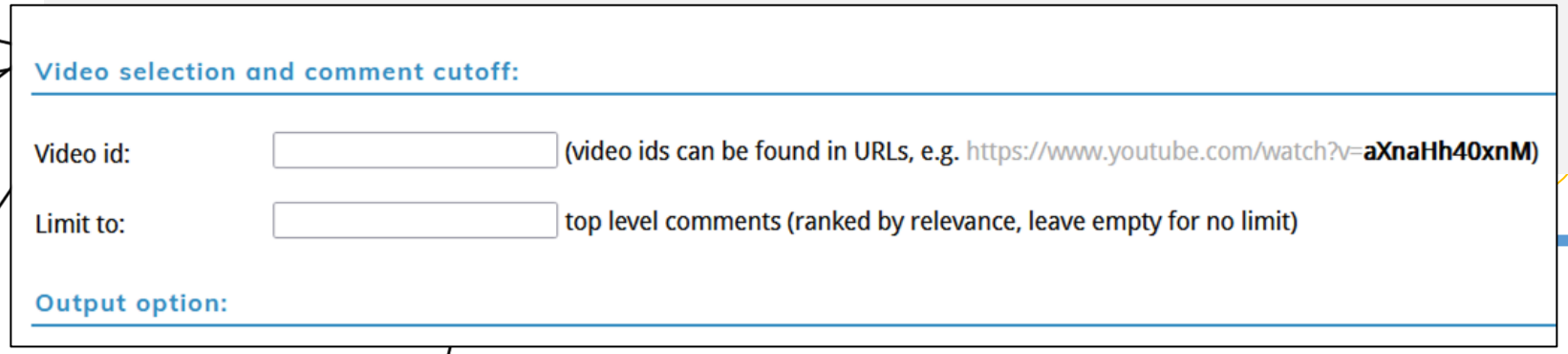
- Travail sur une vidéo extrêmement commentée : « L'histoire coloniale derrière la guerre Israël-Palestine » - *Histoires Crépues* – 17 oct 2023 – 10 981 commentaires
- Utilisation de YouTube Data Tool pour récupérer les commentaires
- Mise en œuvre d’un traitement sur CorText pour en faire une analyse thématique
- Tuto d’usage de CorText

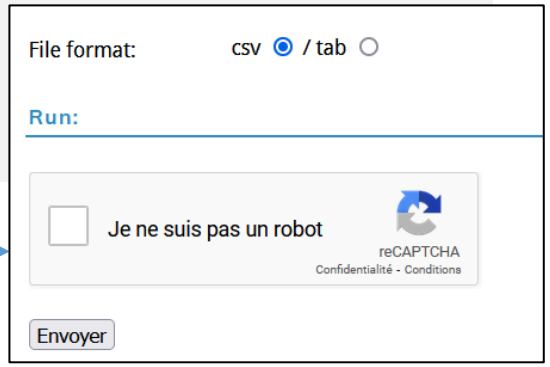
DÉMO DE YOUTUBE DATA TOOL

1. SÉLECTIONNER LE MODULE « VIDEO COMMENT »



2. RENSEIGNER LA « VIDEO ID », LE FORMAT DE SORTIE EN CSV, ET « RUN »





CONSTRUCTION DE VOTRE CORPUS

DÉMO DE YOUTUBE DATA TOOL

3. EXPORTER LES RÉSULTATS (COMMENTS UNIQUEMENT)

Results

Getting comments: 100 200 300 400 500 600 700 752

Digging into thread structure: 0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 290 300 310 320 330 340 350 360 370 380 390 400 410 420 430 440 450 460 470 480 490 500 510 520 530 540 550 560 570 580 590 600 610 620 630 640 650 660 670 680 690 700 710 720 730 740 750

The script retrieved 1403 comments from 752 top level comments.

Generating network.

The following files have been generated:

videoinfo_Ouy82eJLEk0_2024_01_23-11_44_41_basicinfo.csv
videoinfo_Ouy82eJLEk0_2024_01_23-11_44_41_comments.csv
videoinfo_Ouy82eJLEk0_2024_01_23-11_44_41_authors.csv
videoinfo_Ouy82eJLEk0_2024_01_23-11_44_41_commentnetwork.gdf

APARTÉ : LES .CSV, *KÉZAKO* ?

- Des tableurs dont les colonnes sont délimitées (séparées) par des caractères spécifiques (fréquemment virgules, points-virgules)
- Pour ouvrir votre fichier .csv
  - Avec Excel : sélectionner la première colonne > Données > convertir > délimité > virgule > terminer → enregistrez-le pour conserver la conversion
  - Avec LibreOffice : quand le fichier s'ouvre, vérifiez que « séparé par » « virgule » ou « point virgule » sont bien cochés

NB : ne prêtez pas attention à l'encodage pour le moment, CorText s'en sort mieux qu'Excel ☺ (Ex : « *Jâ€™ai beaucoup apprÃ©ciÃ© le discours de notre PrÃ©sident de la RÃ©publique ðŸ‘ðŸ‘ðŸ‡«ðŸ‡·ðŸ‘ðŸðŸ‘ðŸ* »)

DÉMO DE YOUTUBE DATA TOOL

4. [*ATTENTION PARTIE PLUS RÉBARBATIVE*] VOUS OBTENEZ UN FICHIER CSV PAR VIDÉO AVEC DES COLONNES COMME SUIT. PAS SUFFISANT ☺

	A	B	C	D	E	F	G	H	I	J	K
1	id	replyCount	likeCount	publishedAt	authorName	text	authorChannelId	authorChannelUrl	isReply	isReplyTo	isReplyToName
19	UgwrGQ2pPEUaVwzseqF4A		0	17/01/2024	@goumadepa	Macron e	UCA_d1wTDKiVhMM	http://www.youtube	1	UgwrGQ2pPE	@janvandaele8283
20	Ugxdpnwzw7	0	0	22/01/2024	@laurentdeni	un pays sa	UCqloUJBa4IKW_qR	http://www.youtube	0		

- **AJOUTER UNE COLONNE « SOUSCORPUS »** DANS LAQUELLE VOUS INDIQUEZ LA PROVENANCE (COMPTE DU VIDÉASTE, ÉVÉNEMENT EN PARTICULIER, ETC.). POSSIBILITÉ DE FAIRE PLUSIEURS COLONNES EN PLUS.
- ATTENTION : PAS D'ACCENT, PAS D'ESPACE !
- **REFORMATER LA DATE** EN ANNEE-MOIS-JOUR (AU CAS OÙ ON FERAIT TOURNER DE FUTURS SCRIPTS COMME *EPIC EPOCH*)

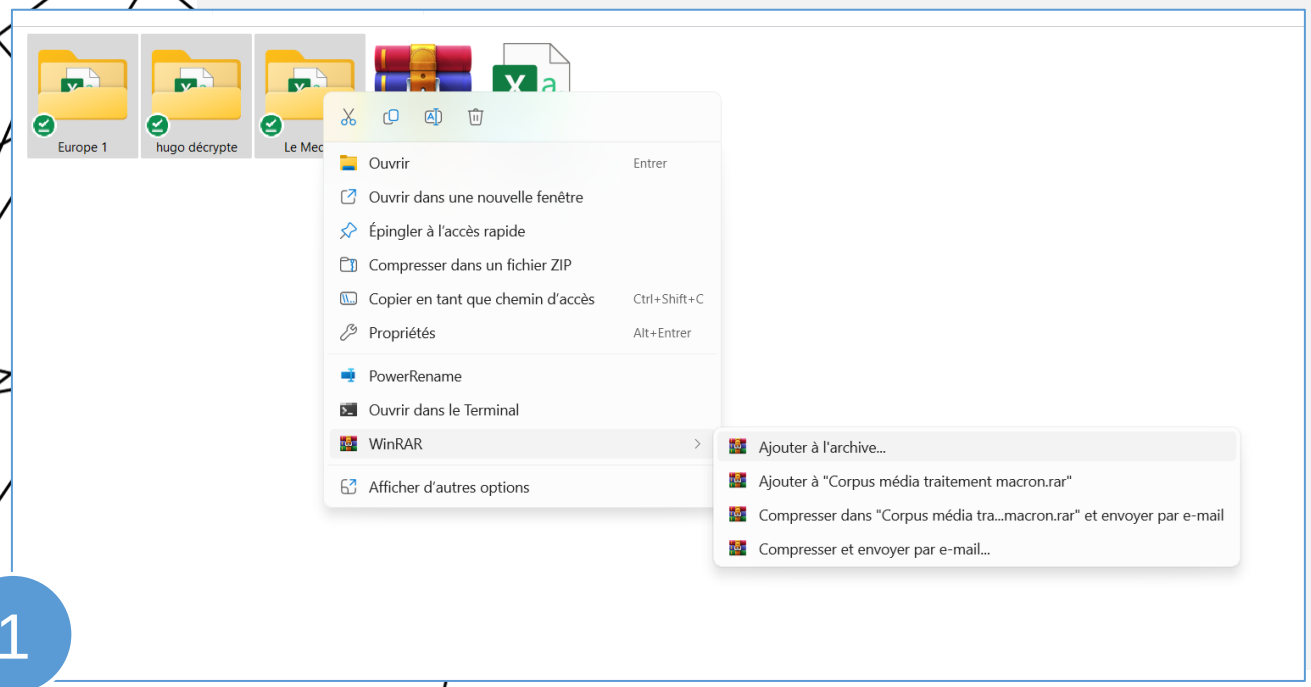
ASTUCE EN CAS DE TROP NOMBREUX FICHIERS : FAIRE UN FICHIER QUI CONCATÈNE TOUS LES AUTRES (PAR COPIER-COLLER BÊTE ET MÉCHANT)

CONSTRUCTION DE VOTRE CORPUS

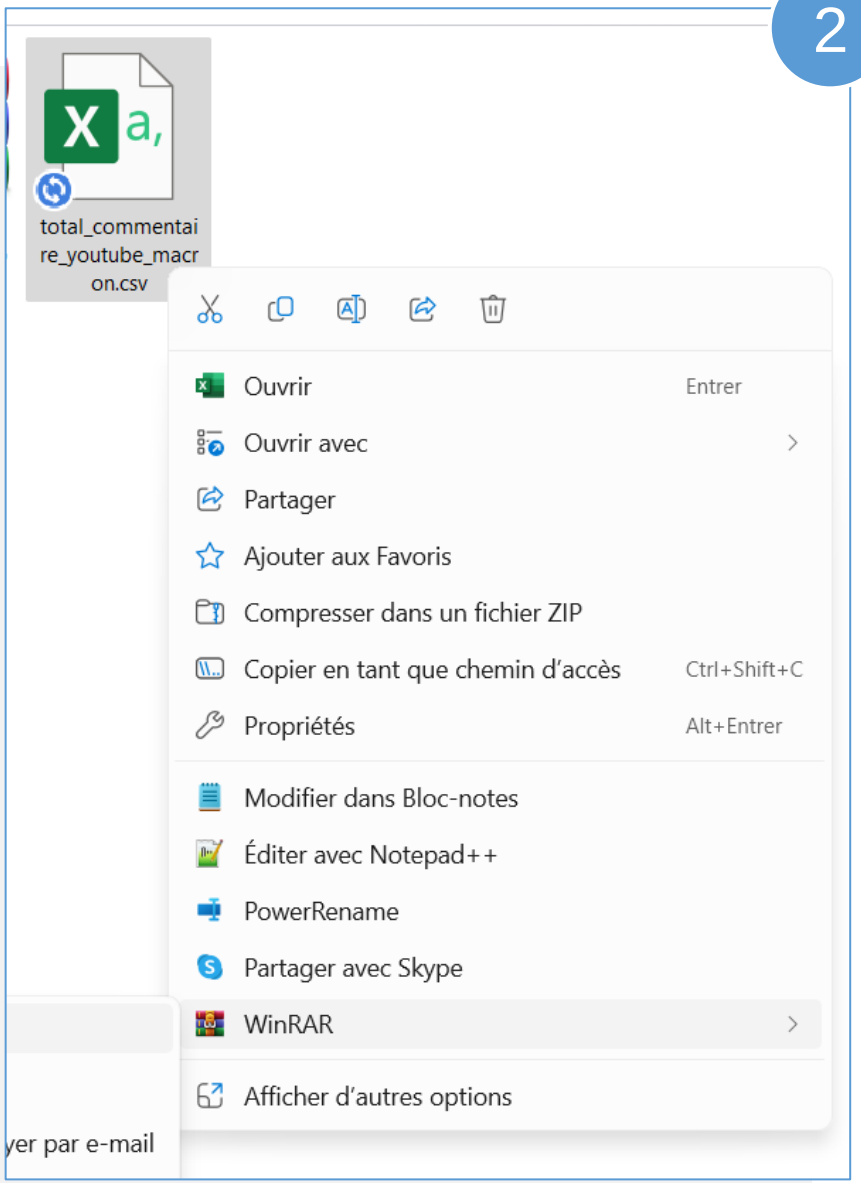
DÉMO DE YOUTUBE DATA TOOL

5. Compresser en format .zip pour import dans CorText

→ Soit vous compressez tous vos fichiers .csv en une seule archive (1)
→ Soit vous compressez seulement votre fichier unifié (2)

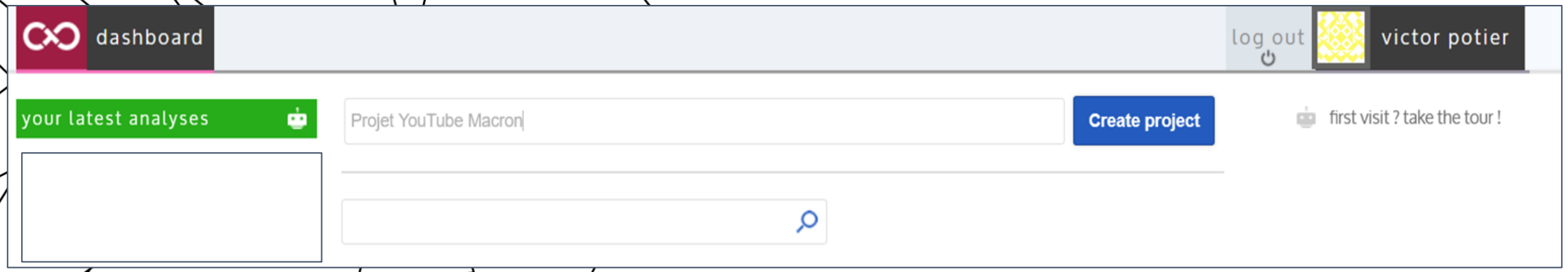


1

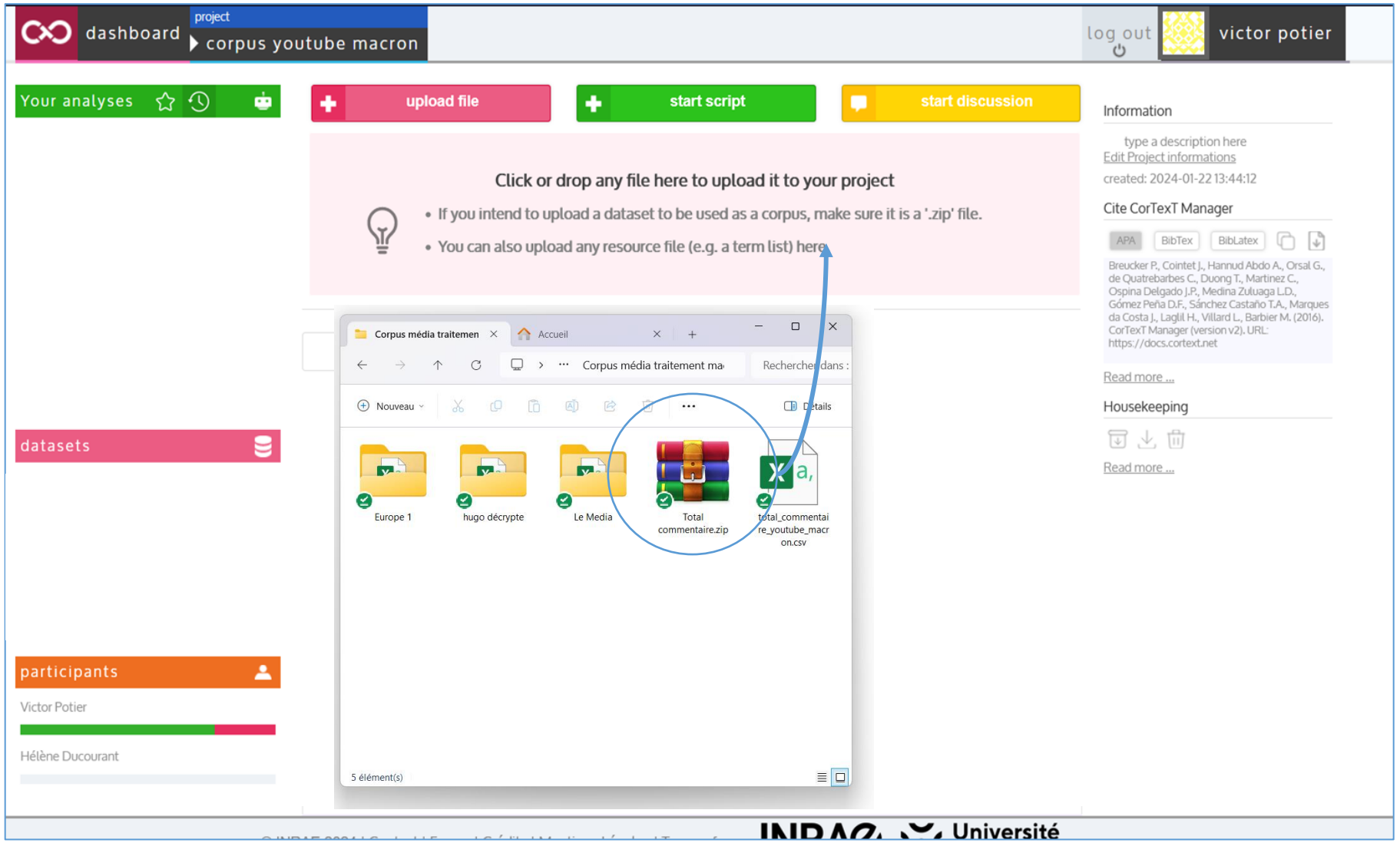


2

CRÉEZ VOTRE PROJET SUR CORTEXT



Donner *un nom* à votre projet



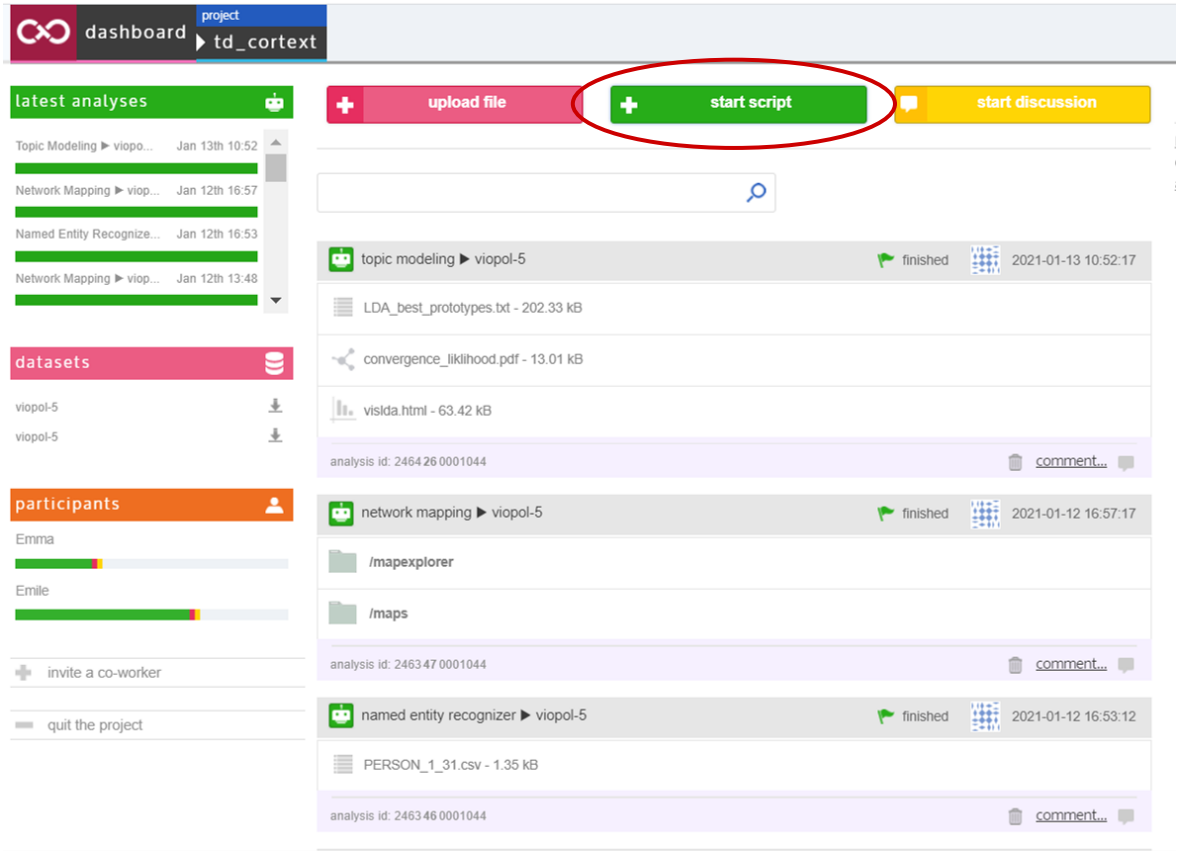
- Pour téléverser notre fichier, cliquez sur “Upload Files” puis sur le rectangle rose qui s’affiche “click here…”
- Sélectionnez votre fichier .zip, fichier qui se trouve dans votre ordinateur
- Votre fichier s’affiche dans votre “board” Cortext

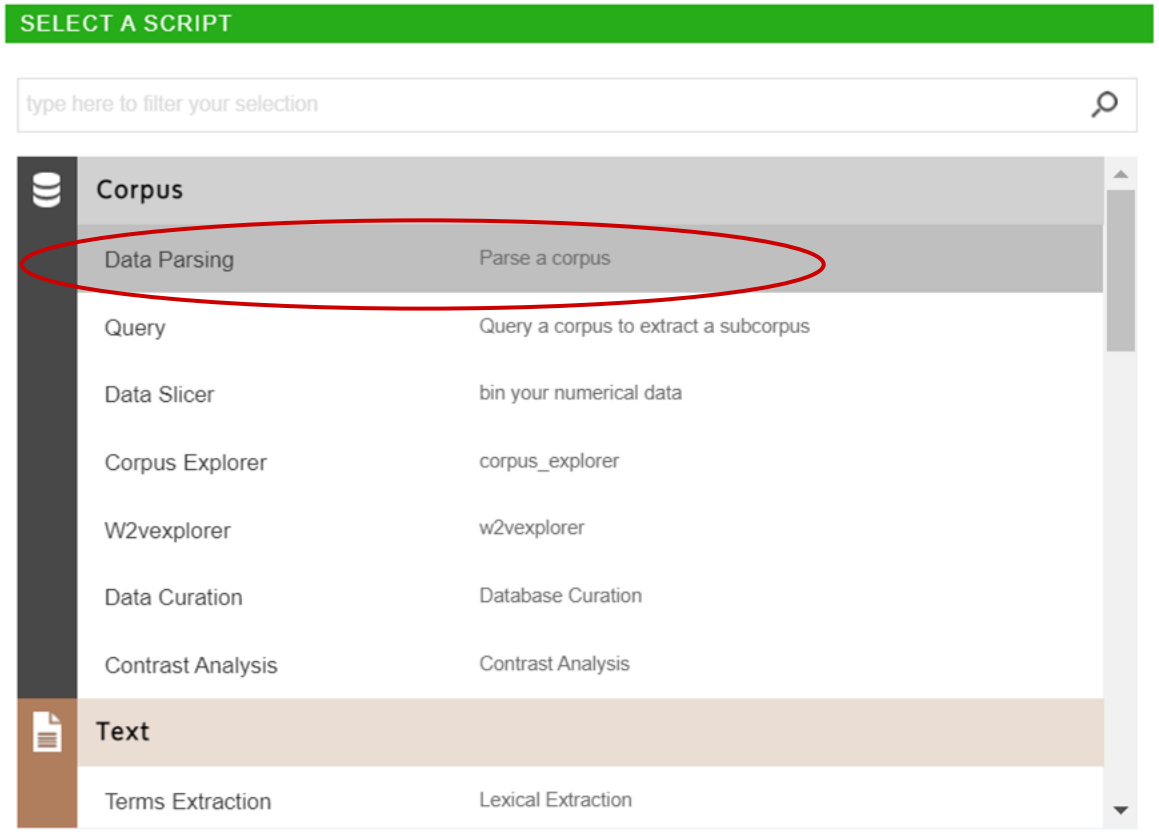
IV. TRAVAIL SUR LES **TERM LIST** : PARSING, INDEXATION, NETTOYAGE ET COMMENTAIRES

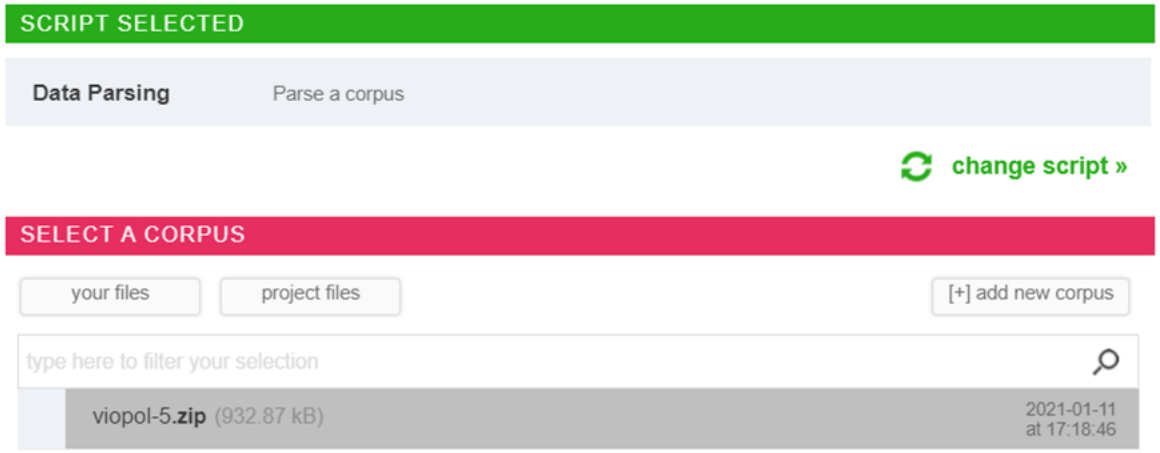
Le premier script : le parsing

Une fois le fichier .zip uploadé sur Cortext, il faut le parser, c’est-à-dire permettre à Cortext de comprendre le découpage des colonnes et des données.

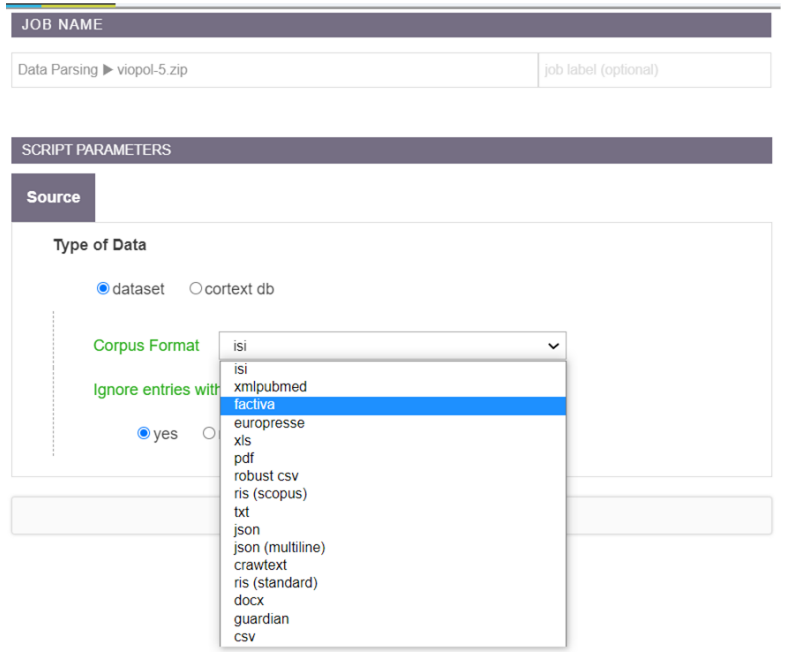
Pour cela, on lance le script “Data Parsing”


dashboard
project
td_cortext
latest analyses
Topic Modeling ▶ viopo... Jan 13th 10:52
Network Mapping ▶ viop... Jan 12th 16:57
Named Entity Recognize... Jan 12th 16:53
Network Mapping ▶ viop... Jan 12th 13:48
datasets
viopol-5
viopol-5
participants
Emma
Emile
invite a co-worker
quit the project
upload file
start script
start discussion
topic modeling ▶ viopol-5
finished
2021-01-13 10:52:17
LDA_best_prototypes.txt - 202.33 kB
convergence_liklihood.pdf - 13.01 kB
visIda.html - 63.42 kB
analysis id: 2464 26 0001044
comment...
network mapping ▶ viopol-5
finished
2021-01-12 16:57:17
/mapexplorer
/maps
analysis id: 2463 47 0001044
comment...
named entity recognizer ▶ viopol-5
finished
2021-01-12 16:53:12
PERSON_1_31.csv - 1.35 kB
analysis id: 2463 46 0001044
comment...


SELECT A SCRIPT
type here to filter your selection
Corpus
Data Parsing
Parse a corpus
Query
Query a corpus to extract a subcorpus
Data Slicer
bin your numerical data
Corpus Explorer
corpus_explorer
W2vexplorer
w2vexplorer
Data Curation
Database Curation
Contrast Analysis
Contrast Analysis
Text
Terms Extraction
Lexical Extraction



Sélectionnez votre fichier.zip que vous voulez parser

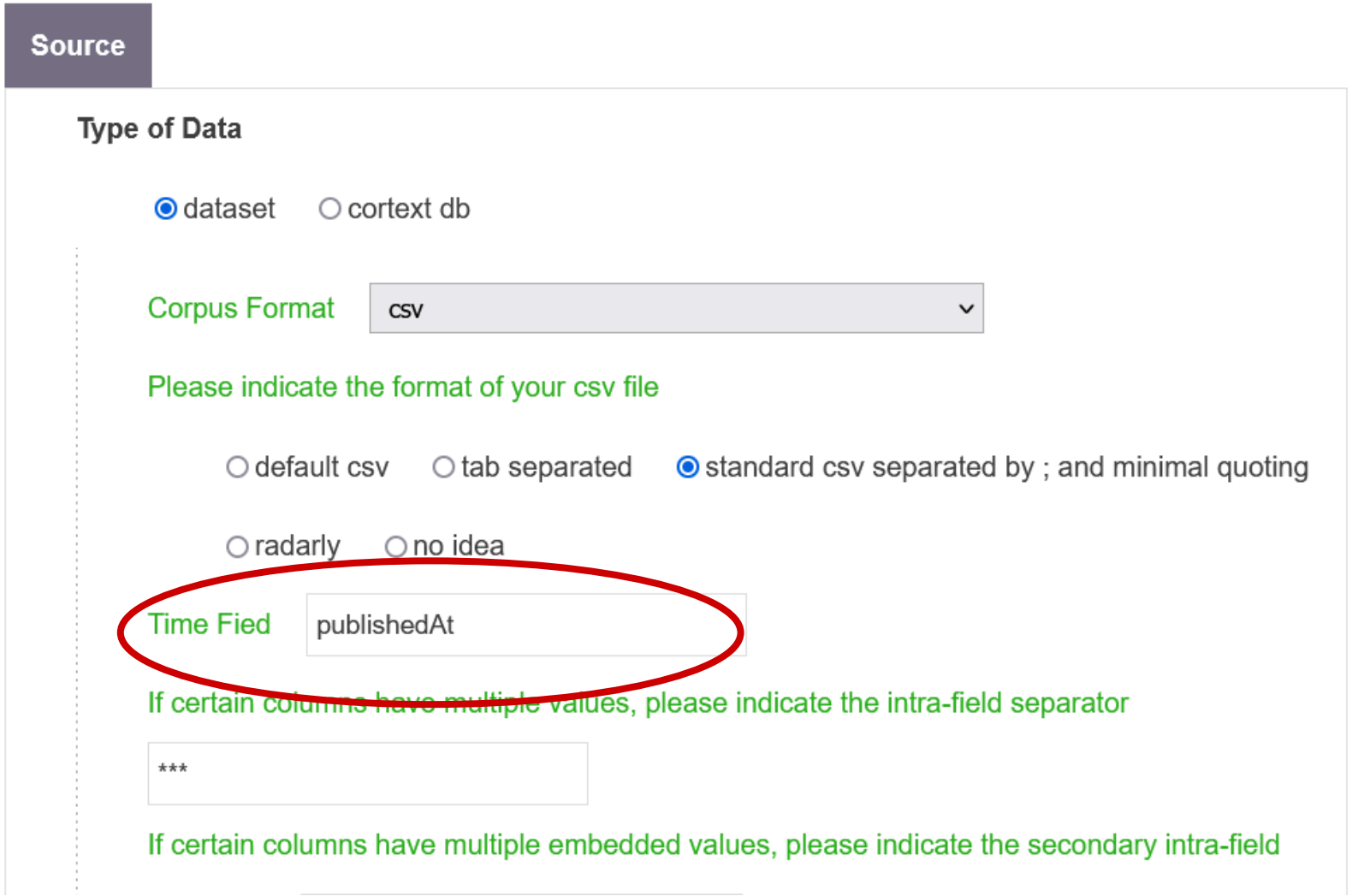


puis sélectionnez le bon format en fonction :

- de votre plateforme de requête (Factiva ou Europresse)

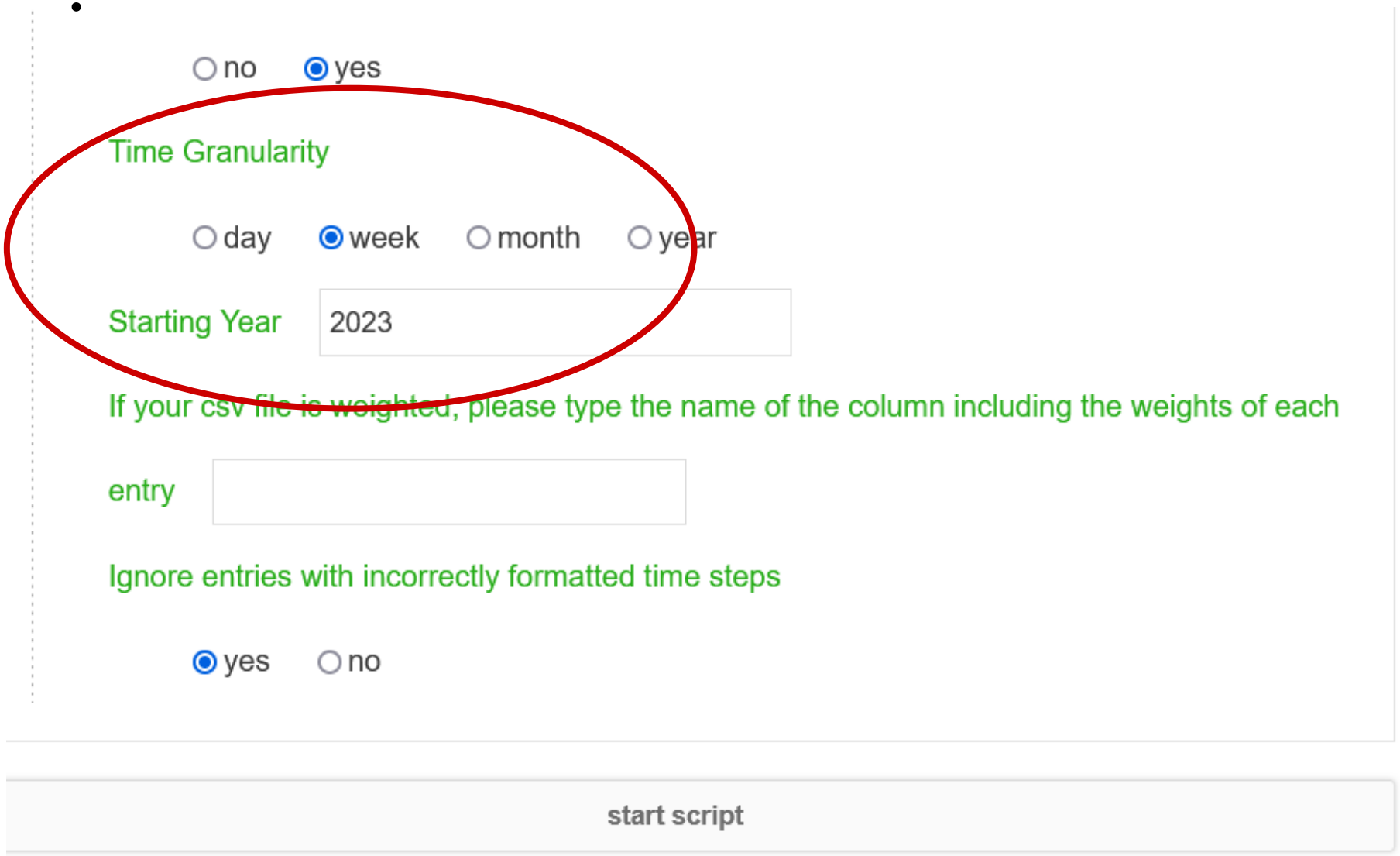
- de votre fichier (CSV dans le cas du corpus YouTube)

ATTENTION À BIEN INDIQUER LA COLONNE DES DATES



Indiquer le nom de la colonne correspondante dans le CSV

ATTENTION À BIEN INDIQUER LA COLONNE DES DATES !



Puis régler la granularité et la date de départ de l'analyse



Votre fichier.zip a été parsé avec succès ! Le drapeau vert l'indique.

Nous allons pouvoir commencer l'exploration de nos données et notre analyse !

“ERROR “ : QUE FAIRE SI “ERROR” S’AFFICHE ?

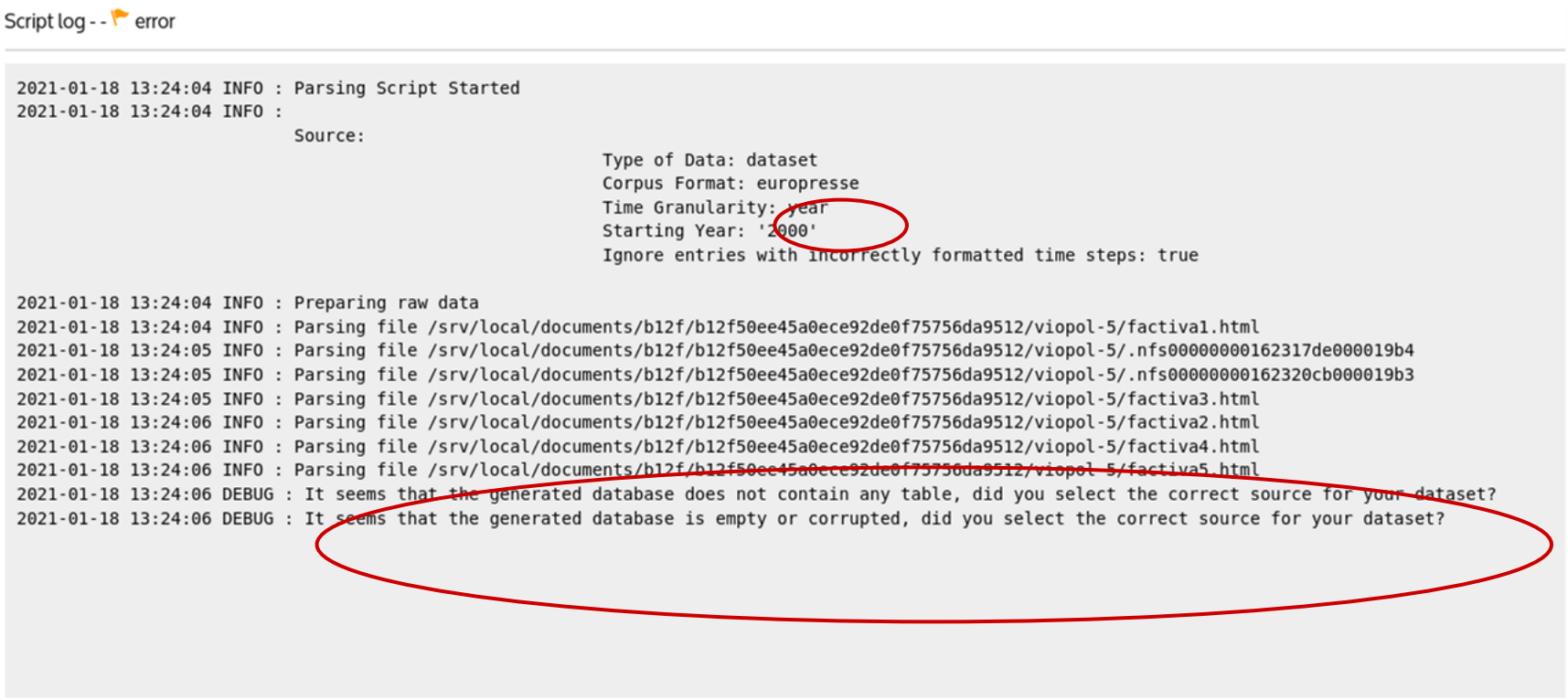
Quand vous n’avez pas le “drapeau vert” qui s’affiche et qu’au contraire, vous obtenez un méchant “error”, il faut aller voir quel est le problème.

> D’une manière générale, utiliser la documentation de Cortext où tous les scripts sont explicités étape par étape

> Lire le “debug log” : cliquer sur “Error”

> Le détail du debuglog s’affiche quand on clique dessus.





En général, à cette étape, si vous rencontrez un problème, c’est 1, car vous avez un accent sur une lettre du nom de fichier 2) parce que votre fichier n’est pas zippé, 3 parce que vous n’avez pas sélectionné dans le menu déroulant le bon élément de description de vos données

EXPLORER SON CORPUS

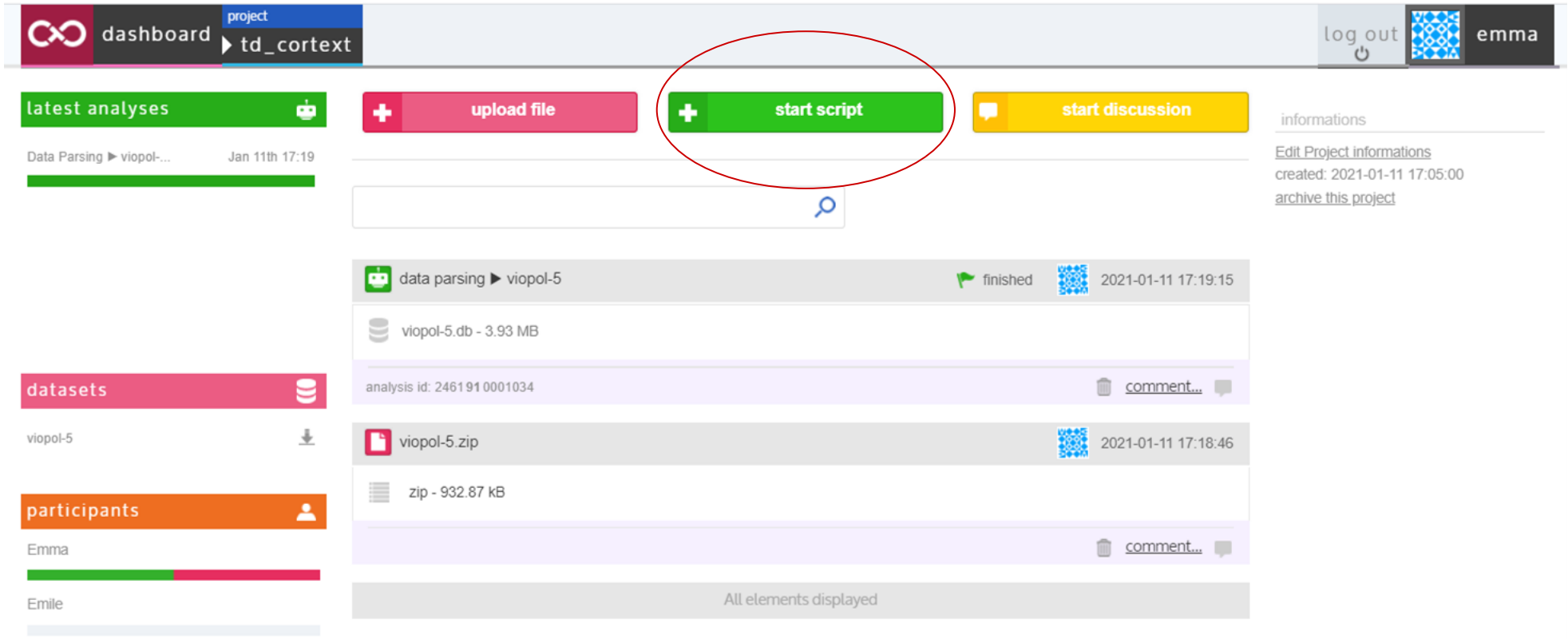
Une fois qu'on dispose d'un corpus "parsé" avec succès, on va pouvoir commencer l'analyse!

Première chose à faire "explorer son corpus" , en gros, cela permet de

- voir à quoi ressemble notre "base de données" ;
- Voir le type de "métadonnées" dont on dispose.

La diapo suivante montre comment faire !

CHOIX DU SCRIPT DANS “START SCRIPT”



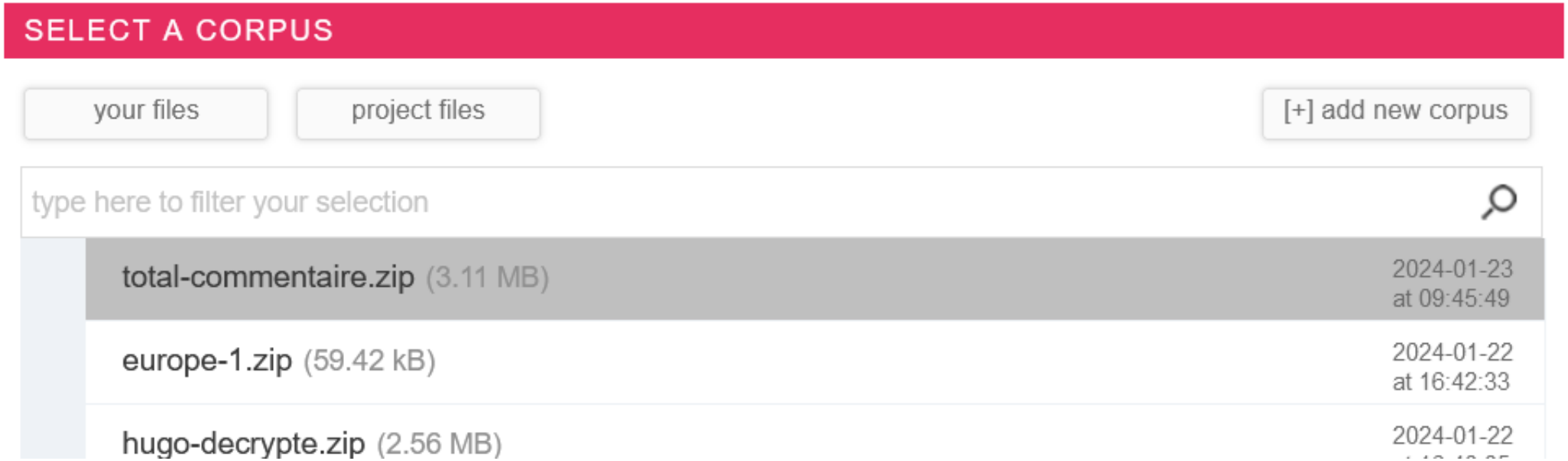
SELECT A SCRIPT

type here to filter your selection

Corpus

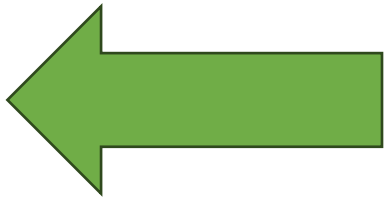
Data Parsing	Parse a corpus
Query	Query a corpus to extract a subcorpus
Data Slicer	bin your numerical data
Corpus Explorer	corpus_explorer
W2vexplorer	w2vexplorer
Data Curation	Database Curation

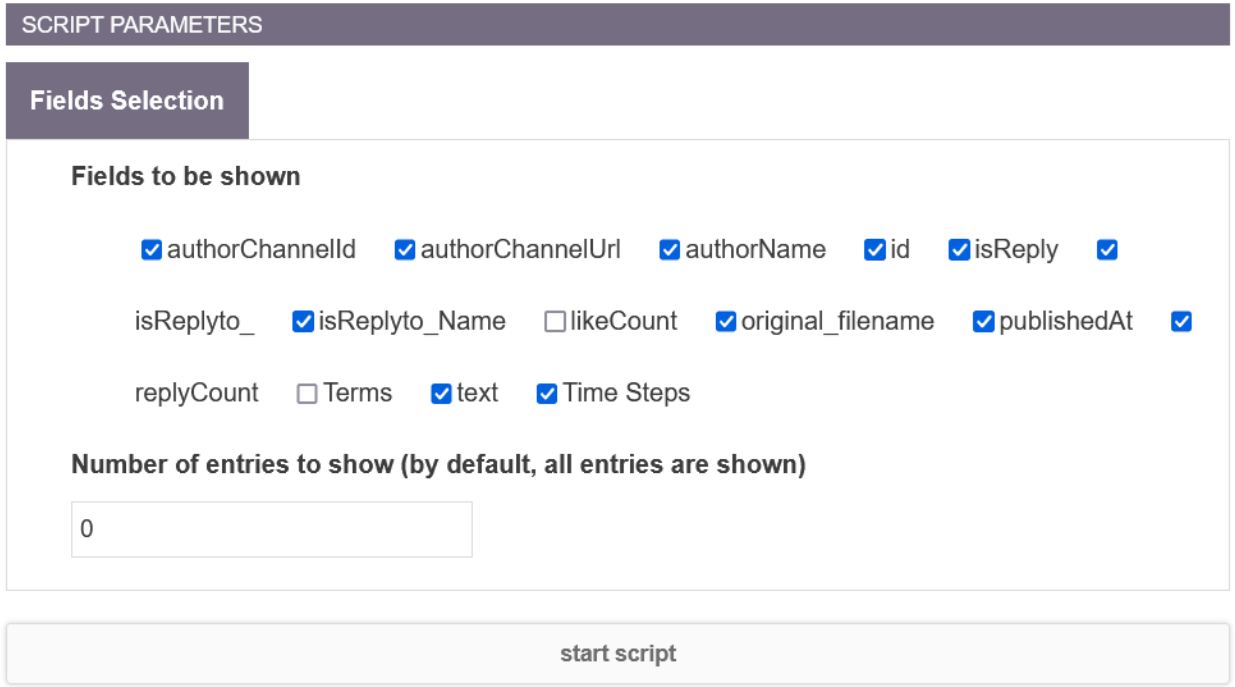
CORTEXT PROPOSE DE “FAIRE TOURNER” LE SCRIPT SUR VOTRE CORPUS.
VOUS DEVEZ SÉLECTIONNER CE DERNIER EN CLIQUANT DESSUS.



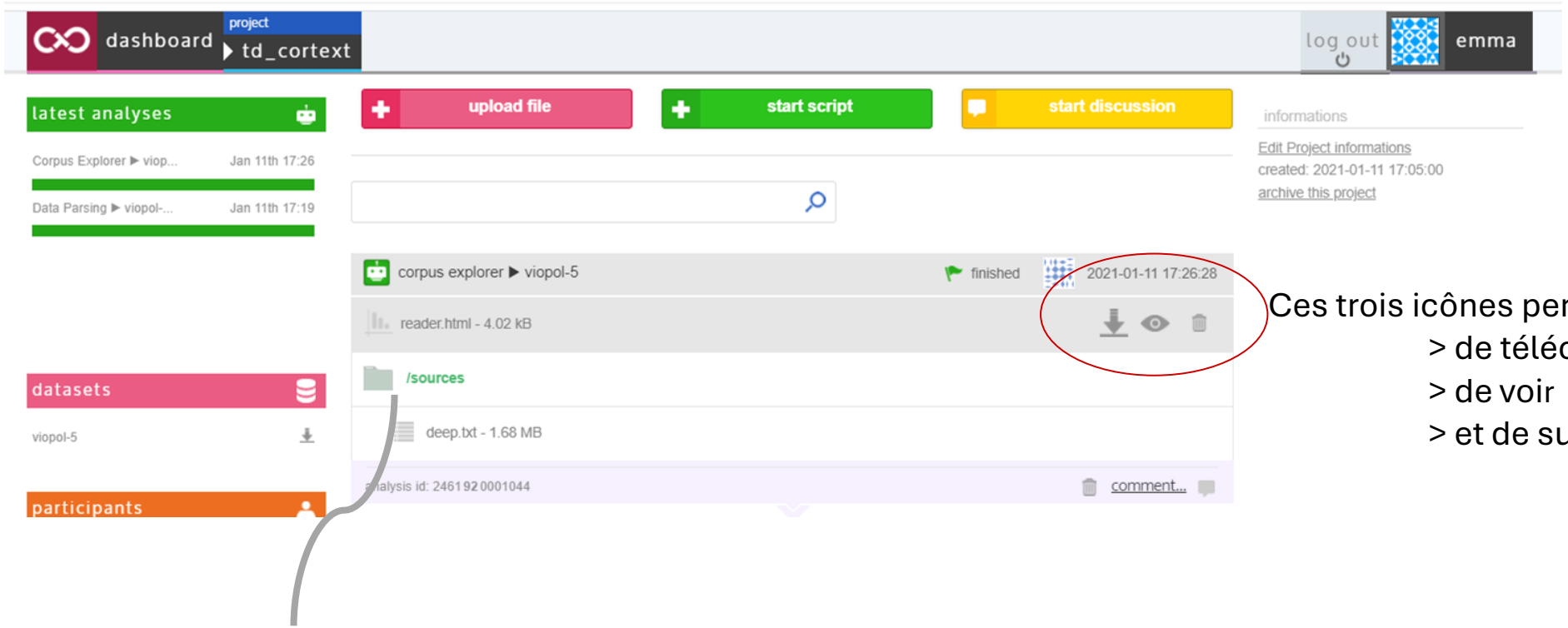
Cliquez ici pour sélectionner votre corpus

Les cases correspondent aux métadonnées dispo. Elles pourront devenir des “variables” de l’analyse. Cochez celles qui vous intéressent (cochez les toutes !)





Une fois que le drapeau vert s'affiche ! c'est bon ! vous pouvez consulter les résultats grâce à l'icône "oeil"

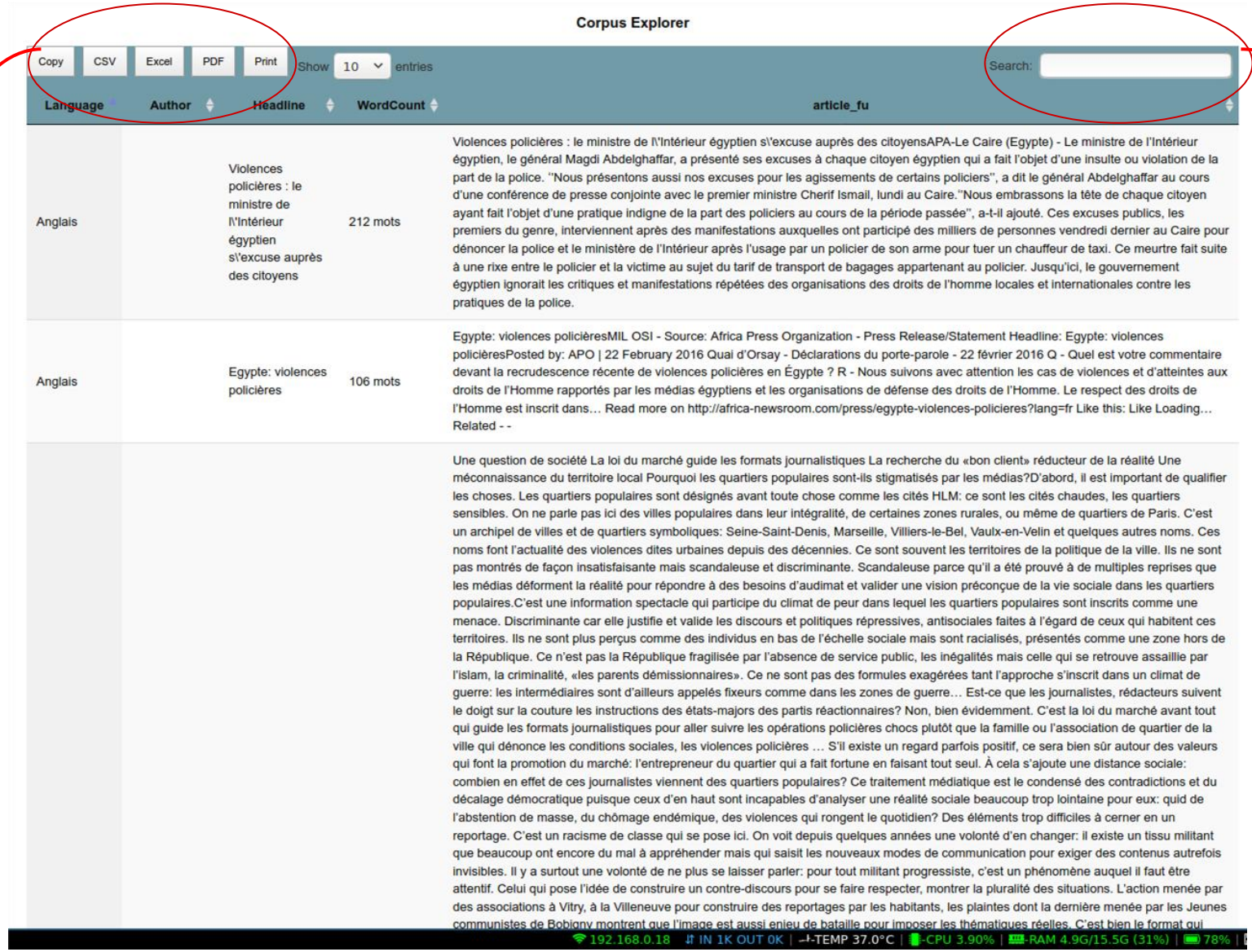


Ces trois icônes permettent
> de télécharger le corpus
> de voir
> et de supprimer le corpus explorer

Ce dossier vous permet d'ouvrir votre corpus dans un format json. Cependant, ce n'est pas utile de le faire pour naviguer dans votre corpus

Et voilà à qui ressemble notre base d'articles sous CorText : un tableau, des colonnes, etc.

Vous permet de télécharger votre corpus (originalement en HTML) en version CSV, Excel, PDF)



Corpus Explorer

Copy | CSV | Excel | PDF | Print | Show 10 entries | Search:

Language	Author	Headline	WordCount	article_fu
Anglais		Violences policières : le ministre de l\'Intérieur égyptien s\'excuse auprès des citoyens	212 mots	Violences policières : le ministre de l\'Intérieur égyptien, le général Magdi Abdelghaffar, a présenté ses excuses à chaque citoyen égyptien qui a fait l'objet d'une insulte ou violation de la part de la police. ''Nous présentons aussi nos excuses pour les agissements de certains policiers'', a dit le général Abdelghaffar au cours d'une conférence de presse conjointe avec le premier ministre Cherif Ismail, lundi au Caire.''Nous embrassons la tête de chaque citoyen ayant fait l'objet d'une pratique indigne de la part des policiers au cours de la période passée'', a-t-il ajouté. Ces excuses publics, les premiers du genre, interviennent après des manifestations auxquelles ont participé des milliers de personnes vendredi dernier au Caire pour dénoncer la police et le ministère de l'Intérieur après l'usage par un policier de son arme pour tuer un chauffeur de taxi. Ce meurtre fait suite à une rixe entre le policier et la victime au sujet du tarif de transport de bagages appartenant au policier. Jusqu'ici, le gouvernement égyptien ignorait les critiques et manifestations répétées des organisations des droits de l'homme locales et internationales contre les pratiques de la police.
Anglais		Egypte: violences policières	106 mots	Egypte: violences policièresMIL OSI - Source: Africa Press Organization - Press Release/Statement Headline: Egypte: violences policièresPosted by: APO | 22 February 2016 Quai d'Orsay - Déclarations du porte-parole - 22 février 2016 Q - Quel est votre commentaire devant la recrudescence récente de violences policières en Égypte ? R - Nous suivons avec attention les cas de violences et d'atteintes aux droits de l'Homme rapportés par les médias égyptiens et les organisations de défense des droits de l'Homme. Le respect des droits de l'Homme est inscrit dans… Read more on http://africa-newsroom.com/press/egypte-violences-policieres?lang=fr Like this: Like Loading… Related - -
				Une question de société La loi du marché guide les formats journalistiques La recherche du «bon client» réducteur de la réalité Une méconnaissance du territoire local Pourquoi les quartiers populaires sont-ils stigmatisés par les médias?D'abord, il est important de qualifier les choses. Les quartiers populaires sont désignés avant toute chose comme les cités HLM: ce sont les cités chaudes, les quartiers sensibles. On ne parle pas ici des villes populaires dans leur intégralité, de certaines zones rurales, ou même de quartiers de Paris. C'est un archipel de villes et de quartiers symboliques: Seine-Saint-Denis, Marseille, Villiers-le-Bel, Vaulx-en-Velin et quelques autres noms. Ces noms font l'actualité des violences dites urbaines depuis des décennies. Ce sont souvent les territoires de la politique de la ville. Ils ne sont pas montrés de façon insatisfaisante mais scandaleuse et discriminante. Scandaleuse parce qu'il a été prouvé à de multiples reprises que les médias déforment la réalité pour répondre à des besoins d'audimat et valider une vision préconçue de la vie sociale dans les quartiers populaires.C'est une information spectacle qui participe du climat de peur dans lequel les quartiers populaires sont inscrits comme une menace. Discriminante car elle justifie et valide les discours et politiques répressives, antisociales faites à l'égard de ceux qui habitent ces territoires. Ils ne sont plus perçus comme des individus en bas de l'échelle sociale mais sont racialisés, présentés comme une zone hors de la République. Ce n'est pas la République fragilisée par l'absence de service public, les inégalités mais celle qui se retrouve assaillie par l'islam, la criminalité, «les parents démissionnaires». Ce ne sont pas des formules exagérées tant l'approche s'inscrit dans un climat de guerre: les intermédiaires sont d'ailleurs appelés fixeurs comme dans les zones de guerre… Est-ce que les journalistes, rédacteurs suivent le doigt sur la couture les instructions des états-majors des partis réactionnaires? Non, bien évidemment. C'est la loi du marché avant tout qui guide les formats journalistiques pour aller suivre les opérations policières chocs plutôt que la famille ou l'association de quartier de la ville qui dénonce les conditions sociales, les violences policières … S'il existe un regard parfois positif, ce sera bien sûr autour des valeurs qui font la promotion du marché: l'entrepreneur du quartier qui a fait fortune en faisant tout seul. À cela s'ajoute une distance sociale: combien en effet de ces journalistes viennent des quartiers populaires? Ce traitement médiatique est le condensé des contradictions et du décalage démocratique puisque ceux d'en haut sont incapables d'analyser une réalité sociale beaucoup trop lointaine pour eux: quid de l'abstention de masse, du chômage endémique, des violences qui rongent le quotidien? Des éléments trop difficiles à cerner en un reportage. C'est un racisme de classe qui se pose ici. On voit depuis quelques années une volonté d'en changer: il existe un tissu militant que beaucoup ont encore du mal à appréhender mais qui saisit les nouveaux modes de communication pour exiger des contenus autrefois invisibles. Il y a surtout une volonté de ne plus se laisser parler: pour tout militant progressiste, c'est un phénomène auquel il faut être attentif. Celui qui pose l'idée de construire un contre-discours pour se faire respecter, montrer la pluralité des situations. L'action menée par des associations à Vitry, à la Villeneuve pour construire des reportages par les habitants, les plaintes dont la dernière menée par les Jeunes communistes de Bobigny montrent que l'image est aussi enjeu de bataille pour imposer les thématiques réelles. C'est bien le format qui

192.168.0.18 | IN 1K OUT 0K | TEMP 37.0°C | CPU 3.90% | RAM 4.9G/15.5G (31%) | 78%

Vous permet de faire une recherche à l'intérieur de votre corpus

POUR INFO, AU DESSUS DU TABLEAU, VOUS POUVEZ…

- télécharger le jeu de données sous plusieurs formats
- afficher un nombre de lignes spécifique sur une page
- filtrer les résultats à afficher à partir d’un mot clé

Nettoyer un corpus, ici pas nécessaire, mais il faut le savoir quand même

→ On peut avoir à travailler avec d'autres corpus de qualité aléatoire et savoir nettoyer un corpus peut être utile…

Pourquoi nettoyer un corpus ?

→ L'intégralité du corpus récupéré est rarement utile, pour des raisons d'efficacité et de temps de traitement de données sous Cortext, il peut être judicieux de **supprimer les variables inutiles** (des colonnes) qui polluent la visibilité des variables centrales.

→ Au cours de l'analyse, on peut réaliser que l'on souhaite se concentrer sur certaines caractéristiques (Analyse par date, par chaîne, par thème, etc.). On peut alors **créer des sous-corpus plus petits** en fonction des variables qui nous intéressent.

NETTOYAGE DE DONNÉES (SCRIPT “DATA CURATION”)

> Ce script Data Curation permet de :

> Supprimer une variable

> Renommer une variable (éviter de mettre de caractères spéciaux (les accents è,é, *, +,)

> Concaténer des variables

> Convertir des variables en un autre type (integer, time, string length)

> Séparer une variable en plusieurs variables

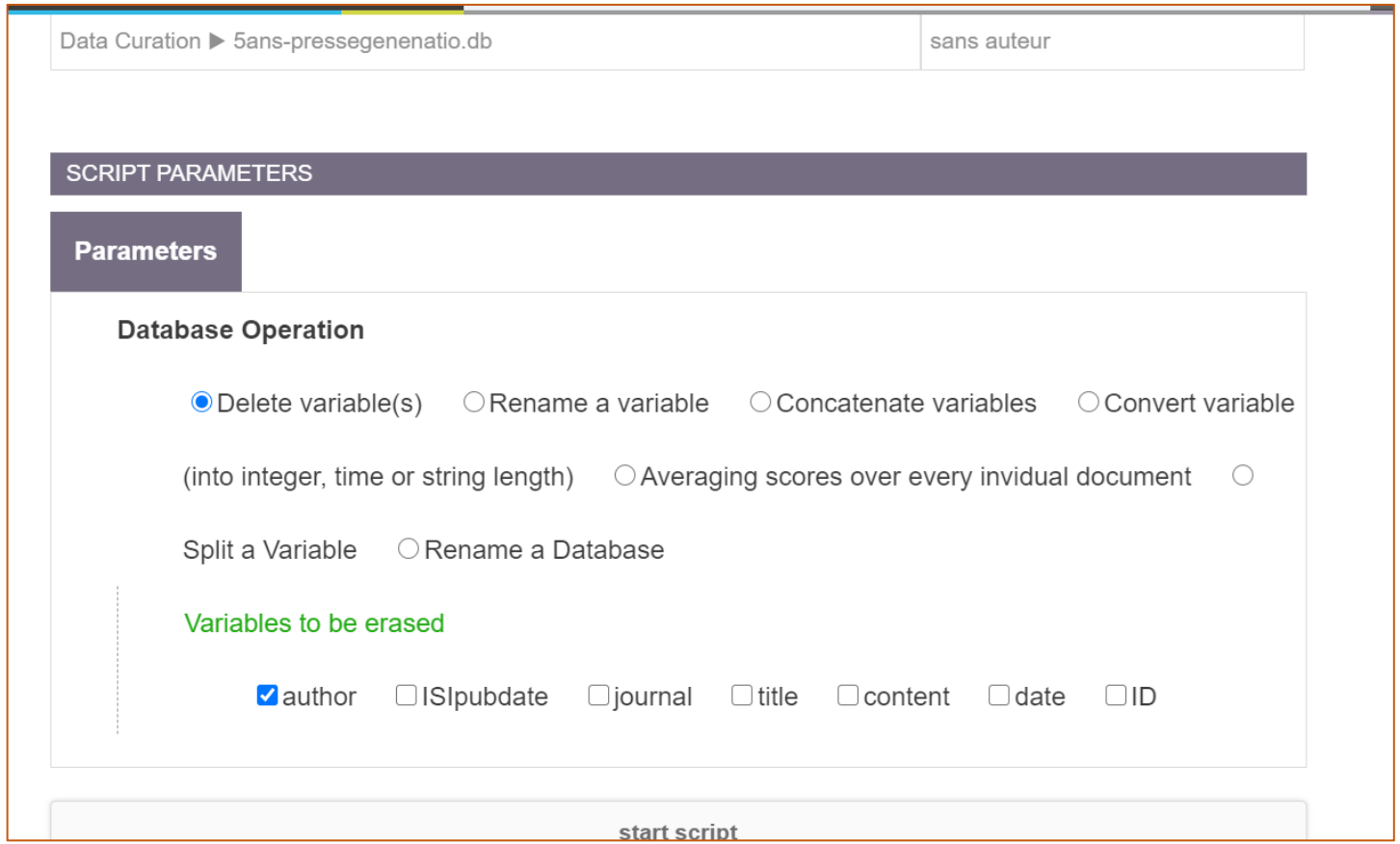
> Renommer le jeu de données

SUPPRIMER UNE VARIABLE

Vous voulez supprimer la variable “auteur”, car vous savez qu’elle ne vous intéresse pas.
“delete” (supprimer en anglais)
“Author’.
“start script”

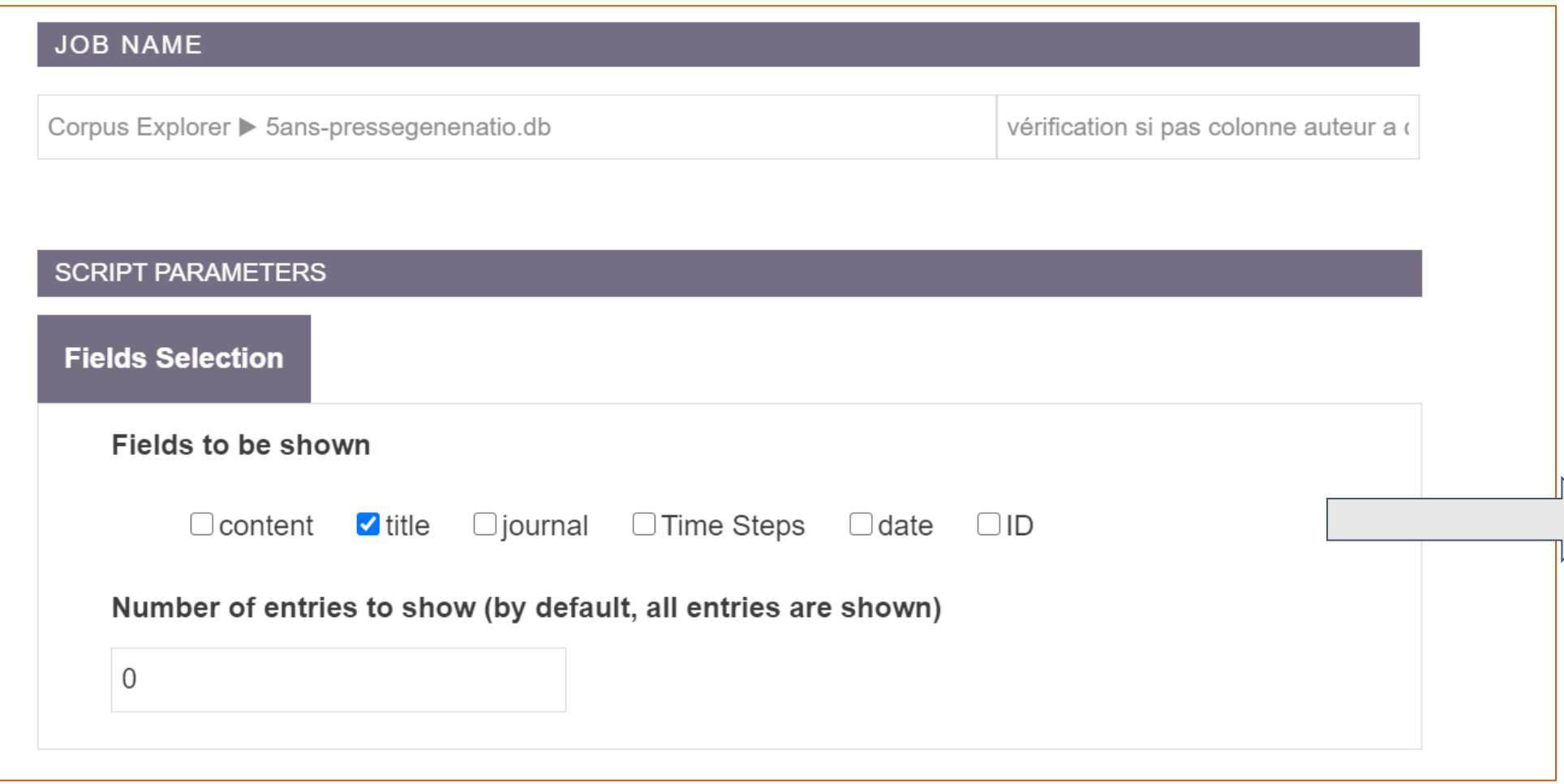
--
Comment faire ensuite pour voir si la suppression a bien eu lieu ?

(réponse diapo suivante !)



FAITES À NOUVEAU UNE EXPLORATION DE CORPUS, LANCEZ LE SCRIPT “ CORPUS EXPLORATION” !

écrire dans “job name” ce que vous êtes en train de faire



Dès cet encadré, on voit que notre variable auteur a disparu et si on lance le script d’exploration, on le voit…

EST-IL POSSIBLE D'ANNULER UN NETTOYAGE ?

Ex. : vous venez de supprimer une variable et vous avez un regret. Comment faire ?

- L'opération est définitive
- pour "revenir en arrière", il faut tout recommencer : uploader votre fichier d'origine, "parsez"-le.
- Pour que vous votre tableau de bord *(dashboard)* "reste propre", supprimez vos scripts précédents en cliquant sur le bouton poubelle puis en acceptant qu'ils soient définitivement supprimés.

NB : Vous ne vous y retrouvez plus sur l'interface cortex ? Cliquez en haut à gauche sur le 8 inversé ou sur le nom de votre projet (selon la situation), vous retrouvez accès à votre "tableau de bord" et aux boutons de scripts, etc.

ANALYSE DES TERMES

Nous allons chercher à connaître les expressions les plus fréquemment utilisées dans les articles de notre corpus.

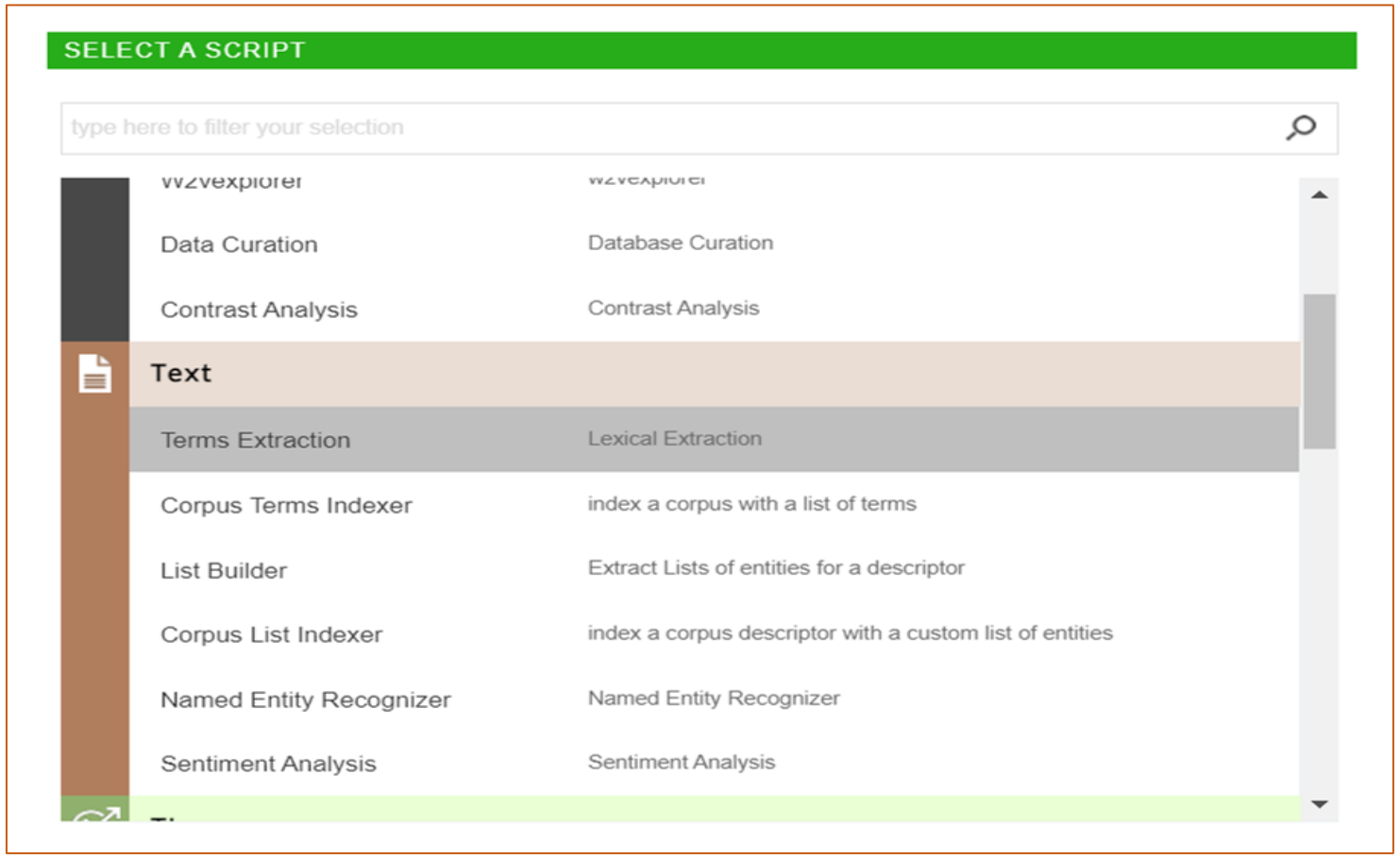
Les “**occurrences**” les plus fréquentes.

LES GROUPES DE MOTS LES PLUS FRÉQUENTS…

- L’objectif de l’analyse textuelle est de construire des analyses autour des mots employés dans les textes d’un corpus… Allons-y !
- À l’aide de cortex, nous allons chercher à connaître les mots les plus cités, à avoir la liste des “occurrences” les plus fréquentes.
- En dehors de CorText, vous trouvez l’habitude de voir des visualisations des occurrences fréquentes un peu partout.

QUEL SCRIPT POUR OBTENIR LA LISTE DES EXPRESSIONS/GROUPES DE DEUX OU TROIS MOTS LES PLUS CITÉES ?

→ “**TERM EXTRACTIONS**”



NB : Nous ne voulons pas obtenir un “nuage de mots” (i.e. représentation visuelle), mais une liste (i.e. une feuille de tableur)

SCRIPT “TERMS EXTRACTION” :

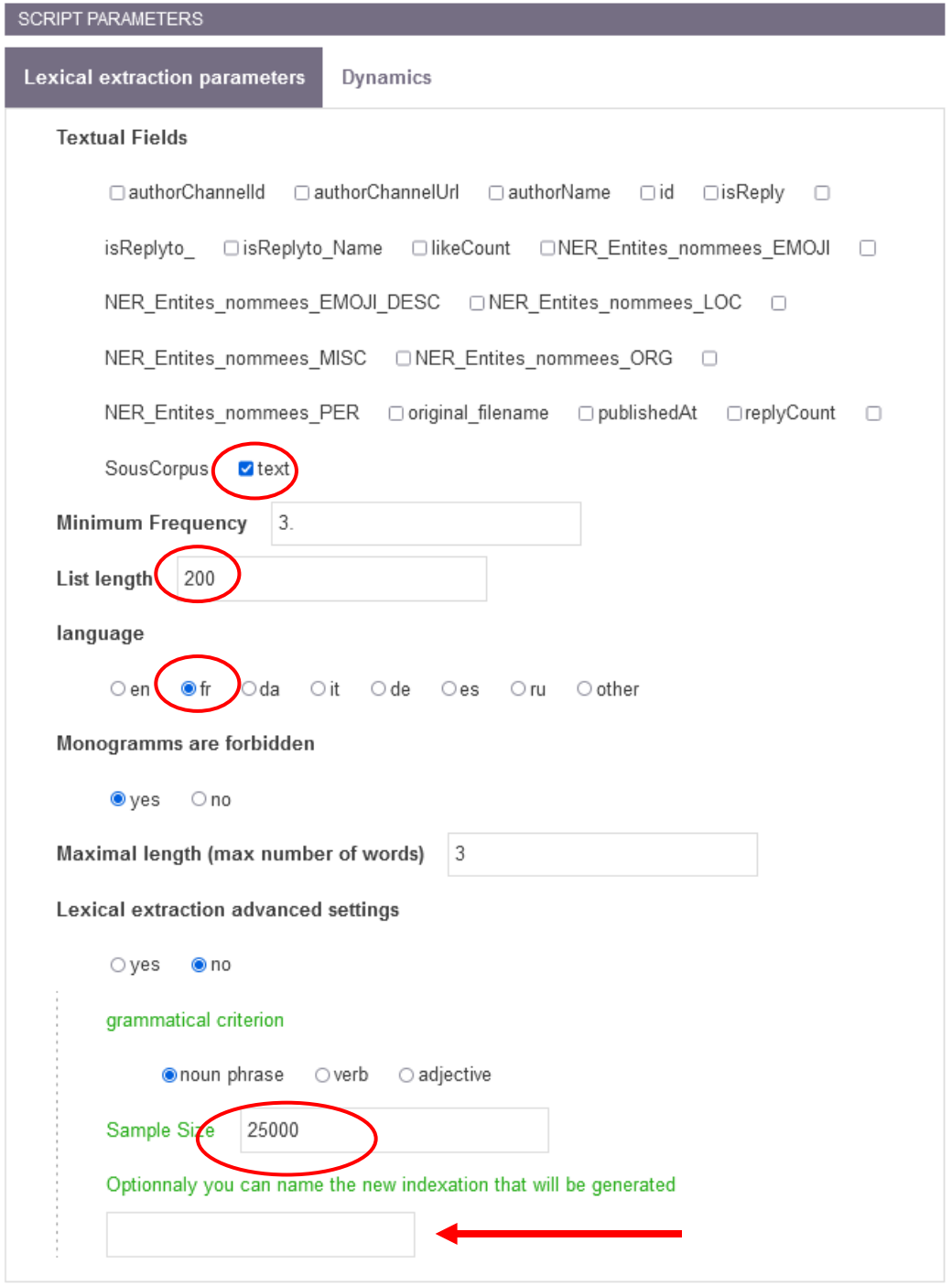
LES MOTS/EXPRESSION LES PLUS CITÉS

On précise les champs suivants : on veut le contenu des commentaires (text), la liste des 200 mots les plus fréquents (lengh list=200), en français (fr).

On ne change pas les paramètres par défaut, ce qui est déjà pré-défini

Ex monograms forbidden, noun phrase, max lengh 3 (cf diapo suivante)

Et à la fin, tout en bas, je vous invite à **nommer votre action** dans le champ prévu . Par exemple “premiere liste 200 mots comm YouTube macron”.



LES “PARAMÈTRES PAR DÉFAUT” : QUELQUES EXPLICATIONS

- monograms forbidden : interdit les mots uniques : on prend que **des groupes de mots, pas de mots seuls**.
- max length 3 : groupe de 3 mots max dans le groupe de mots/expressions (ex Emmanuel Macron : 2 mots. Faute aux jeux vidéo : 3 mots
- Attention : Le lancement d’un script cortex peut prendre plusieurs minutes (10 min !)

QUAND LE SCRIPT EST FINI, ON CLIQUE SUR L'ŒIL ET ON OBTIENT CECI

dashboard | project: corpus youtube macron | log out | victor potier

Editing: extracted_terms_list_top200_Cval3.0_chi2_samplingTrue_periods1.tsv - 26.7 kB

Edit the file as a simple spreadsheet. Click on column headers to sort the whole table, and resize column by dragging its border. You can also right-click on a cell to add or delete a row. When you are done editing the file, simply click save and the document will be saved in its current state with the custom name you wish.

New name (if you want to create a new file) extracted_terms_list_top200_Cval3.0_chi2_samplingTrue_periods1.tsv **Save**

	Stem	Main form	Forms	n	C-value	Gfidf	Specificity chi2	Occurrences	Cooccur
1	Stem	Main form	Forms	n	C-value	Gfidf	Specificity chi2	Occurrences	Cooccurre
2	etat unis	Etats Unis	Etats Unis	2.00	33.05	1.05	7715.86	20.00	
3	url-remplac 👉	👉 url-remplacée	👉 url-remplacée	2.00	23.60	1.00	11934.22	15.00	
4	anné prochain	année prochaine	année prochaine|&|prochaines années	2.00	17.31	1.00	5963.71	11.00	
5	genr person	genre de personnes	genre de personnes|&|personne et du genre|&|genre de personne|&|personnes dans ton genre	3.00	17.31	1.00	5183.87	11.00	
6	publiqu servic	services publiques	services publiques|&|service publique	2.00	28.33	1.00	5197.75	18.00	
7	guerr ukrain	guerre en Ukraine	guerre en Ukraine|&|guerre en ukraine|&|guerre d'Ukraine|&|guerre pour l'Ukraine|&|Ukraine en guerre|&|guerre entre l'Ukraine|&|guerre contre l'Ukraine|&|guerre de Ukraine|&|ukraine en guerre	3.00	89.70	1.04	6180.35	55.00	
8	climat urgenc	urgence climatique	urgence climatique	2.00	18.88	1.00	7550.45	12.00	
9	actuel heur	heure actuelle	heure actuelle	2.00	20.46	1.00	5885.84	13.00	
10	central nucléair	centrales nucléaires	centrales nucléaires|&|centrale nucléaire	2.00	34.62	1.05	5638.26	21.00	
11	hugo équip	équipe Hugo	équipe Hugo|&|Hugo et son équipe|&|Hugo et toute l'équipe|&|Hugo et l'équipe|&|Hugo et à toute l'équipe|&|Hugo et ton équipe|&|Hugo et à ton équipe|&|Hugo et toute son équipe|&|équipe d'Hugo|&|équipe et à toi Hugo|&|Hugo et toute ton équipe|&|Hugo et tes équipe|&|Hugo et votre équipe|&|Hugo et à toute son équipe|&|Hugo et à l'équipe|&|Hugo et toutes l'équipe|&|Hugo et à toute ton équipe	2.00	73.96	1.00	11223.91	47.00	
12	interview président	interview avec le président	interview avec le président|&|interview du président|&|interview d'un président|&|interview le président|&|interview de vous avec le président|&|interview de président|&|président pour cet interview|&|interview un président|&|interview de notre président	4.00	28.33	1.00	7514.46	18.00	
13	opinion publiqu	opinion publique	opinion publique	2.00	18.88	1.00	4203.07	12.00	
14	bel interview	belle interview	belle interview	2.00	18.88	1.00	9234.71	12.00	
15	et le	ET LES	ET LES	2.00	20.46	1.00	6079.96	13.00	
16	argent public	argent public	argent public	2.00	18.88	1.00	7892.92	12.00	

FOCUS SUR LES NOMS DES COLONNES : STEM, MAIN FORM,

. Stem (trad “tige”), **main form** (trad. “forme principale”) et une troisième colonne qui nous donne un aperçu de la façon dont travail CorText (regroupement de termes proches : exemples quartiers populaires quartier populaire une seule et même ligne).

. Une colonne dénombre **“occurrences”**

> Cliquez sur une colonne pour classer les terms par ordre croissant ou décroissant

> Si un terme ne vous pas pertinent, cliquez droit et supprimez-la ligne (**remove row – supprimer ligne**)

. NB : N’oubliez pas d’enregistrer si vous changez des choses dans la liste : save as… votre liste, vous pouvez la renommer avec un nom plus parlant pour vous. Nous en aurons besoin

TERMS EXTRACTIONS

Vous avez réussi ? Vous venez de faire une étape importante.

> En soi, cette liste est un résultat, il faut prendre le temps de la lire, de se laisser surprendre par les termes, de s'interroger sur nos surprises, etc. Il faut se demander le "type de groupes de mots qui reviennent" (Ex. des noms de personnes, de lieux, des thèmes, etc.)

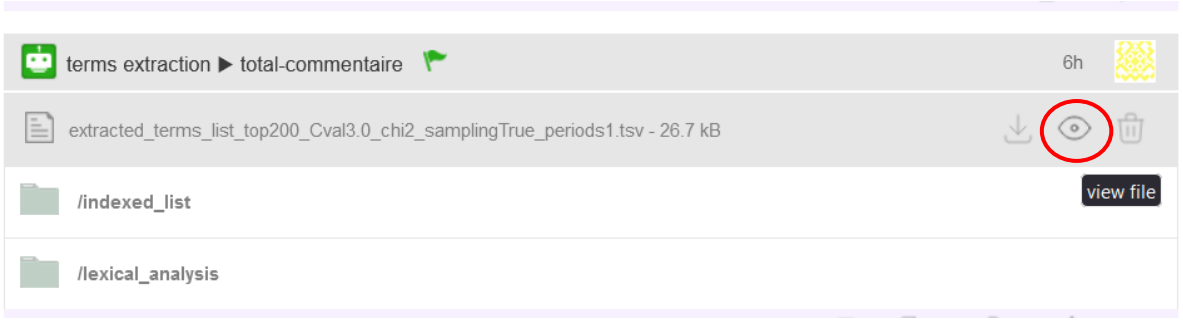
> En soi, cette liste **mérite donc d'être commentée**.

> Mais cette liste est aussi une **étape préparatoire** à d'autres scripts (indexation et network mapping)...

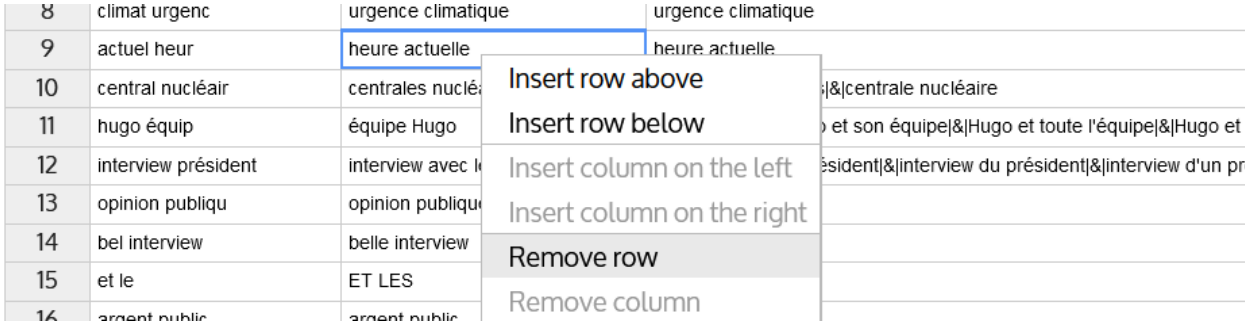
Produire un premier commentaire de la liste

- **Identifier des sous-catégories** : des vocabulaires très marqués dans le militantisme de gauche, des références religieuses, des marques d'interactions entre la communauté et le vidéaste (« Salut Hugo », « Merci Hugo ! », etc.) qui sont propres au format, etc. = cela amène déjà à la formulation de *précatégories*, à prendre en notes à côté.
- On remarque aussi que certaines choses n'ont rien à faire là ou sont en double (« c'est ça », « de toute façon » / « jeu vidéos, jeux videos, jeu videos, jeux vidéo », etc.) = demande à ***nettoyer la liste de termes***

NETTOYER LA LISTE : DEUX OPTIONS



On clique sur « view file » – icône en forme d'œil



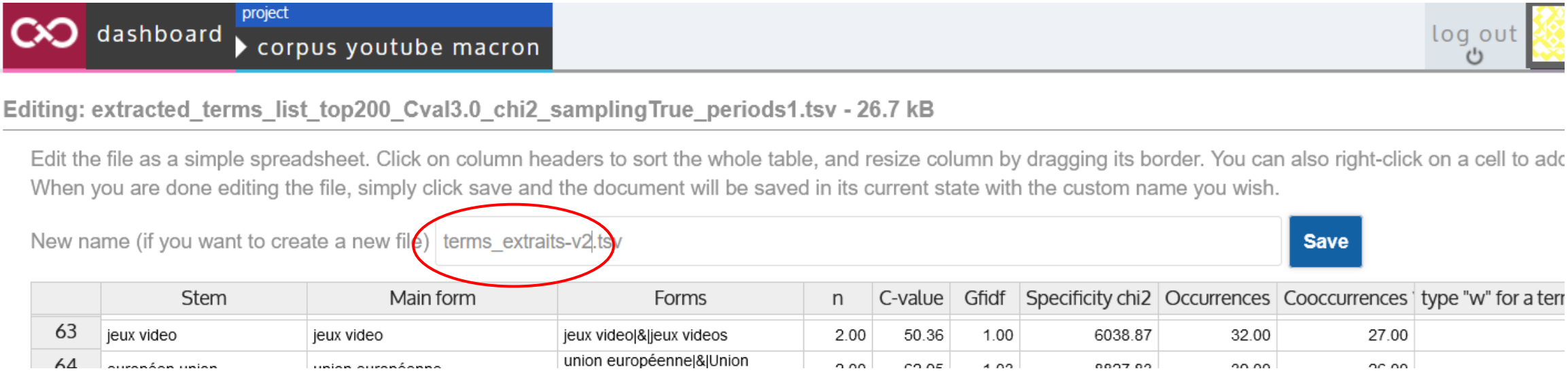
Soit on supprime une ligne inintéressante – « remove row »

Soit on agrège deux lignes similaires en double-cliquant sur la case de la colonne « Forms » dans la première, en ajoutant |&| et en copiant le contenu de la case de la 2nde ligne dans la première… sans oublier d'effacer la 2nde ensuite

	Stem	Main form	Forms
63	jeux video	jeux video	jeux video
64	jeux videos	jeux videos	jeux videos
65			union européennel&|Union

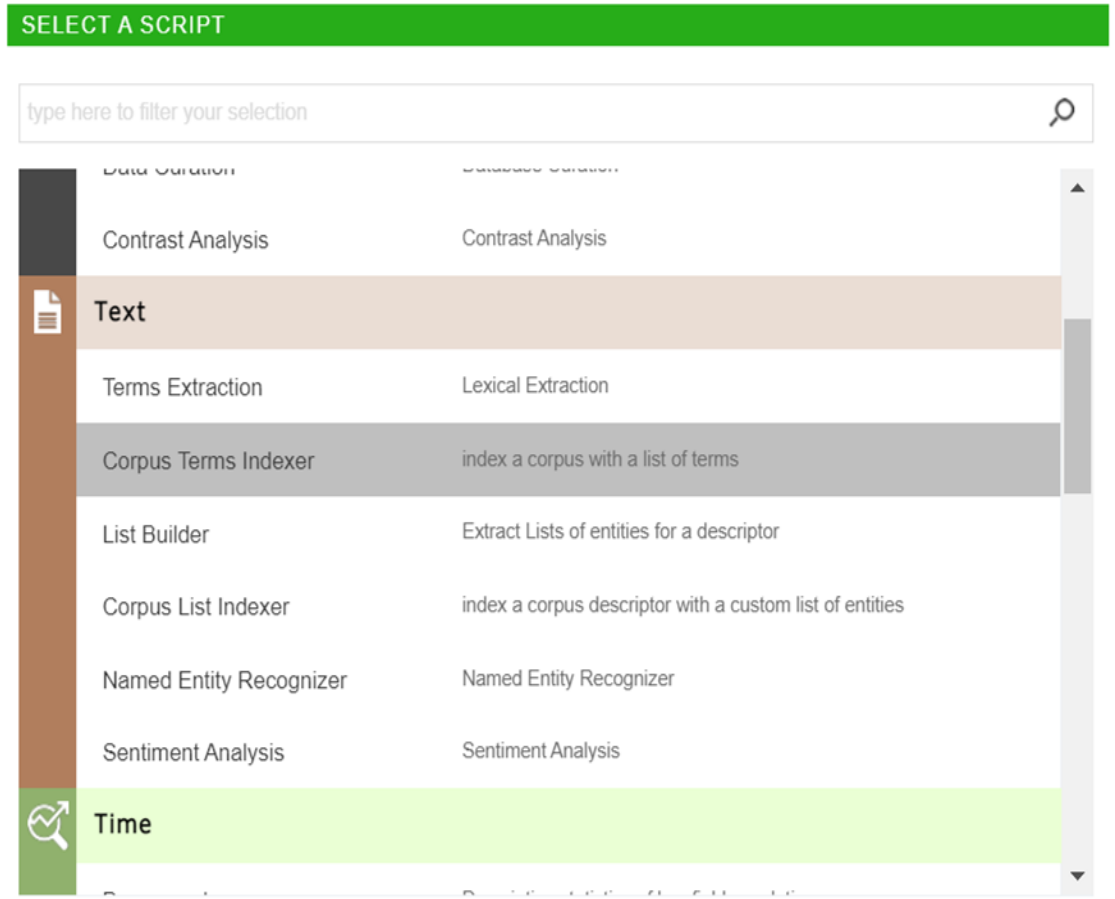
	Stem	Main form	Forms
63	jeux video	jeux video	jeux video|&|jeux videos
64	européen union	union européenne	union européennel&|Union européenne|&|Union Européenne

TRÈS IMPORTANT, RENOMMEZ VOTRE NOUVELLE LISTE NETTOYÉE



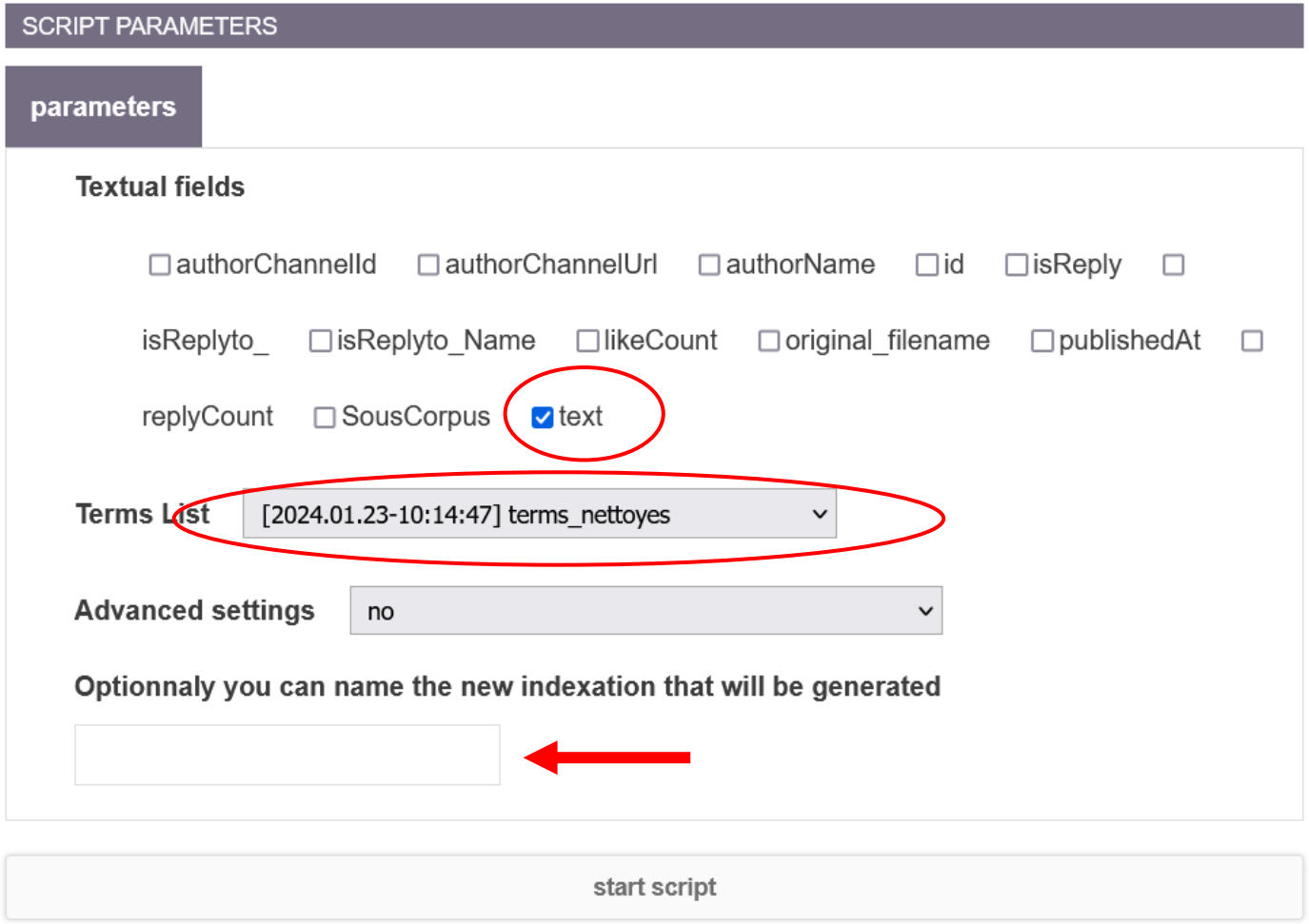
→ Puis n'oubliez pas de cliquer sur « ***Save*** »

INDEXER UNE LISTE À UN CORPUS (APRÈS AVOIR MODIFIÉ VOTRE LISTE INITIALE)



On utilise le script “*Corpus Terms Indexer*”

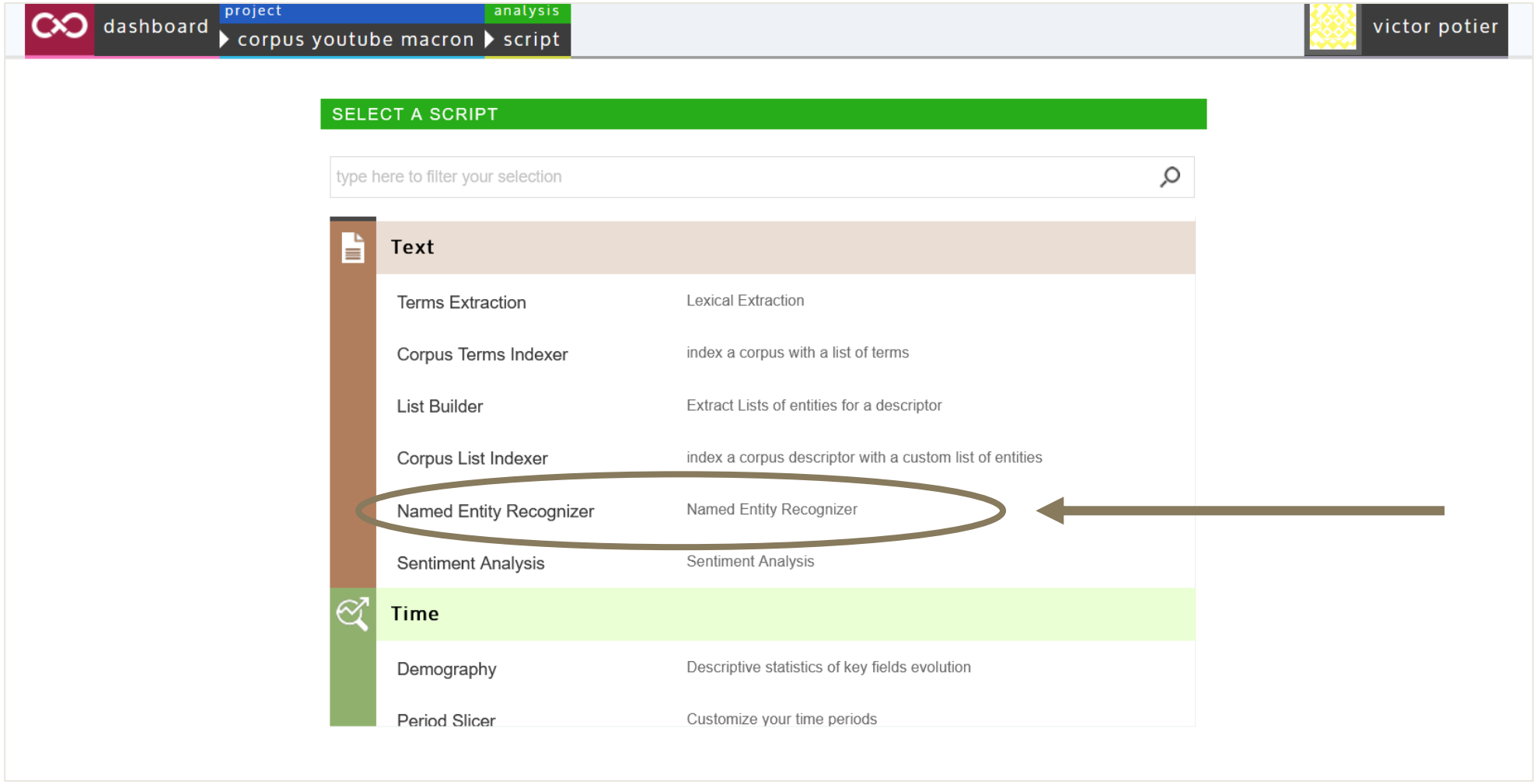
INDEXER UNE LISTE À UN CORPUS (APRÈS AVOIR MODIFIÉ VOTRE LISTE INITIALE)



On précise bien la variable qui nous intéresse pour ce *term extraction* (ici *« text »* - le contenu des commentaires)

*Pensez bien à **donner un nom** à cette opération*

Tutoriel script *Named entity recognizer*



Tutoriel script *Named entity recognizer*

- Permet notamment d'identifier des noms de personnes, de lieux, etc.
- Génère une liste de termes
- On l'utilise pour reconstituer une liste de termes et indexer le corpus avec, pour l'ajouter à de futures exploitations (network mapping notamment)

Fun fact : ça marche aussi avec les emoji

	entity	frequency
1	entity	frequency
2	😂	1880.00
3	🤣	491.00
4	😅	437.00
5	❤	436.00

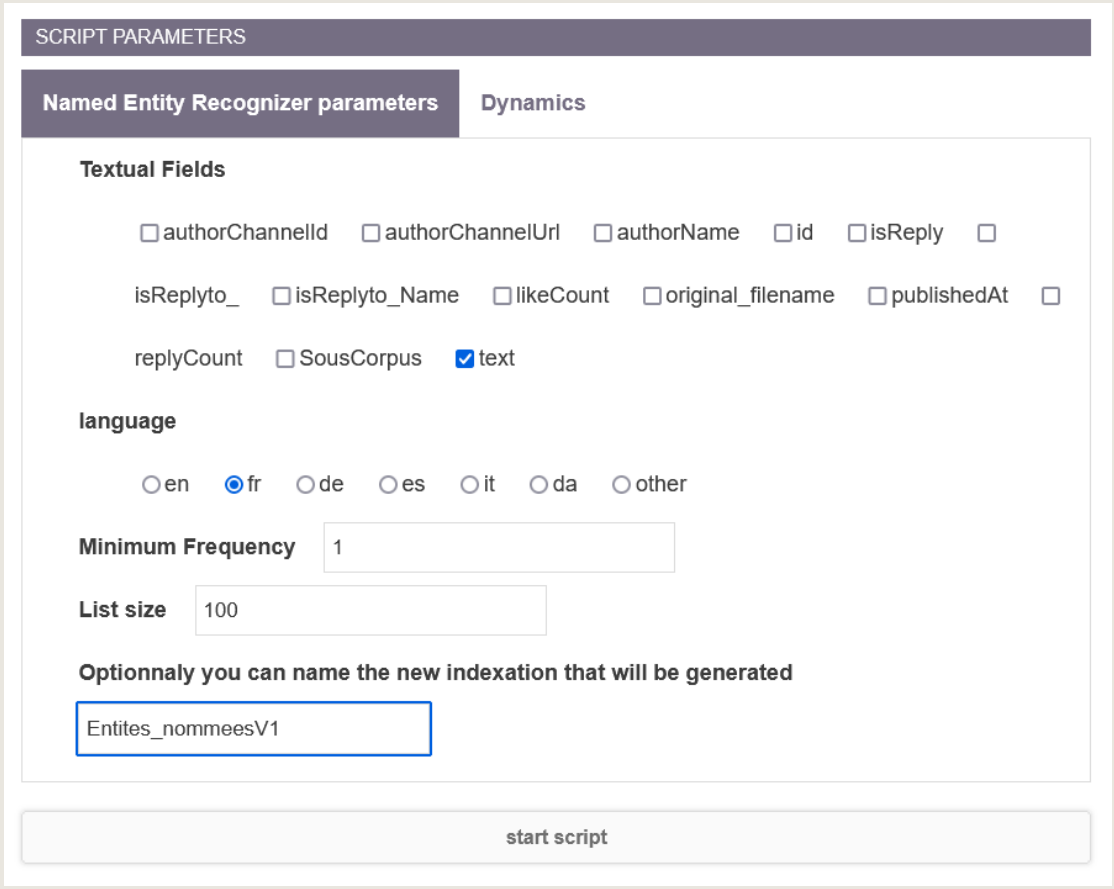
START SCRIPT

On sélectionne Text > « Named Entity Manager »

SELECT A CORPUS

On sélectionne son corpus

SCRIPT PARAMETERS



- On sélectionne le « textual field » = « text » ici
- Langage = FR (important !)
- Le type d'entité choisie : personnes, organisation, emoji, etc. si on est en EN (pas nécessaire en FR)
- Le nombre minimum de fréquences d'occurrence
- La taille de la liste

➔ On n'oublie pas de nommer la nouvelle indexation qui sera générée !

ON OBTIENT PLUSIEURS LISTES SELON LE TYPE D'ENTITÉ RECONNUES

REGARDONS LES PERSONNES : PER_0_0.TSV



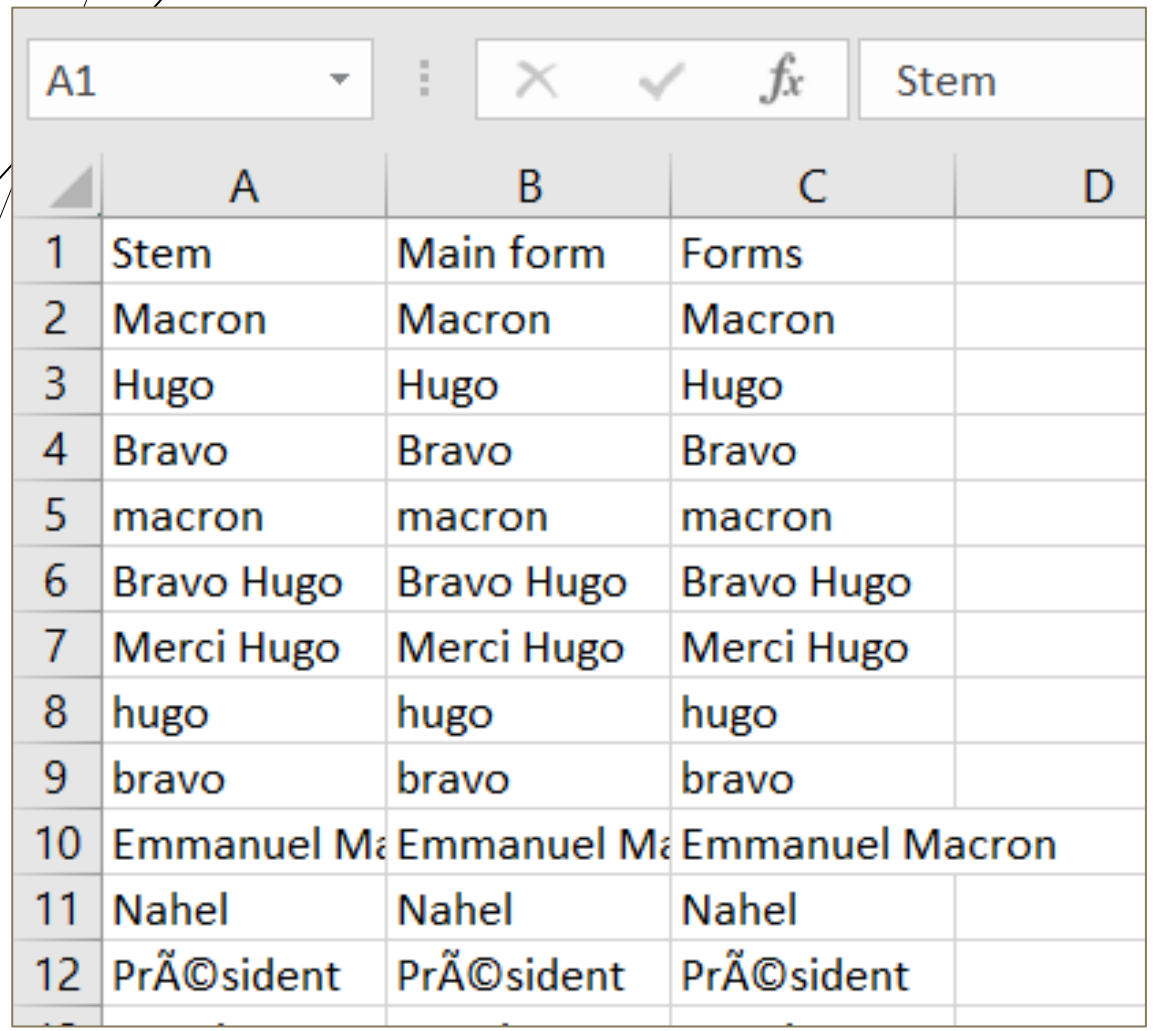
ON OBTIENT PLUSIEURS LISTES SELON LE TYPE D'ENTITÉ RECONNUES

	entity	frequency
1	entity	frequency
2	Macron	1668.00
3	Hugo	1279.00
4	Bravo	343.00
5	macron	198.00
6	Bravo Hugo	154.00
7	Merci Hugo	129.00
8	hugo	128.00
9	bravo	123.00
10	Emmanuel Macron	120.00
11	Nahel	110.00
12	Président	104.00
13	Degois	78.00
14	Poutine	72.00
15	Paul	70.00
16	Manu	66.00
17	Françoise	63.00
18	Dommage	51.00
19	Françoise Degois	51.00
20	Mr Macron	49.00
21	Benella	49.00

Plusieurs erreurs : multiples Macron, des merci hugo, etc.

Il faut la nettoyer

ON OBTIENT PLUSIEURS LISTES SELON LE TYPE D'ENTITÉ RECONNUES



A1 | Stem

	A	B	C	D
1	Stem	Main form	Forms	
2	Macron	Macron	Macron	
3	Hugo	Hugo	Hugo	
4	Bravo	Bravo	Bravo	
5	macron	macron	macron	
6	Bravo Hugo	Bravo Hugo	Bravo Hugo	
7	Merci Hugo	Merci Hugo	Merci Hugo	
8	hugo	hugo	hugo	
9	bravo	bravo	bravo	
10	Emmanuel Ma	Emmanuel Ma	Emmanuel Macron	
11	Nahel	Nahel	Nahel	
12	PrÃ©sident	PrÃ©sident	PrÃ©sident	

Comment faire ?

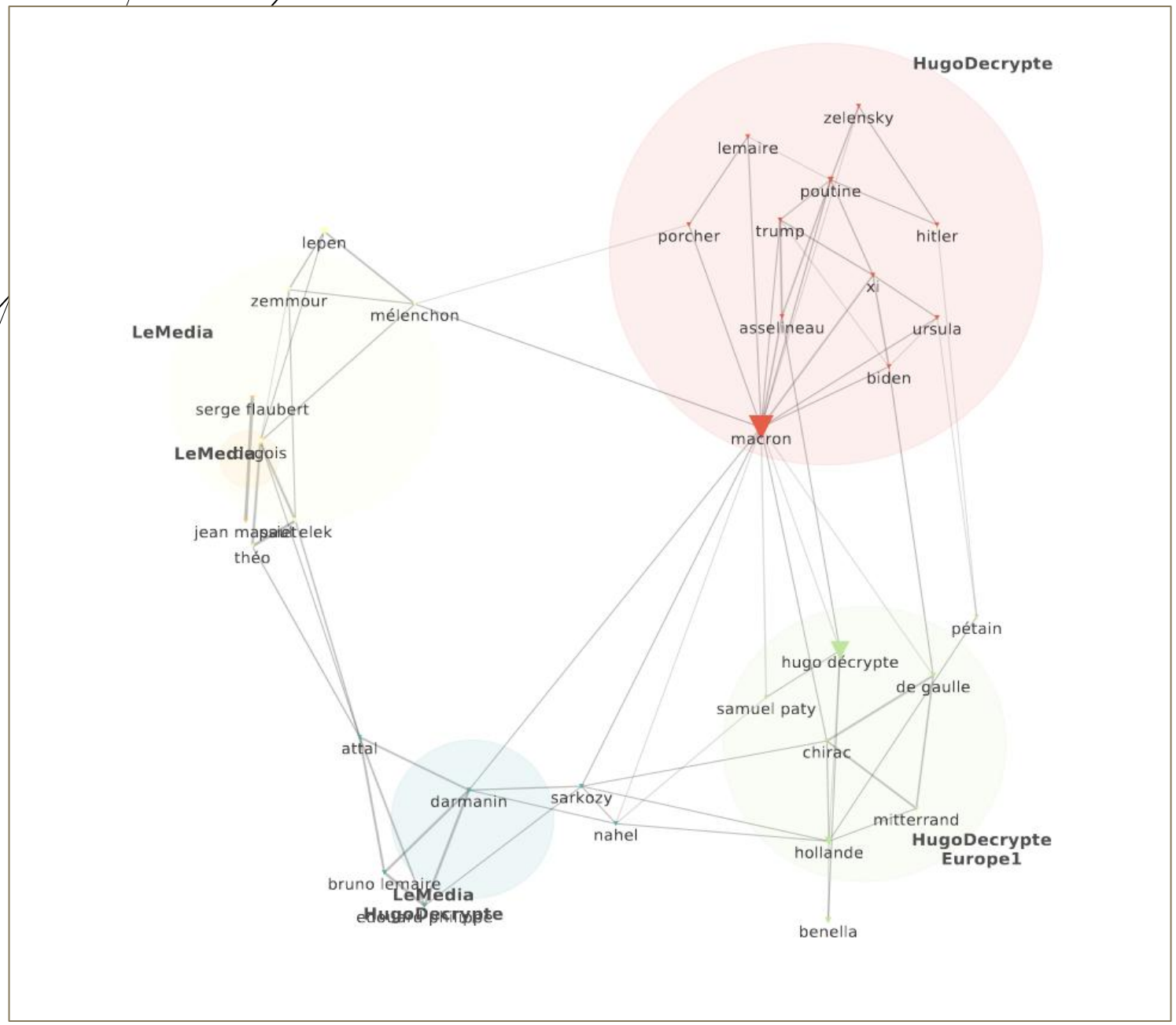
- on va la télécharger sur le PC depuis CorText
- Ouvrir Excel depuis son PC > ouvrir fichier (CTRL – O) > sélectionner le fichier TSV
- Renommer les trois premières colonnes du tableur « Stem », « Main Form » et « Forms »… pour en faire une term list !
- *Respectez les majuscules*
- Copiez-collez les résultats de la liste d'entités dans les 3 colonnes

ON OBTIENT PLUSIEURS LISTES SELON LE TYPE D'ENTITÉES RECONNUES

	stem	main form	forms
1	stem	main form	forms
2	macron_4	macron	macron|&|emmanuel macron|&|mr macron|&|monsieur macron|&|emmanuel|&|m. macron|&|président macron|&|e. macron|&|manu
3	hugo_7	hugo décrypte	hugo|&|hugo décrypte
4	nahel_56	nahel	nahel|&|naël|&|nael
5	Degois_12	Degois	degois|&|françoise degois|&|mme degois|&|madame degois|&|f. degois
6	Poutine_13	Poutine	poutine
7	Hollande_59	Hollande	hollande
8	De Gaulle_22	De Gaulle	de gaulle
9	Mélenchon_81	Mélenchon	mélenchon
10	Benella_20	Benella	benella
11	Darmanin_28	Darmanin	darmanin
12	lepen_83	lepen	lepen|&|le pen|&|marine le pen
13	Jésus Christ_32	Jésus Christ	jésus christ
14	Porcher_96	Porcher	porcher|&|thomas porcher
15	Paul Elek_24	Paul Elek	paul elek|&|paul

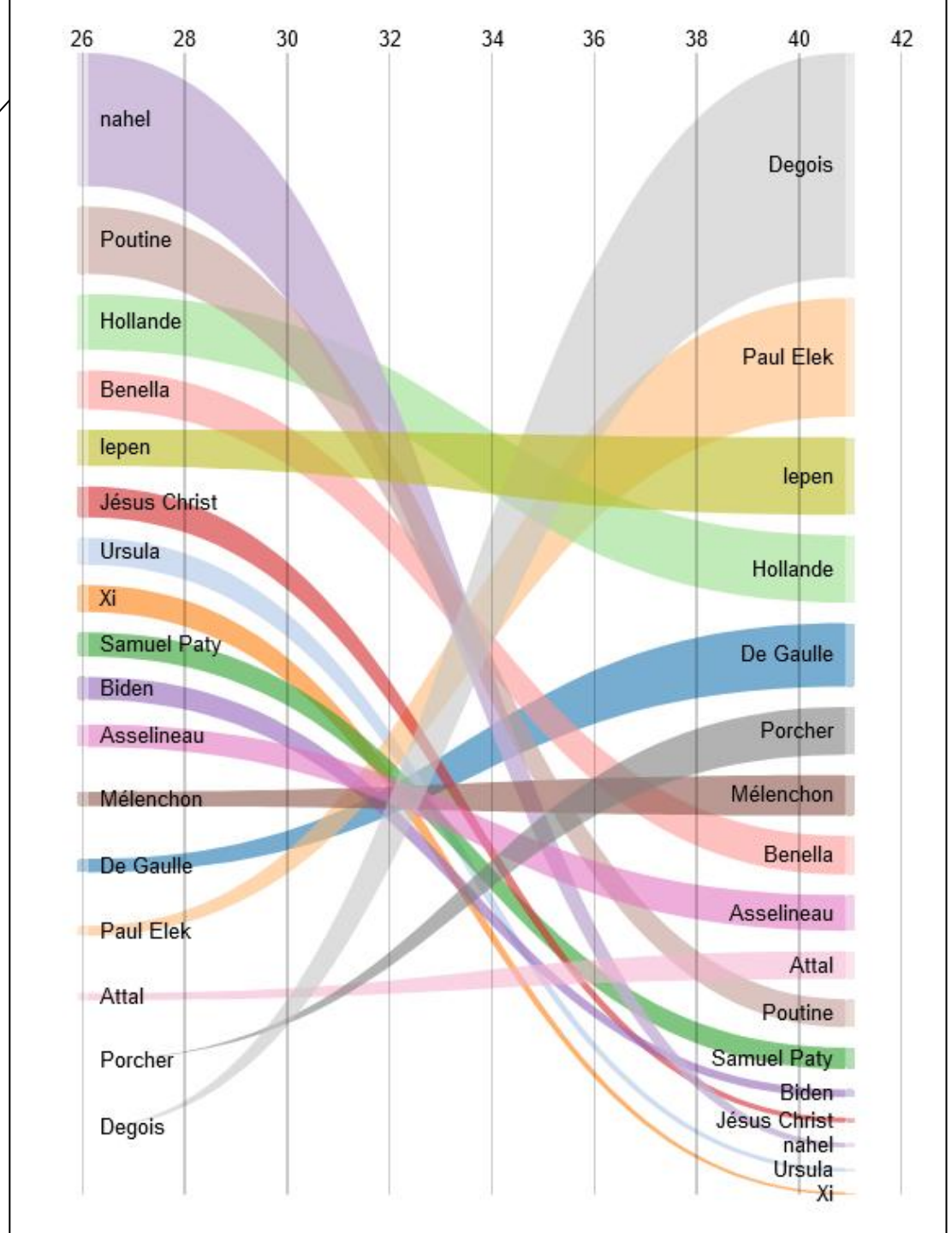
Retour sur CorText

- On upload la nouvelle liste de terms
- On identifie bien qu'il s'agit d'un fichier csv
- On peut la nettoyer comme une liste de terme classique ➔ *Remove Row* ou en concaténant des entrées « |&| »
- On n'oublie pas de renommer sa nouvelle liste. Ex. entitee_nommee_v2
- On peut indexer le corpus avec en utilisant le script *corpus term indexer*



FINALITÉ ?

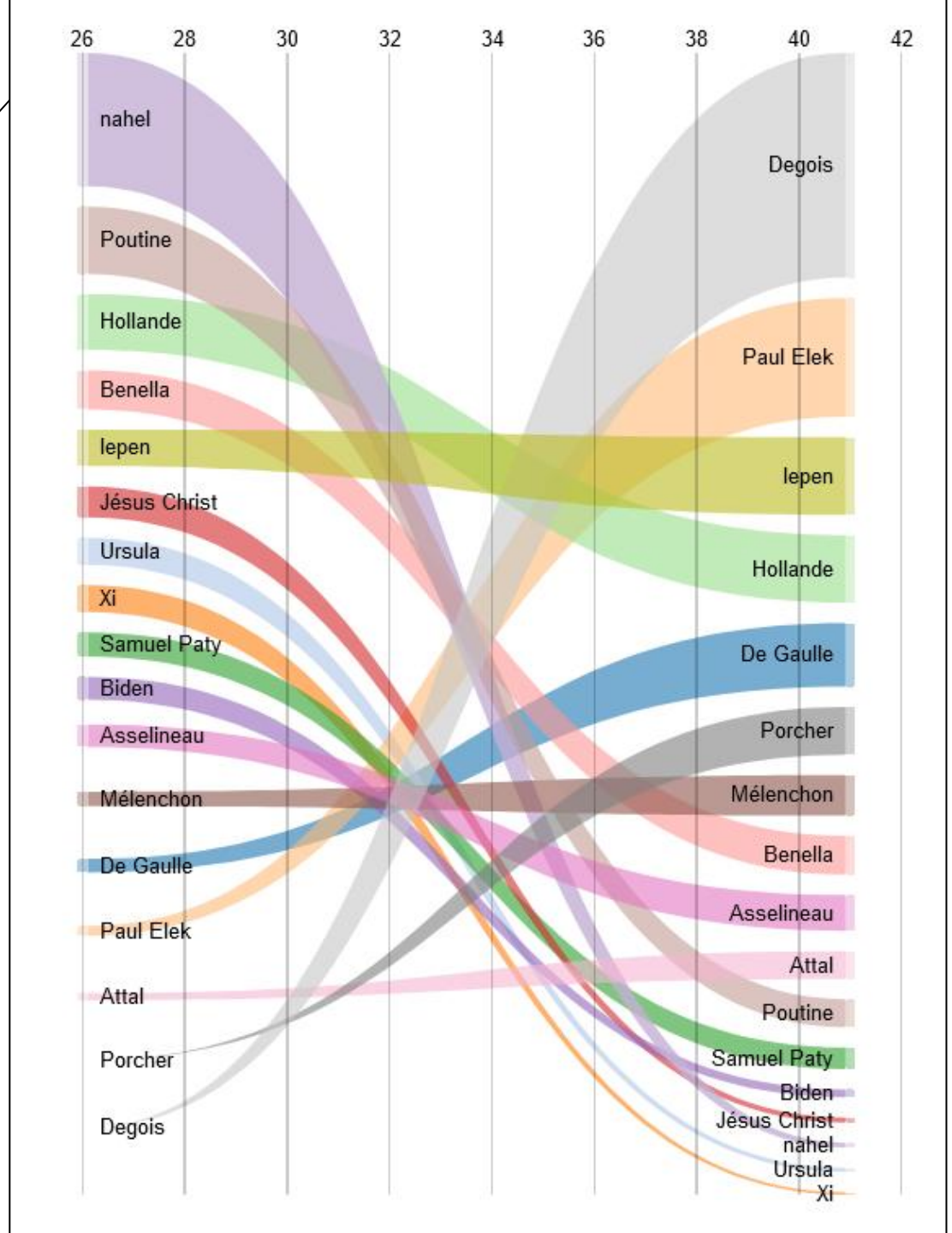
Représenter par exemple quelle chaîne fait parler plus de telles personnes ?



FINALITÉ ?

Petit exercice :

1. Comment fait-on pour enlever Macron et Hugo Décrypte ?



FINALITÉ ?

Petit exercice :

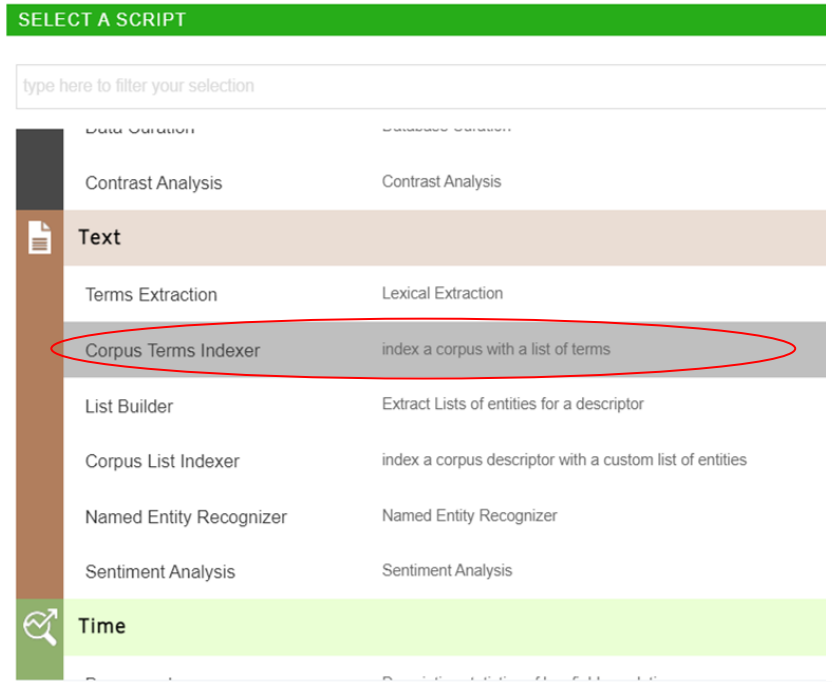
1. Pourquoi Degois et Paul Elek apparaissent soudainement ?

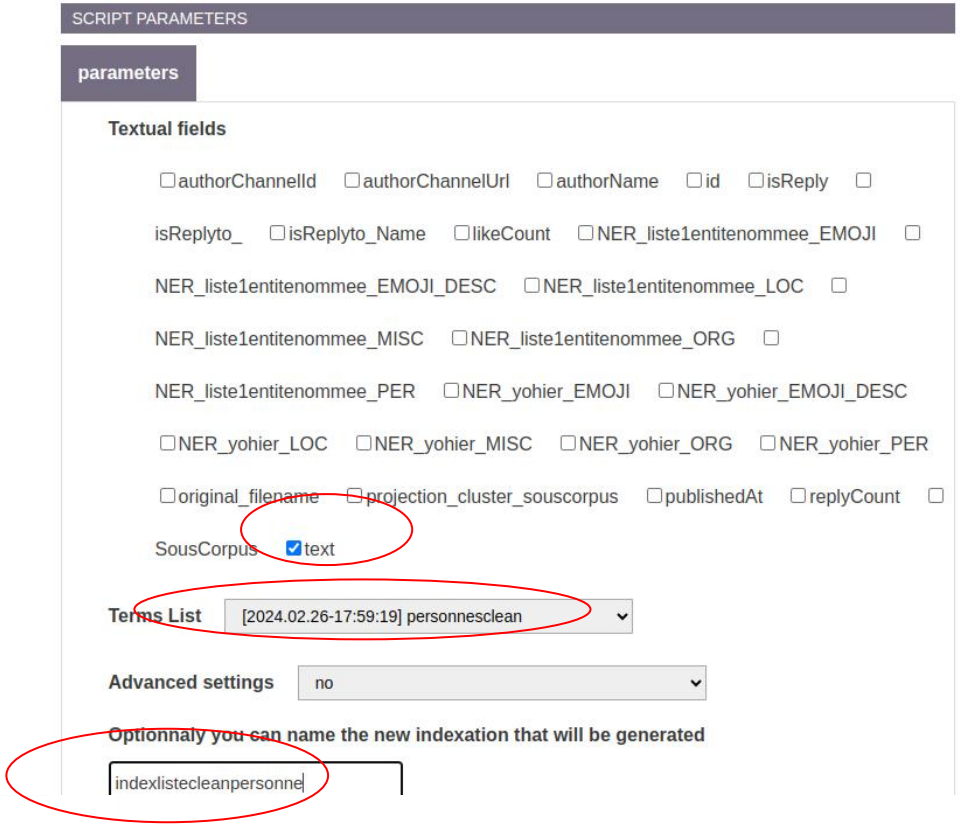
Le mapping “de base”

- Le Network mapping propose une représentation spatiale des co-occurrences
- Définition : la **cooccurrence** est la présence simultanée de plusieurs mots dans le même énoncé.
- En mettant dans le script “network mapping” la liste des termes les plus fréquents en champs 1 et en champs 2 (field), on va obtenir une représentation graphique - non pas des occurrences, mais **des co-occurrences.**

→ Une représentation graphique des réseaux de co-occurrences d’un corpus permet de voir quels **sont les mots les plus utilisés ensemble** (ce sont les **“noeuds”** qui sont liés ensemble). Cela permet alors de dégager des groupes thématiques, qu’on appelle aussi **“clusters”**.

Rappel : ne pas oublier d'indexer une liste à un corpus



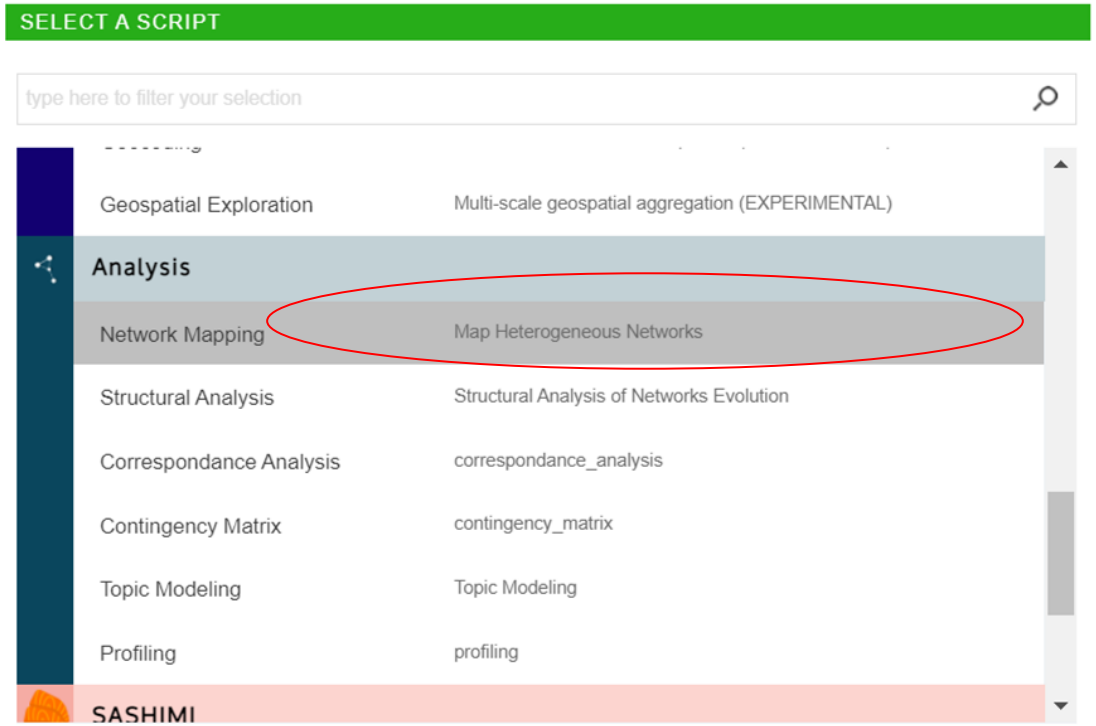


On utilise le script "Corpus Terms Indexer" en précisant la variable qui nous intéresse pour ce term extraction (ici "text")

On peut donner un nom à cette opération (en bas de l'écran)

Une fois que vous obtenez le drapeau vert, c'est que ça a fonctionné ! Vous pouvez passer à la seconde étape : network mapping

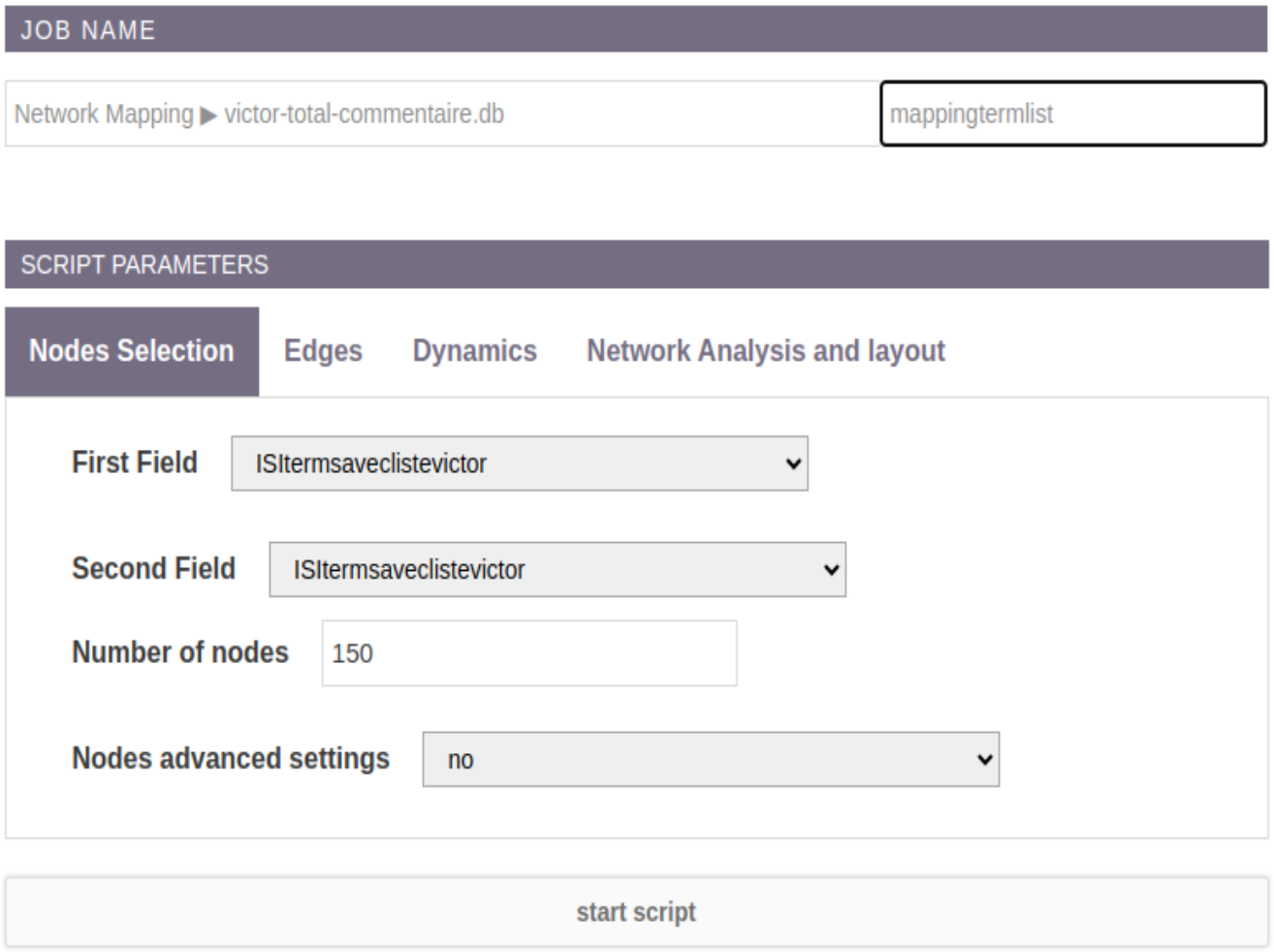
Network mapping **des expressions les plus fréquentes**



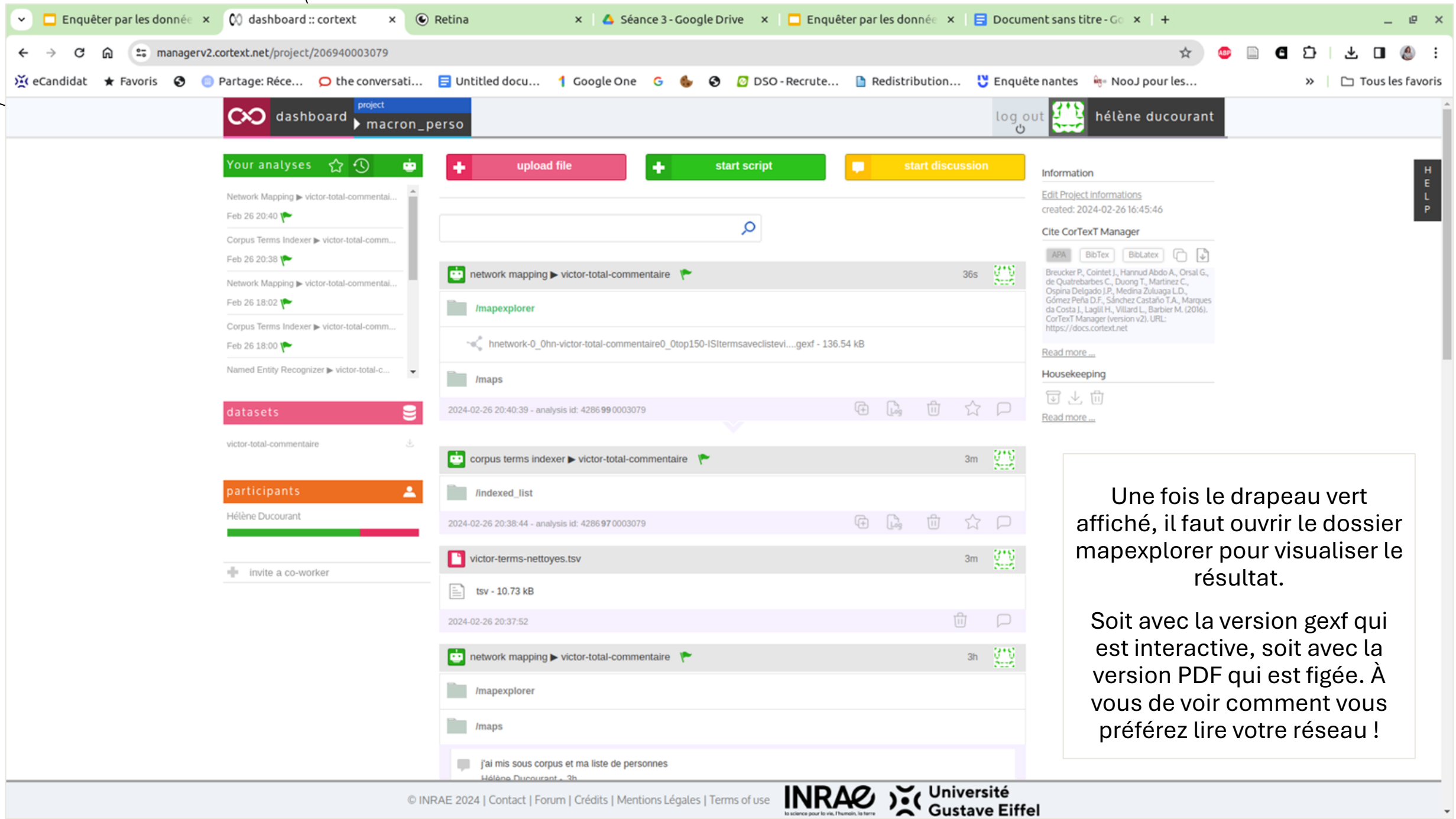
Le premier network mapping à réaliser, c'est celui des expressions les plus fréquentes (la **term list** nettoyée).

Il faut d'abord préciser la base de données sur laquelle vous travailler. Puis il faudra lui demander d'aller chercher la liste de termes par deux fois : **pour le premier champ ET pour le second**. Ne changez rien ! il vous propose par défaut votre term list des mots les plus fréquents ! Elle commence par **"ISIterms"**. On verra plus tard ce que l'on peut changer…

Pour info, Node veut dire Nœud, combien de nœuds voulez-vous dans votre graphique ? Là encore, on ne change rien.



On choisit la liste de terme nettoyée que l'on place en first field et second field
(ISI apparaît devant)

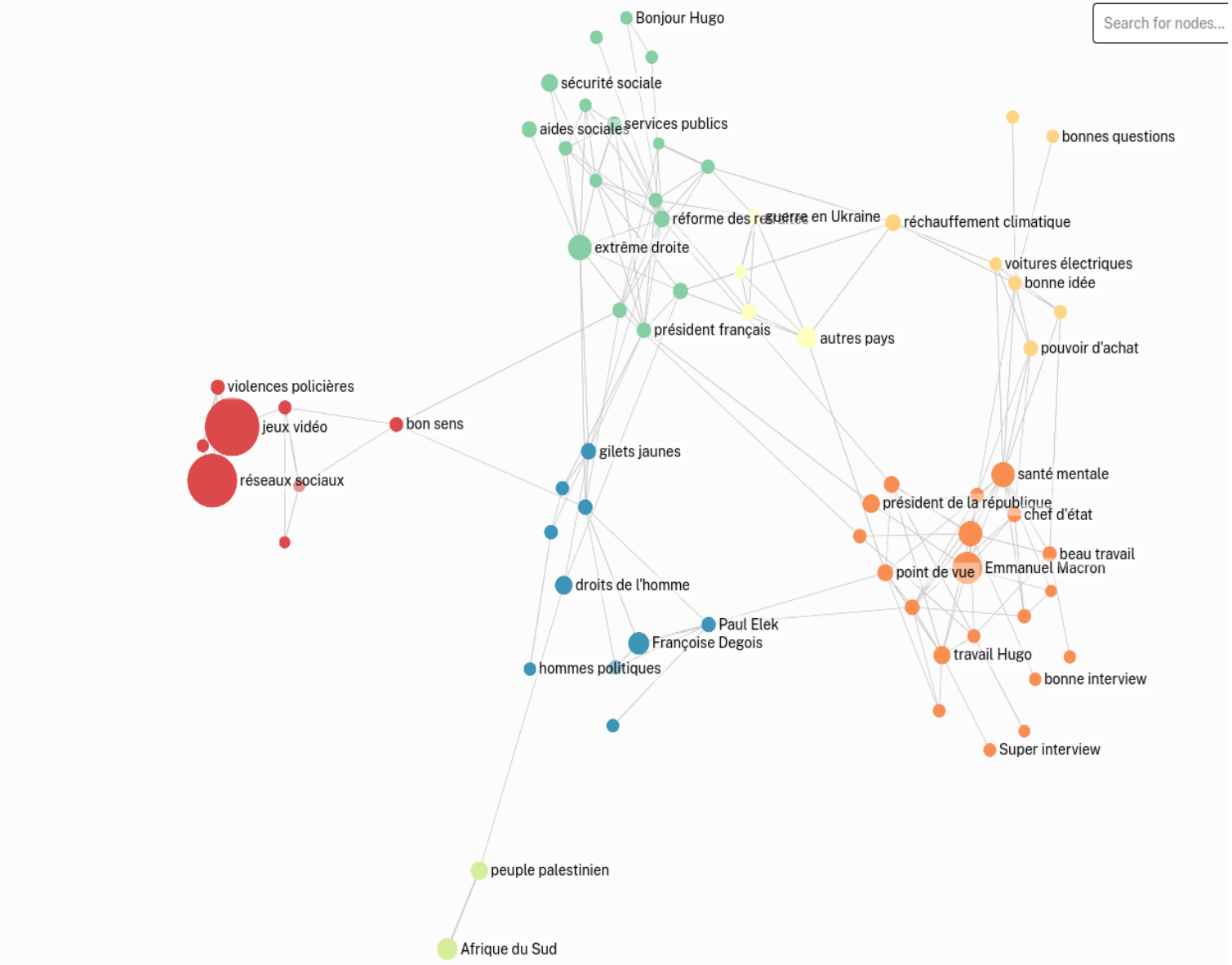


Une fois le drapeau vert affiché, il faut ouvrir le dossier mapexplorer pour visualiser le résultat.

Soit avec la version gexf qui est interactive, soit avec la version PDF qui est figée. À vous de voir comment vous préférez lire votre réseau !

https://retina.cortext.net/#/graph/?url=https%3A%2F%2Fassets.cortext.net%2Fdocs%2F676a4c37b22852167bd3e84e460e463d&r=d
On peut télécharger la carte sémantique pas nécessaire de faire une capture d'écran.

Il faut se creuser la tête pour donner sens aux clusters -
les expressions proches = elles co-occurent plus fréquemment que la moyenne dans les commentaires



Qu’est-ce que ce mapping vous dit de votre corpus ?

Gardez en tête votre objectif, votre question de départ.

Comment commenter une telle carte de co-occurrences - rien de simple !

1 - les éléments les plus proches sont le plus souvent que la moyenne ensemble dans les mêmes commentaires.

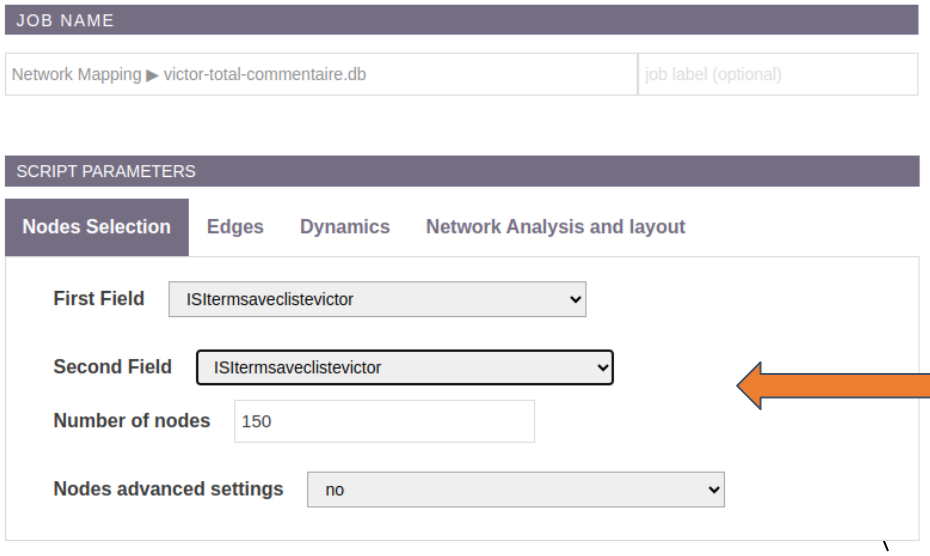
2 - Il ne faut pas forcément commenter tous les clusters

3 - On peut commenter les groupes de termes qui sont éloignés sur la carte. (moins souvent que la moyenne ensemble).

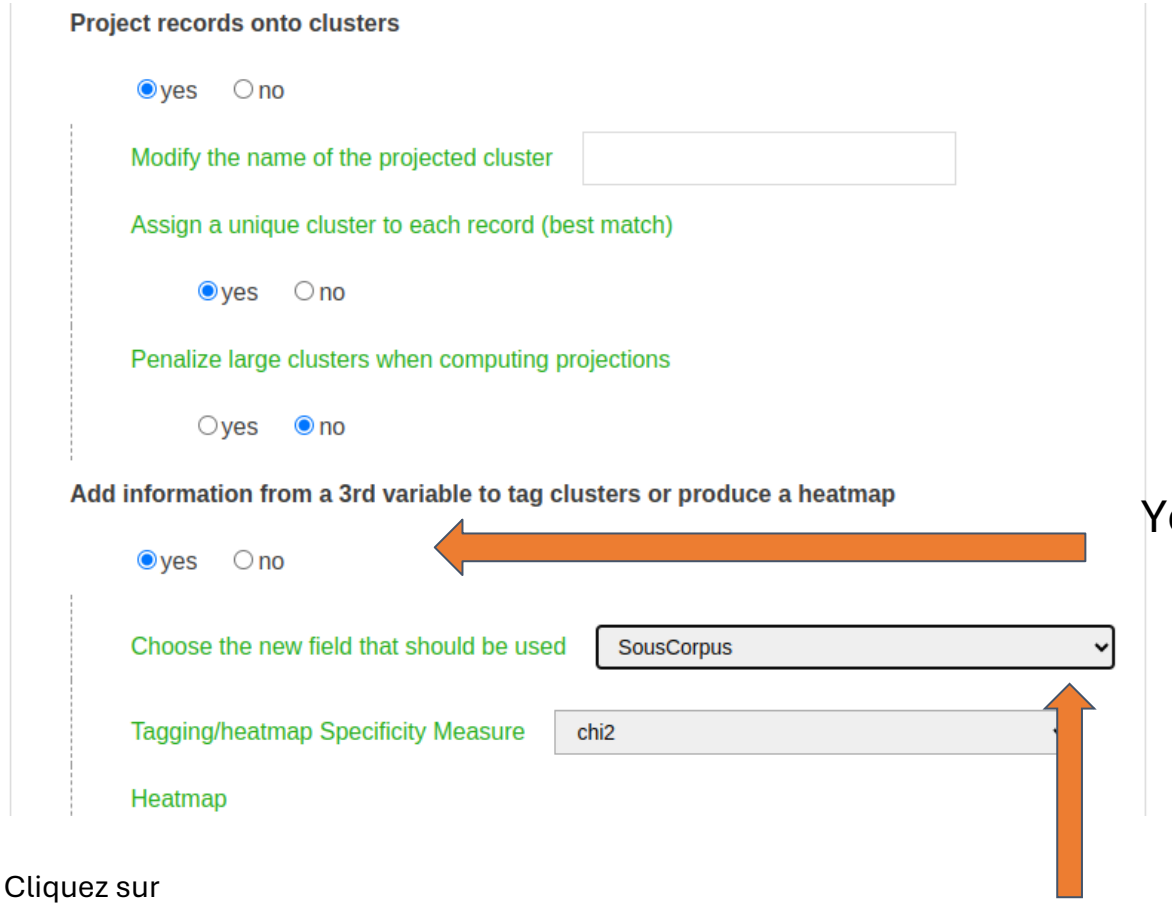
Produire un autre type de network mapping
“Ajouter une nouvelle variable”

- Et si on ajoutait des infos complémentaires ? Bonne idée !
  - On pourrait demander à CorText de nous dire sous quels médias/chaînes… se trouvent les co-occurrences ?
  - On pourrait chercher à mettre en relation votre liste de personnes et les clusters de co-occurrences des termes les plus fréquents ?

Faire cela dans le script Network Mapping s’appelle **“ajouter une troisième variable illustrative”**. Voyons ensemble comment faire.



Cliquez sur
Network Analysis



Yes

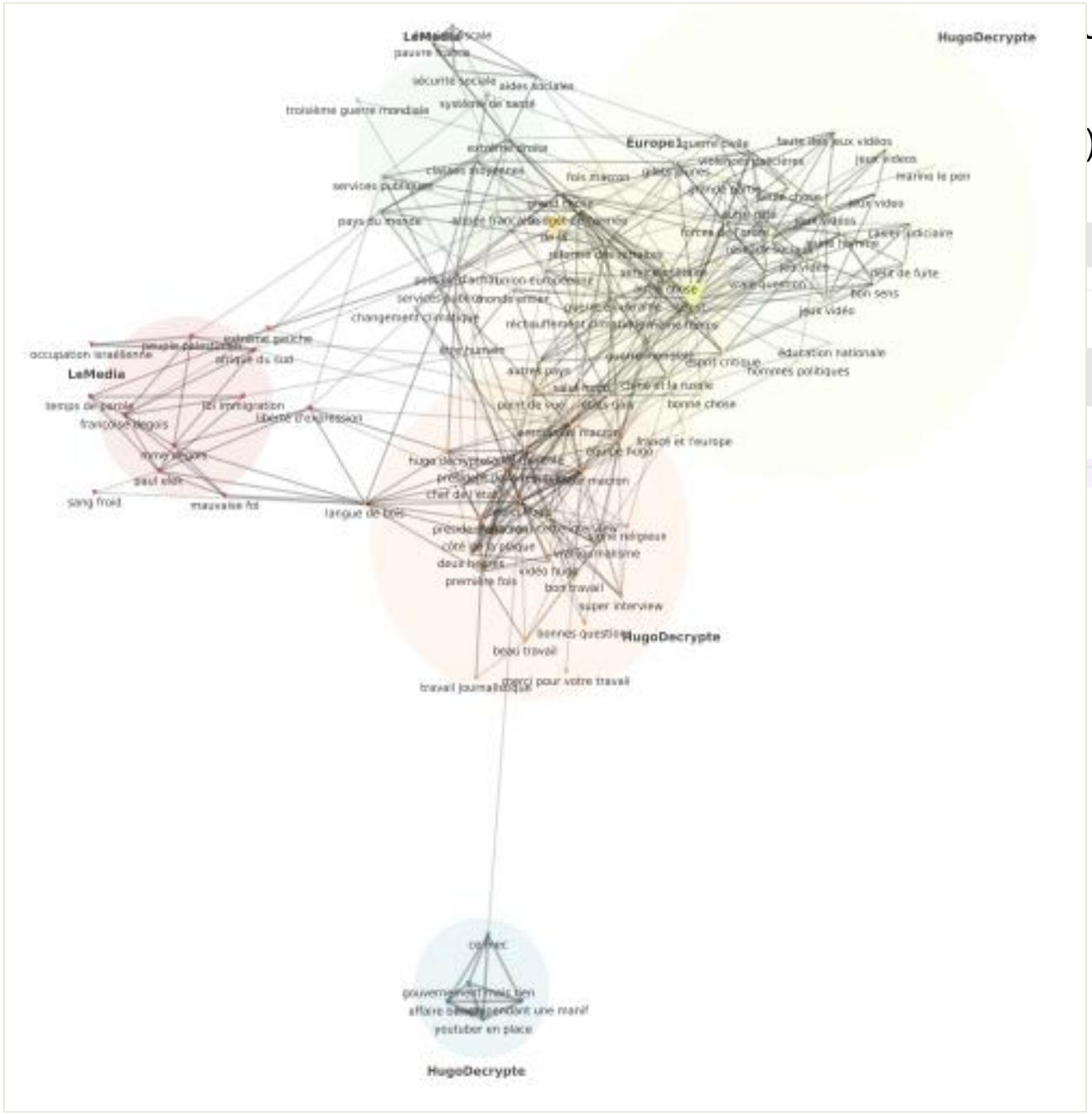
Choisissez une variable : ex sous-corpus, une liste que vous avez créée…

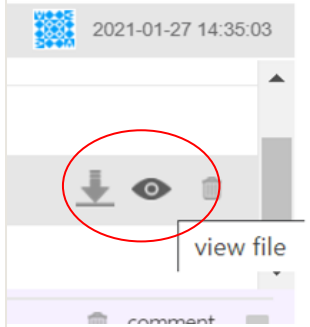
On refait le script “network mapping”, et là, on utilise l’onglet “network analysis and layout” pour changer certains paramètres du script

ur voir apparaître la 3e

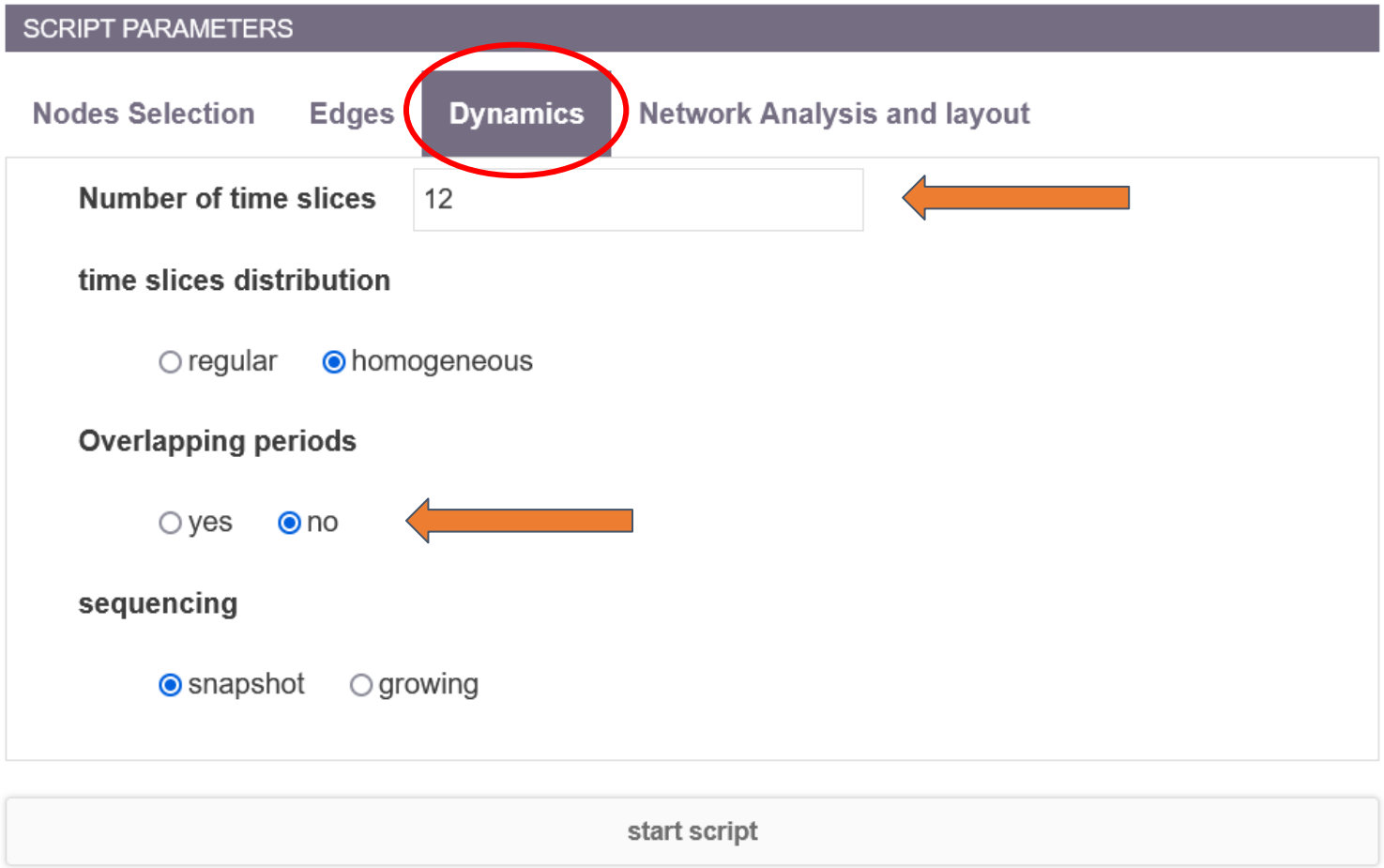
visualiser le network

)

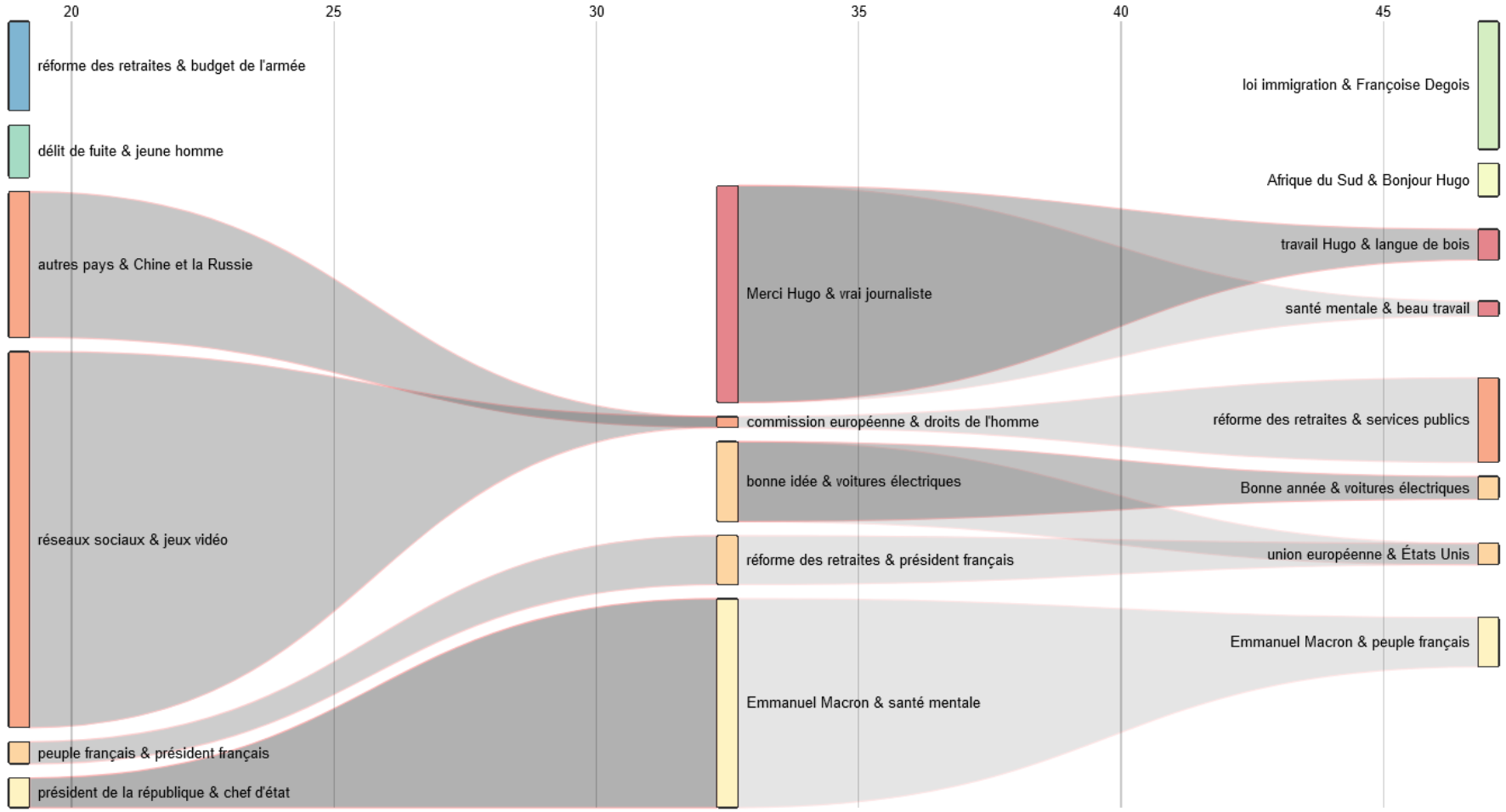


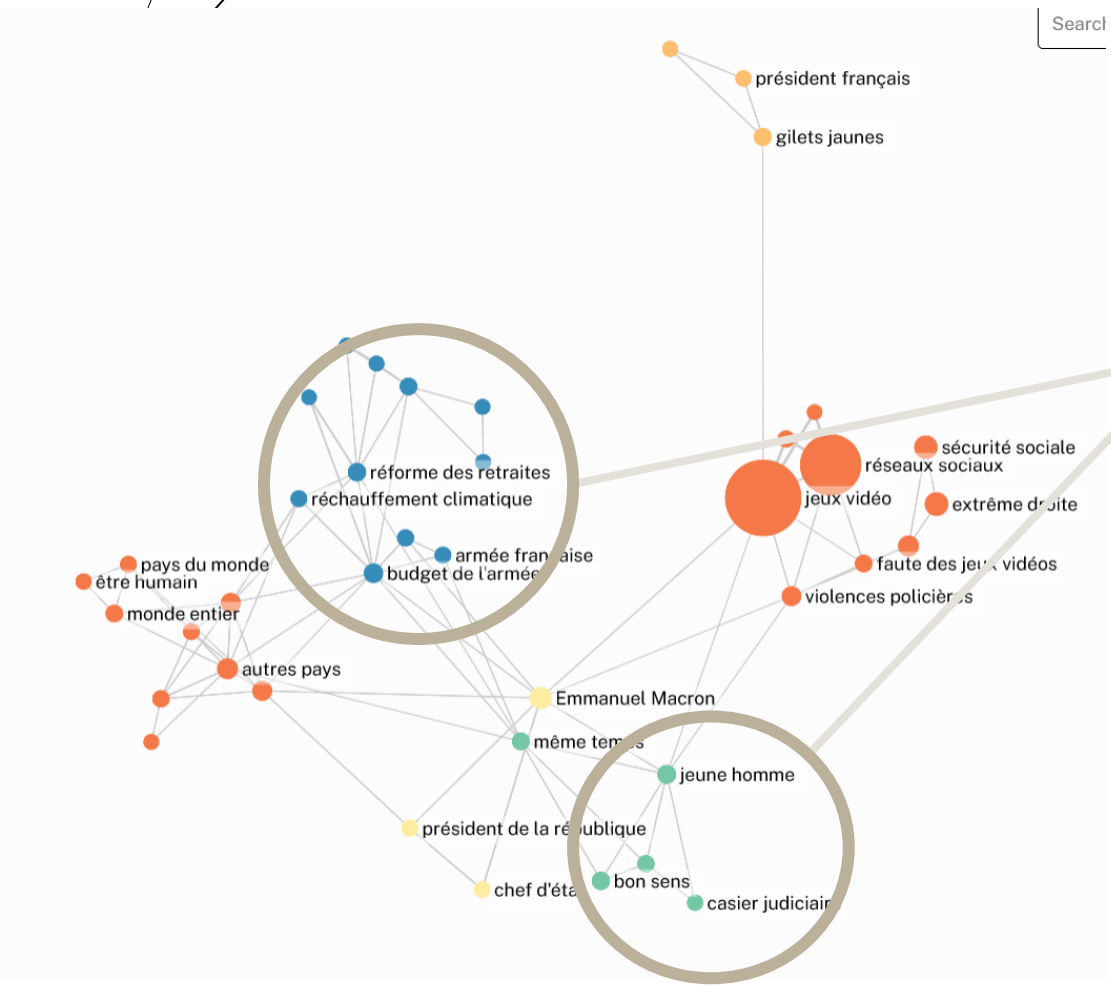


Faire une visualisation temporelle et dynamique des clusters

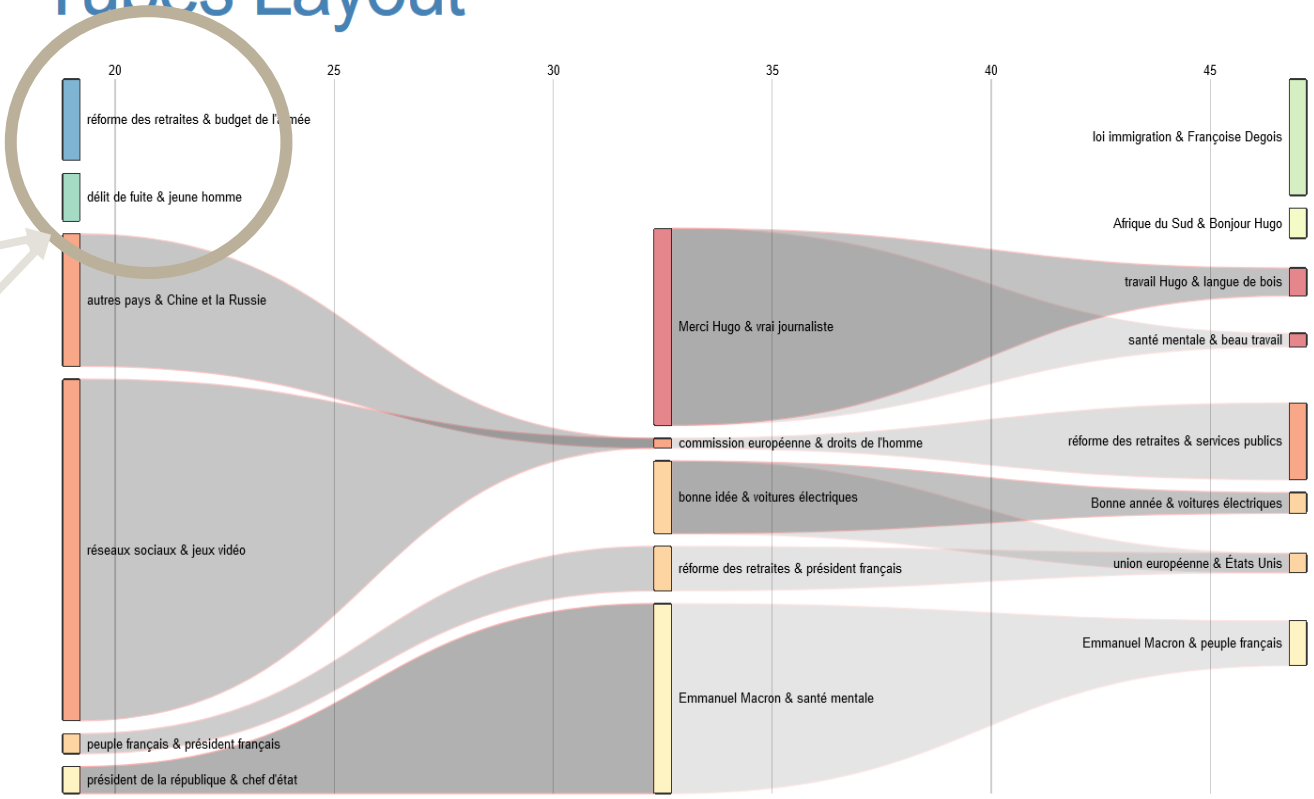


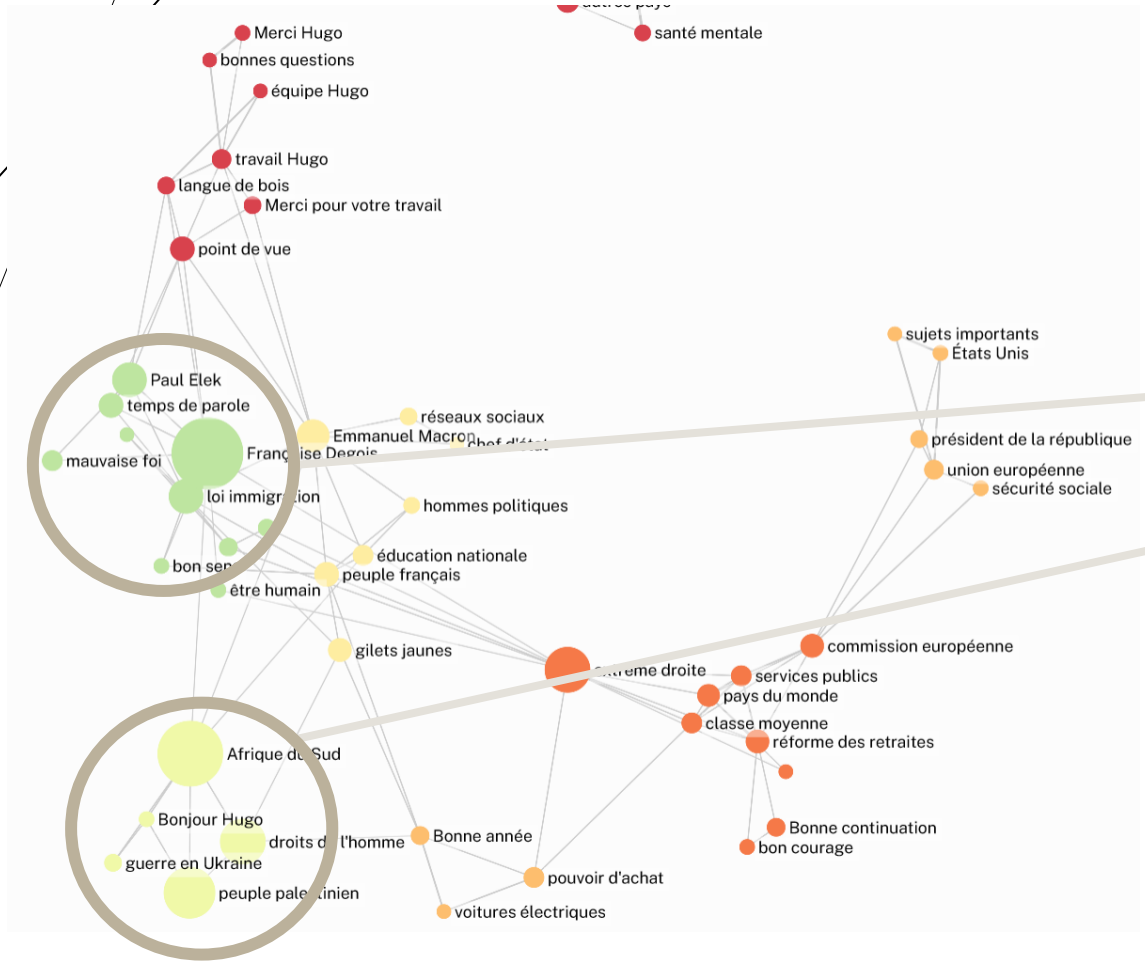
Tubes Layout



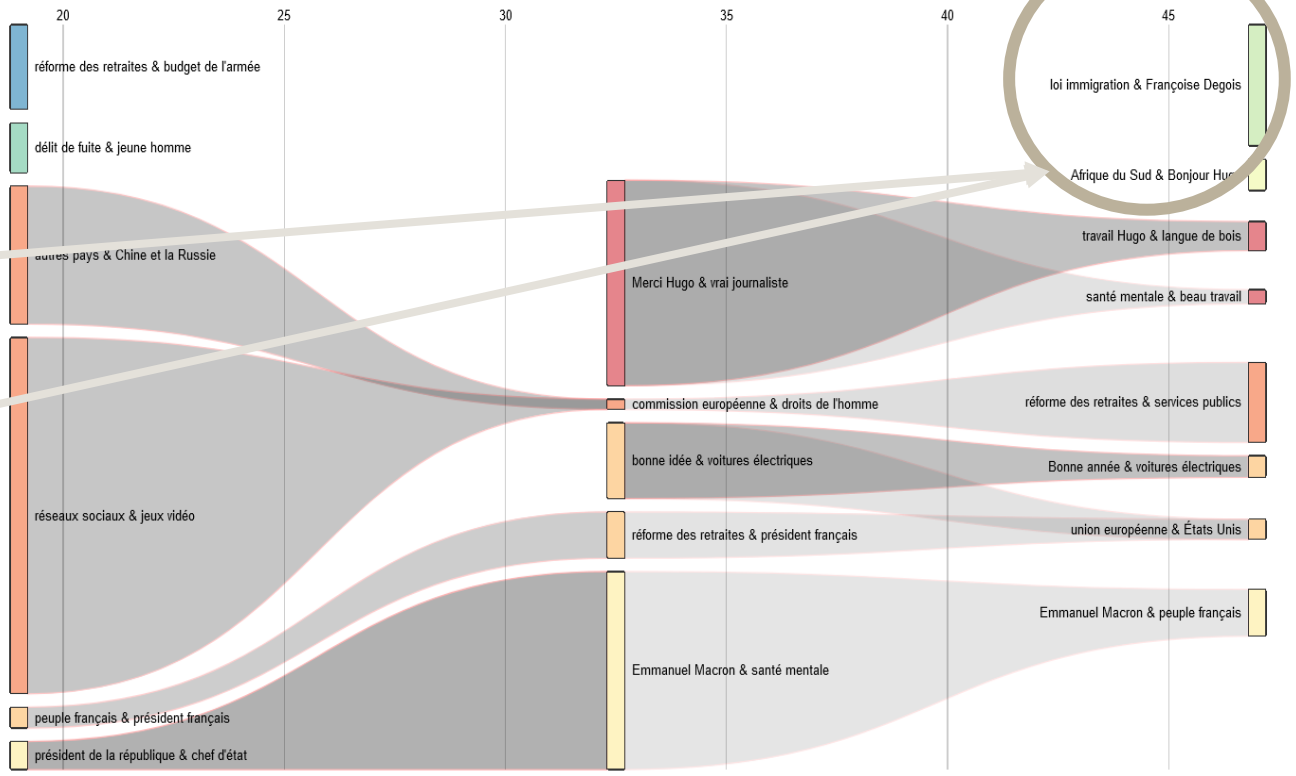


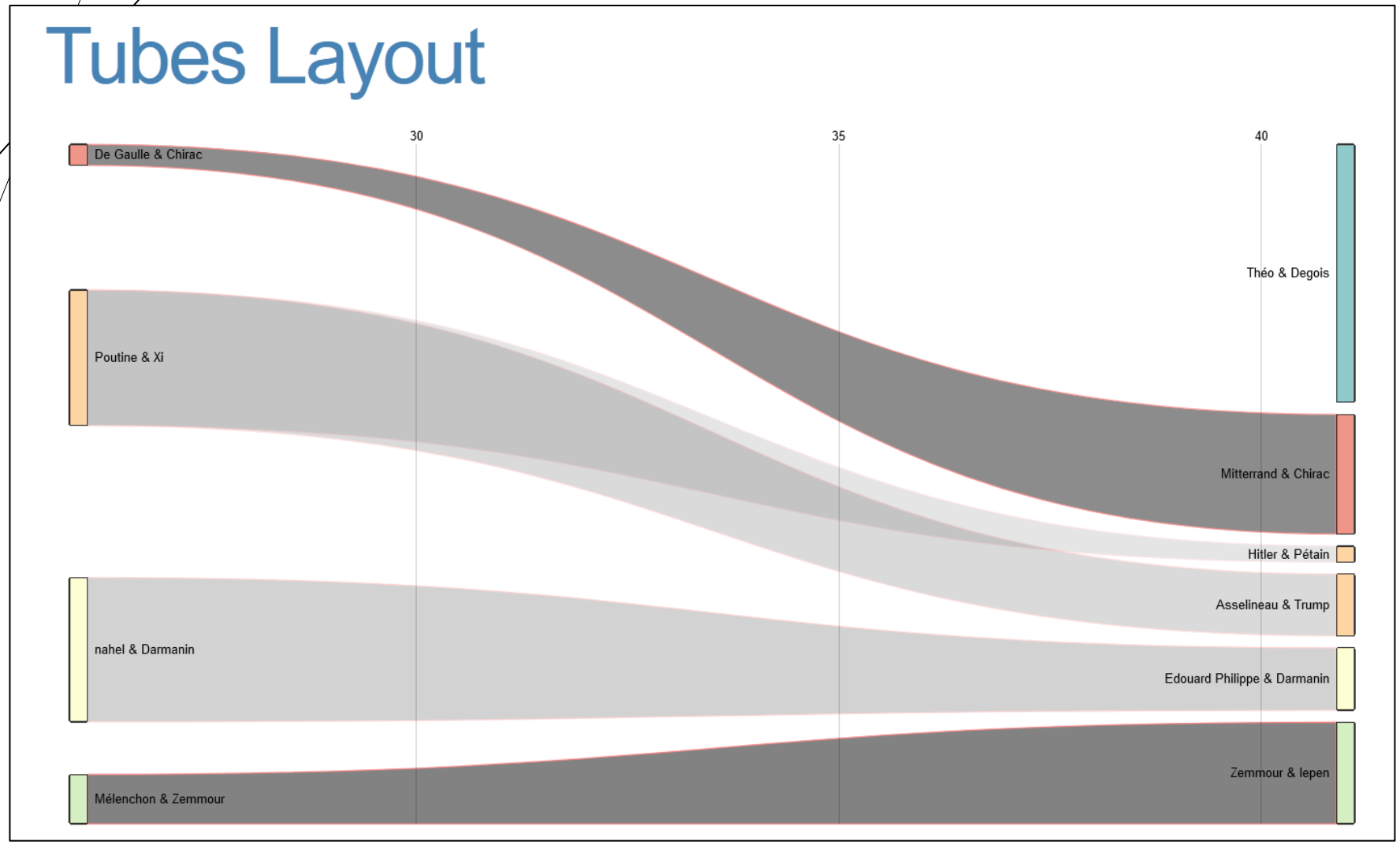
Tubes Layout





Tubes Layout

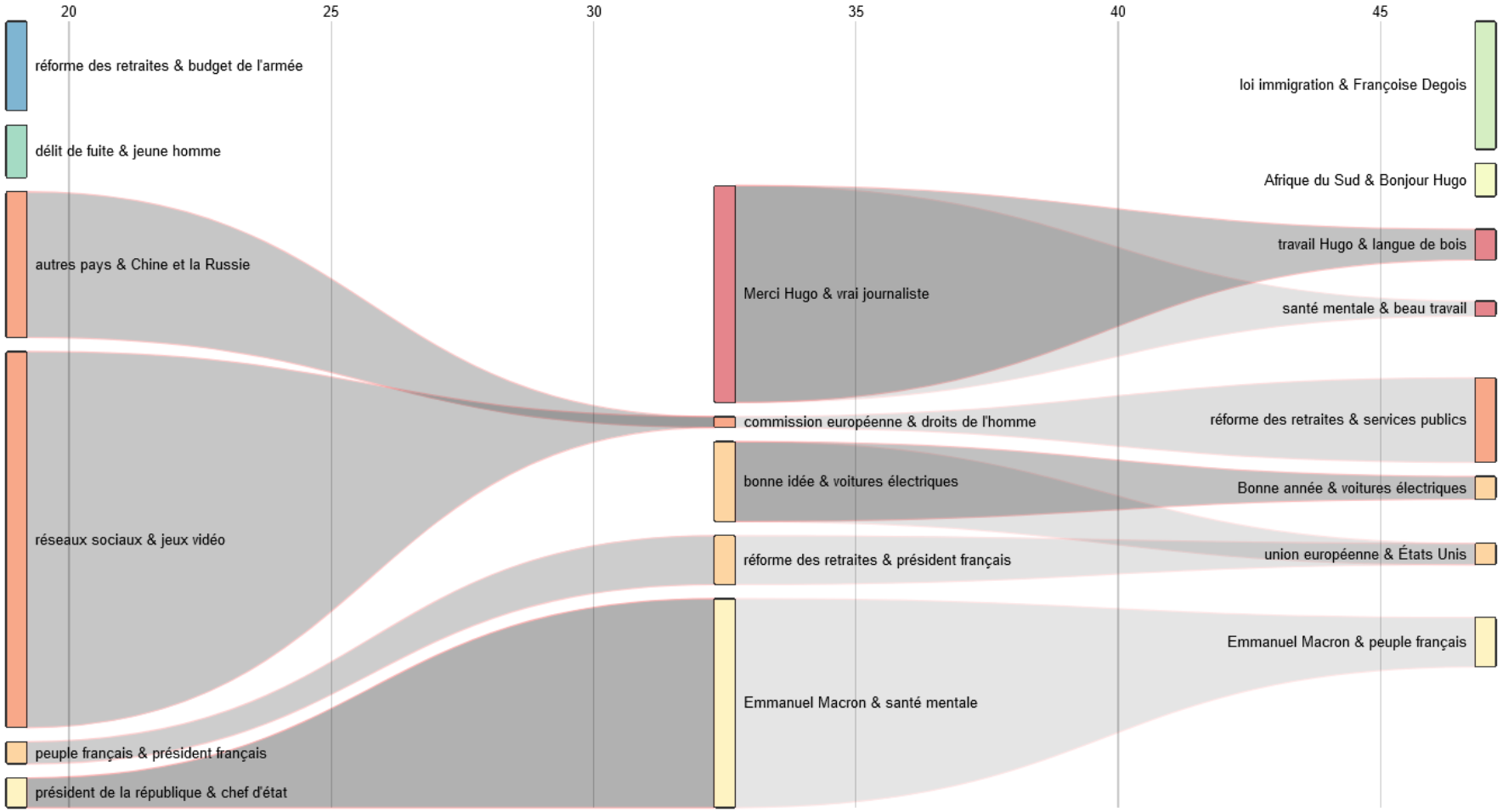




EN BONUS

On peut sortir une petite visualisation de l'évolution dans le temps des clusters de cooccurrences (ici sur la liste des noms sur l'ensemble des chaînes)

Tubes Layout



EN BONUS

Idem avec les sujets