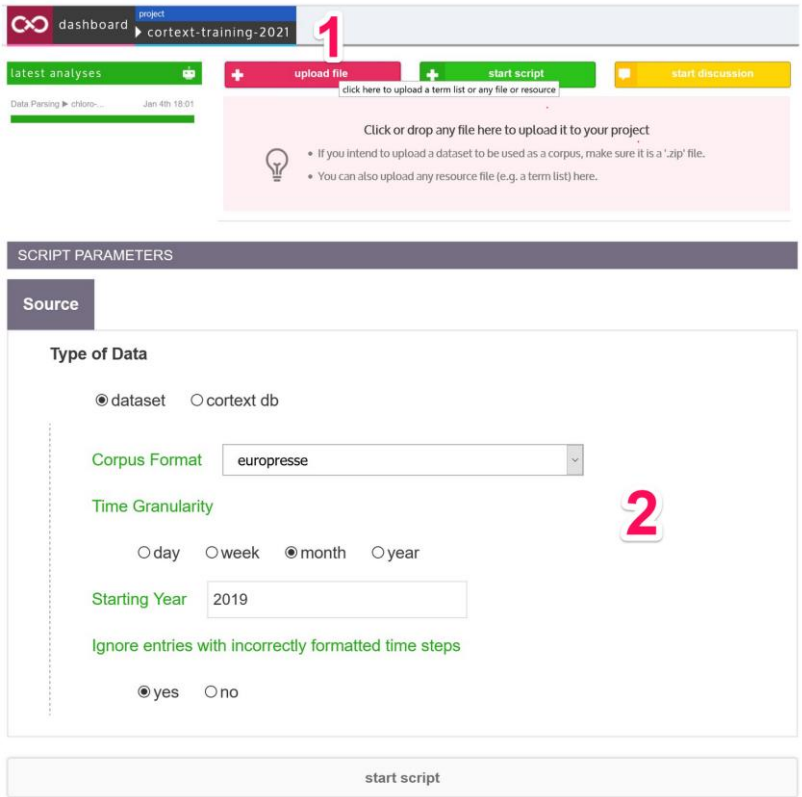
1/ Choisissez un corpus qui déterminera la question que vous allez vous poser

Espace scientifique (Web Of Science) : production académique sur la chloroquine

- Tous les articles scientifiques publiés en langue anglaise, entre janvier 2014 et décembre 2020, et accessible sur le la plateforme du Web of Science (datastet **chloro-sci-2014-2020-v2.zip** parser ISI sur Cortext Manager)
- Pour aller plus loin : tous les articles scientifiques publiés en langue anglaise, entre 2001 et 2020, et accessible sur le la plateforme du Web of Science (dataset **chloro-sci-2001-2020.zip** et parser ISI sur Cortext Manager). Ce corpus est déconseillé pour une première découverte de CorText Manager.

Inscrivez-vous et créez-vous un projet sur CorText Manager : https://managerv2.cortext.net/
Télécharger le corpus : https://docs.cortext.net/trainings/cortext-lisis/2-methodes/01-dataset/
Uploadez et parser le corpus



ISI pour le format les notices d'articles scientifiques (Web Of Science)

2/ Puis choisissiez une dimension d'analyse et une question

Dimensions d'analyse

- **Analyse sémantique** (réseaux de mots et identification des thèmes)
- **Analyse sociale** (réseaux de chercheurs, d'organisations, de lieux géographiques)
- Il est possible de croiser les deux (**socio-sémantique**) et d'utiliser la **dimension temporelle**

Exemples de questions, choisissiez-en une ou construisez votre propre question à partir de ces exemples

- Dans les articles qu'elles ont été les **sources bibliographiques mobilisées** dans les travaux des chercheurs sur ces sujets et quelles sont les « écoles de pensées » (relations directes | Corpus WOS : Network Mapping sur la variable Cited Ref) ?
- Quelles sont les **espaces géographiques** dont les chercheurs ont été les plus actifs en 2020 sur ces sujets (relations directes | Corpus WOS : Network Mapping sur la variable cities) ?
- Quelles sont les **organisations** dont les chercheurs ont été les plus actifs en 2020 sur ces sujets, et **comment collaborent-elles** (relations directes | Corpus WOS : Network Mapping sur la variable Research institutions) ?
- …

3/ Travailler le texte

Les groupes nominaux ne sont pas tous signifiants. Un travail de relecture du vocabulaire est utile pour améliorer la précision des analyses qui pourront être effectuées à partir des résultats produits. Plusieurs types de situations sont généralement identifiables :

- **le bruit :** les groupes nominaux extraits dans des sections parasites du texte (mise en page, tableaux et figures, édition, noms des journaux…)

University Press	University Press|&|Press University
op cit	op cit
Tableau II	Tableau II|&|tableau II
Harvard University Press	Harvard University Press|&|University Press Harvard
…	…

- **les mots vides de sens**

premier chapitre	premier chapitre|&|chapitre premier
Par exemple	Par exemple|&|Par cet exemple|&|exemple Par
juste titre	juste titre
…	…

- **les mots génériques**

première moitié	première moitié
année suivante	année suivante|&|années suivantes|&|année ou la suivante
autres groupes	autres groupes|&|autre groupe|&|groupe autre|&|groupe à l'autre|&|groupe à un autre…
parties prenantes	parties prenantes|&|partie prenante
mise en valeur	mise en valeur|&|valeur dont la mise
…	…

- **les groupes nominaux ayant une signification identique nécessitant d'être réunis**

18e siècle	18e siècle
XVIIIe siècle	XVIIIe siècle|&|xviiie siècle|&|XVIIIe siècles
Moyen Âge	Moyen Âge|&|âge moyen|&|moyen âge|&|Moyen Â ge
Moyen Age	Moyen Age|&|MOYEN AGE
…	…

Ces étapes nécessitent une expertise sur les sujets manipulés. Elles constituent le socle sur lequel s'appuient ensuite les analyses

3/ Les mesures de proximité

Dans le script **Network Mapping** il est demandé de préciser la mesure qui sera utilisée pour calculer la proximité / similarité entre deux variables (**onglet Edge** « promixity mesure » ou **l'onglet « Network Analysis and layout »** quand « Add information from a 3rd variable to tag clusters or produce a heatmap » est activé).

- *Pour aller plus loin :* https://docs.cortext.net/analysis-mapping-heterogeneous-networks/mapping/#tagging-heatmap-specificity-measure

proximity measures	type of network	normalisation	special properties
raw	interaction network (*e.g.* social network)	no	-
X^2	homogeneous & heterogeneous	yes	normalization tend to create links toward higher degree nodes
MI	homogeneous & heterogeneous	yes	Inspired from information theory
Cramer	homogeneous & heterogeneous	yes	-
cosine	homogeneous network (eg. semantic)	yes	Classical measure (originating from scientometrics)
distributional	homogeneous network (eg. semantic)	yes	very robust measure (coming from computational linguistics)
cosine_het	affiliation network (eg. users sharing the same hashtags)	yes	two fields are required but the final network is homogeneous
dot_product_het	affiliation network (eg. users sharing the same hashtags)	no	two fields are required but the final network is homogeneous

Pour les réseaux homogènes et hétérogènes

Raw correspond à la valeur brute (le compte, la fréquence). Mesure de cooccurrence brute. Par exemple, on comptera 1 pour la paire {carottes, poireaux} à chaque fois que carottes et poireaux apparaîtront ensemble dans une recette. Mesure pertinente pour la construction de réseaux de collaborations (aucune correction particulière de l'information est nécessaire ; respect des données). Repose sur l'hypothèse qu'un lien correspond à une interaction effective.

- est généralement à privilège pour **les réseaux sociaux** (collaborations entre des individus ou entre des organisations)
- *pour aller plus loin : https://docs.cortext.net/metrics-definitions/#raw*

Chi² : mesure l'intensité du lien entre deux termes en appréciant l'écart par rapport à la valeur attendue. Le Chi²est donc mesure de spécificité. La valeur attendue entre deux termes est égale à la somme de l'ensemble des cooccurrences du premier terme (avec, donc, l'ensemble des autres termes) multipliée par la somme de l'ensemble des cooccurrences du second terme (avec, donc, l'ensemble des autres termes), sur la somme de l'ensemble des cooccurrences entre elles (somme des lignes plus la somme des colonnes de la matrice de cooccurrences, soit, en fait, le nombre total de cooccurrences observées). Lorsqu'il est positif, l'écart entre la valeur réelle des cooccurrences de deux termes et entre la valeur attendue indique une surreprésentation du lien entre ces deux termes et donc une **spécificité**.

- est généralement à privilège pour **dégager la spécificité d'une valeur** dans un contexte (par exemple avec **l'onglet « Network Analysis and layout »** quand « Add information from a 3rd variable to tag clusters or produce a heatmap » est activé)
- *reference : https://docs.cortext.net/metrics-definitions/#chi2*

Pour les réseaux homogènes

Distributional : issue de la linguistique récente, elle permet de faire apparaître des relations pertinentes, bien que rares. La proximité distributional s'appuie sur la mesure directe d'Information Mutuelle précédemment présentée. Pour un mot donné, l'Information Mutuelle est la quantité d'information apportée par la présence de ce mot dans le contexte d'apparition d'un autre mot. La mesure Distributional, pour deux termes (i et j), compare donc les vecteurs à n dimensions d'Information Mutuelle de ces deux termes, autrement dit la similarité des contextes d'apparition de ces termes. Cela permet de détecter des synonymes, c'est-à-dire des termes qui ne cooccurrent pas forcement mais qui ont des contextes d'apparition identiques. Elle a donc une propriété d'interchangeabilité : carotte et porreau étant des légumes, ils ont des caractéristiques communes, des possibilités d'associations avec d'autres ingrédients proches ainsi que des modalités de cuissons similaires.

- est généralement à privilège pour **l'analyse sémantique** : très performant pour extraire la structures sous-jacentes des textes, en présentant les mots qui jouent des fonctions similaires dans les textes
- *reference: https://docs.cortext.net/metrics-definitions/#distributional*

Cosinus : la mesure de similarité cosinus a été introduite par Salton (Gerard Salton & McGill, 1983) avec l'idée que la similarité entre deux documents peut se mesurer par la comparaison des deux vecteurs (G Salton, Wong, & Yang, 1975) formés par la liste de termes et des fréquences de ces termes pour ces documents. Cette mesure est couramment utilisée dans la fouille de données textuelles. Elle est également populaire en scientométrie, et plus particulièrement dans l'analyse des co-citations (Eck & Waltman, 2009; Hamers et al., 1989) . Appliquée à un graph, la mesure de similarité cosinus vise à comparer les profils de cooccurrences de deux nœuds (mot, citation, auteur…) dans un ensemble de documents. Pour deux termes i et j , leur similarité s'appuie sur les deux lignes des termes de la matrice initiale des cooccurrences, c'est-à-dire sur l'ensemble des cooccurrences de ces deux termes. Les profils de cooccurrences des deux termes sont traités comme des vecteurs à n dimensions (où n correspond au nombre total de termes avec lesquels ils cooccurrent) dont on mesure l'angle. Plus l'angle est petit, plus les profils sont similaires (plus donc les deux listes de termes avec lesquels ils cooccurrent, ainsi que les fréquences associées, sont proches).

- Avantage de cette mesure : ne privilégie pas les termes fréquents (nœuds importants) aux termes rares. Ainsi, deux termes peuvent avoir une proximité élevée même s'ils cooccurrent peu ensemble ou peu avec d'autres termes.
- *Référence : https://docs.cortext.net/metrics-definitions/#cosine*

Pour les réseaux hétérogènes

cosine_het	affiliation network (eg. users sharing the same hashtags)	yes	two fields are required but the final network is homogeneous
dot_product_het	affiliation network (eg. users sharing the same hashtags)	no	two fields are required but the final network is homogeneous

- *https://docs.cortext.net/metrics-definitions/#heterogeneous-dot-product*
- *https://docs.cortext.net/metrics-definitions/#heterogeneous-cosine*

4/ Protocole méthodologique et résultats

- Reportez dans un document vos étapes, et ajouter vos résultats.
- Comparez avec les propositions de résultats : *https://docs.cortext.net/trainings/cortext-lisis/2-methodes/03-resultats/*