Comment utiliser CorText

Rédacteurs :
Frédérique Becquet

Pablo Ruiz

Comment utiliser CorText

Frédérique Mélanie[1] Pablo Ruiz[2]

(1) Ingénieur d'études au Lattice (Langues, Textes, Traitements informatiques et Cognition)
(2) Doctorant au Lattice (Langues, Textes, Traitements informatiques et Cognition)

januar 2017

Summary

Introduction

Ce document s'adresse aux personnes qui n'ont pas l'habitude d'utiliser Cortext ou autre outil d'analyse ou visualisation de corpus. Ce tutoriel décrit étape par étape la création d'un projet dans CorText : comment effectuer des extractions lexicales, obtenir des visualisations de corpus.

Nous attirons l'attention du lecteur sur le fait que ce document est un guide de prise en main de l'outil. Il y trouvera des descriptions de notre utilisation du logiciel, des exemples issus de notre expérience personnelle, des modèles de corpus utilisés - formats d'import et export. Ce document ne contient pas d'analyse de corpus, d'explications détaillées des visualisations obtenues.

Nous invitons le lecteur à parfaire sa connaissance de l'outil en consultant la documentation en ligne : https://docs.cortext.net/. Quand nécessaire, nous nous référons à cette documentation.

L'URL de CorText est https://managerv2.cortext.net/. *CorText Manager v2* est en accès libre, mais il est nécessaire de se créer un compte pour pouvoir l'utiliser. La première fois que vous allez sur la plateforme cliquez sur « Subcribe » (Figure 1) puis sur « Register » (Figure 2), afin de créer et enregistrer votre profil utilisateur.

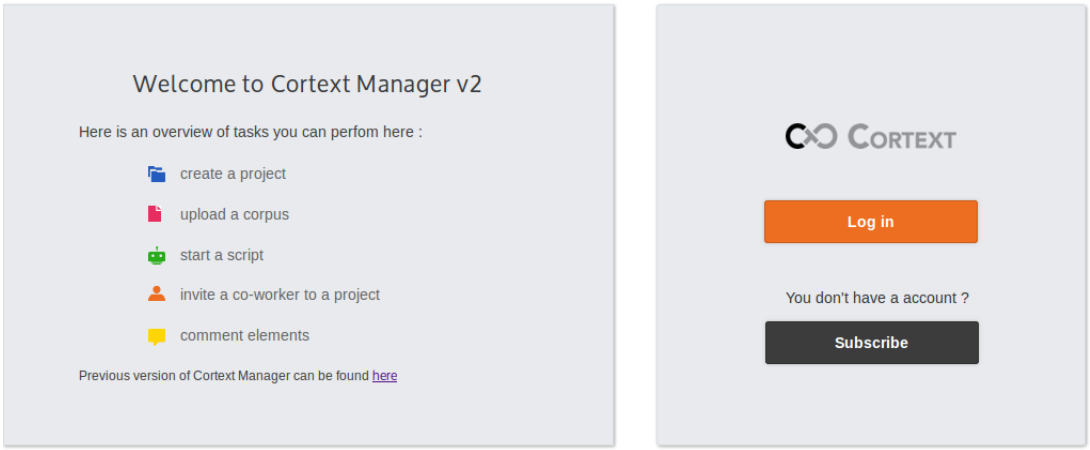


Figure 1: Page de lancement

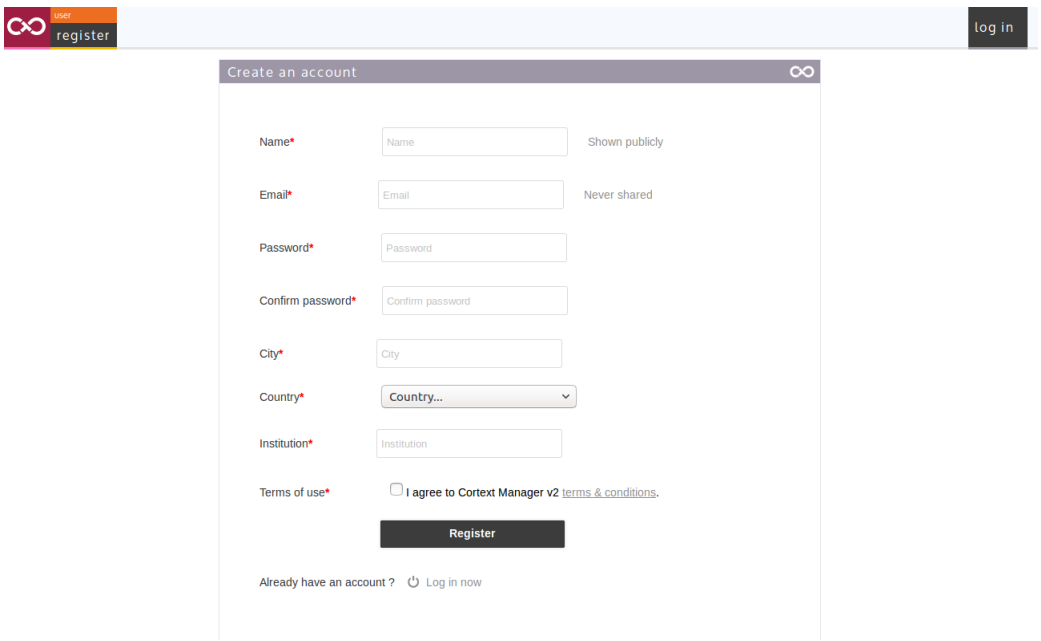


Figure 2: Créer un compte

Maintenant nous pouvons commencer!

1 Prise en main de l'outil

Cortext permet 3 types de tâches : l'exploration de corpus, l'extraction de termes et la construction de réseaux.

- l'exploration de corpus (*explor the corpus*)
  Quand vos données sont téléchargées dans Cortext, il est possible de visualiser les données, faire des requêtes sur les données.
- l'extraction de termes (*extract terms*).
  Dans Cortext, il est possible d'extraire un lexique à partir des données téléchargées, selon un certain nombre de critères. Tout lexique ainsi constitué doit être indexé au texte pour permettre la constitution d'un réseau.
  A noter, Cortext permet l'indexation de tout lexique, c'est-à-dire qu'il permet l'indexation de lexiques constitués en dehors de l'outil.
- construction de réseaux (*build network*)
  Il s'agit de visualiser selon un certain nombre de critères statistiques, les termes qui sont proches, de les représenter graphiquement.

La gestion d'un projet dans Cortext comporte 3 étapes :

1.1 Créer un projet

Pour créer un projet, il suffit d'entrer le nom du projet dans la case *Type the name of your new project*, à la place du texte.

☝ Click on *Create project*

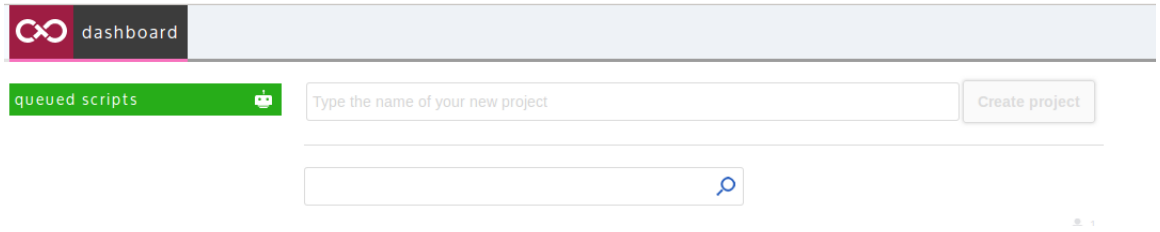


Figure 3: Page d'accueil (initiale et vierge)

La page d'accueil contient maintenant le nom du projet (Figure 4).

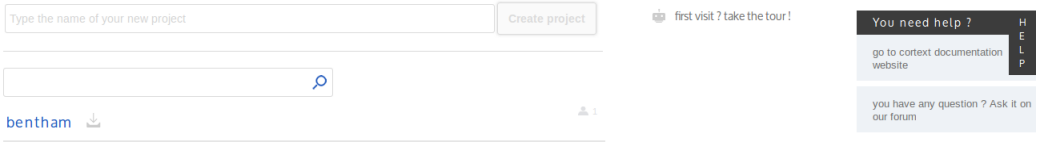


Figure 4: Page d'accueil (avec projet)

⚠ Une fois le projet créé, il est impossible de le supprimer ou le renommer.

Entrer dans le projet

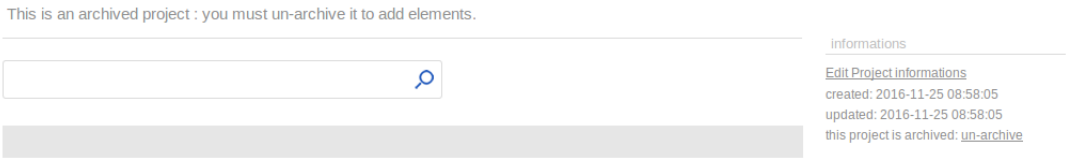
En cliquant sur le nom du projet - dans notre exemple Bentham, on entre dans le projet. On peut travailler, enrichir le projet. A ce stade de la gestion du projet, l'étape suivante serait le téléchargement du corpus (voir le chapitre 1.2).

Archiver le projet

En cliquant sur cette icône , le projet est archivé.

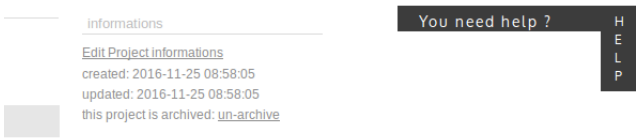
- une étoile apparaît à côté du nom du projet

- Quand le projet est archivé, son contenu est bloqué. Il est impossible de modifier un élément du projet. Pour pouvoir à nouveau faire une modification, il faut désarchiver (cliquer sur *unarchive*).



Demander de l'aide

Pour ce faire, cliquer sur HELP.



Cliquer sur You need help? donne accès :

- à la documentation en ligne https://docs.cortext.net/. Notre tutoriel réfère au besoin à cette documentation en ligne.
- au forum Cortext

1.2 Télécharger des données

La page d'accueil d'un projet contient 4 rectangles - comme le montre la Figure 5.

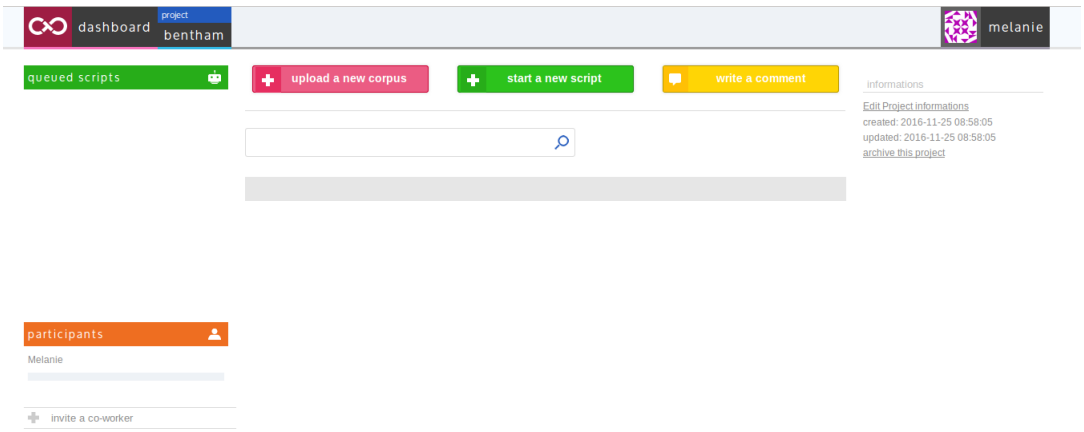


Figure 5: Page d'accueil du projet

- (a) file d'attente de scripts (*queued script*): à la première connexion, la liste est vide.
- Les 2 rectangles suivants réfèrent à 2 types de processus:
  - (b) télécharger un nouveau corpus (*upload a new corpus*)
  - (c) lancer un script (*start a new script*)
- (d) ajouter un commentaire (*write a comment*): il est possible d'ajouter un commentaire soit à l'ensemble du projet, soit à une tâche donnée, à un script.

Après avoir créé le projet, l'étape suivante est bien évidemment le téléchargement du corpus. Si aucun corpus n'est téléchargé, aucune tâche ne peut être lancée. Si malgré tout vous essayez de lancer un script, un message d'alerte vous invite à télécharger un corpus (*you don't have any corpus yet* cf. Figure 6).

☝ Cliquer sur télécharger un nouveau corpus (*upload a new corpus*).

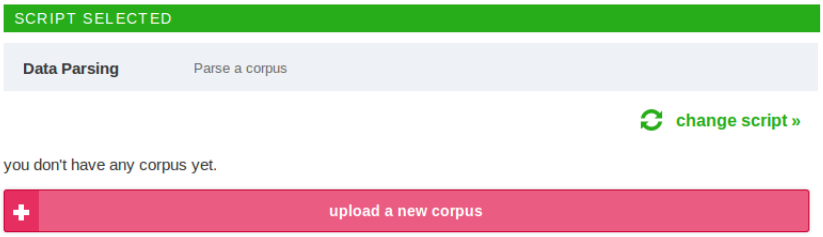


Figure 6: Télécharger un corpus (message d'alerte)

Les formats d'import

L'ensemble des formats qu'il est possible d'importer dans Cortext est défini à cette adresse : https://docs.cortext.net/upload-corpus/

Dans l'exemple suivant, nous utilisons le format csv [1]. Notre corpus est constitué d'environ 29000 fichiers. Chaque fichier est formaté selon le modèle de la Figure 7.

1. Pour être précis, il s'agit ici du format tsv. En effet, les colonnes sont délimitées par une tabulation - comme l'indique le *t* plutôt que par une virgule - comme l'indiquerait le *c* mis pour *commas* dans l'abréviation csv

- La première ligne contient le nom des champs. La seconde contient les données. L'ensemble du fichier est donc contenu sur une seule et unique ligne.
- Les colonnes sont séparées par une tabulation.
- Chaque passage à la ligne est matérialisée par 3 étoiles. [2]

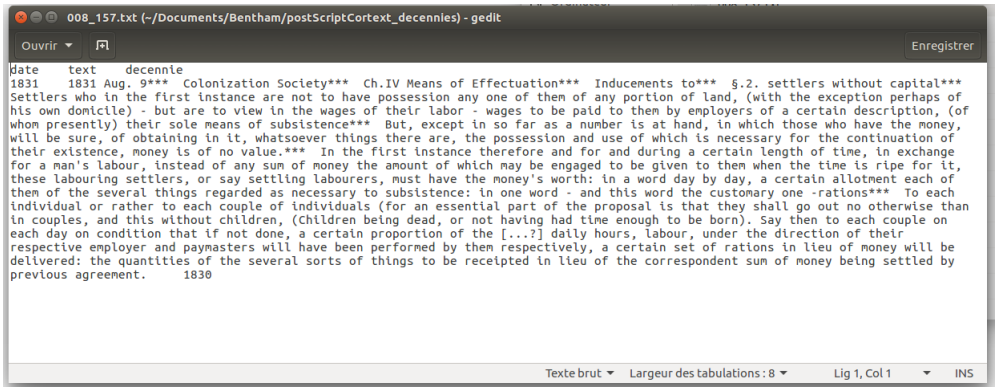


Figure 7: Exemple de format csv

Déposer ou télécharger des données

Concrètement deux solutions sont disponibles pour importer les données :

(a) glisser déposer un document (*drag and drop a document*) (Figure 8).

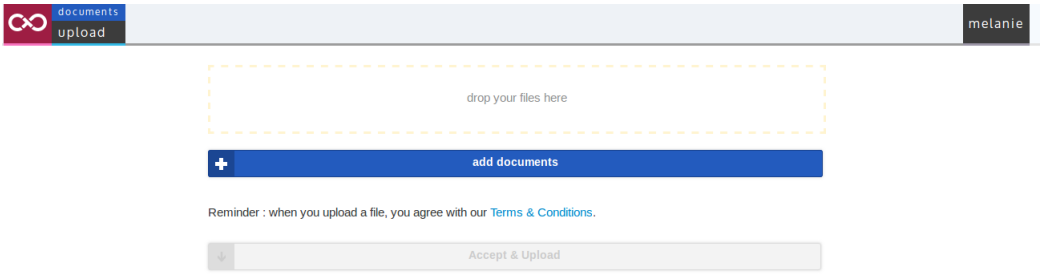


Figure 8: Télécharger un document (glisser et déposer)

Comme l'indique la fonction, il est possible de glisser un fichier depuis l'interface de son ordinateur dans l'interface de Cortext (à l'endroit indiqué : *drop your files here*). [3]

(b) télécharger un document depuis la fenêtre d'exploration de votre ordinateur. Il faut choisie un à un les fichiers à télécharger.

2. A l'importation, CorText considère chaque ligne comme un enregistrement de la base de donnée. C'est pourquoi le contenu textuel doit être une seule et même ligne. Les sauts de lignes dans le texte sont quand à eux matérialisés par un caractère (ou ensemble de caractères) discriminant. Par défaut, il s'agit de 3 étoiles, mais il est possible à l'importation de choisir un autre signe discriminant (cf. p.10 Figure 13).

3. Il est impossible de *glisser déposer* un dossier. Si vous essayer de le faire ce message apparaît : *your folder is empty, please select files again without it.*

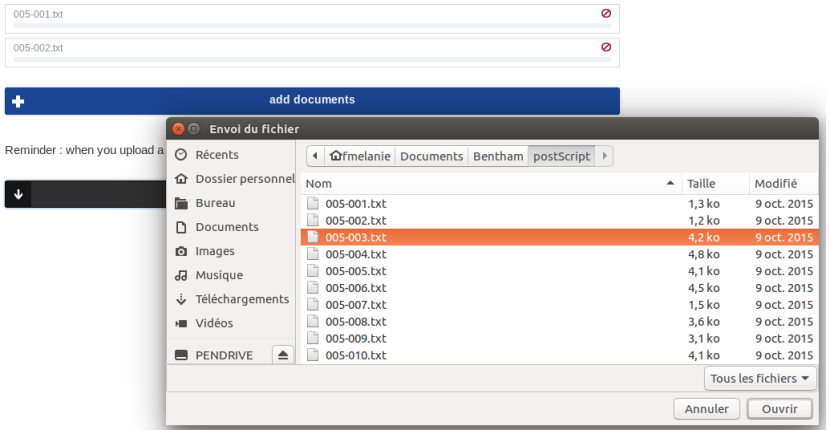


Figure 9: Télécharger un document (importer un fichier)

⚠ Il est possible d'importer en une fois l'ensemble des fichiers d'un répertoire, pour cela il suffit de zipper le répertoire. Dans notre exemple, notre dossier .zip (Figure 9) contient environ 29000 fichiers (.txt).

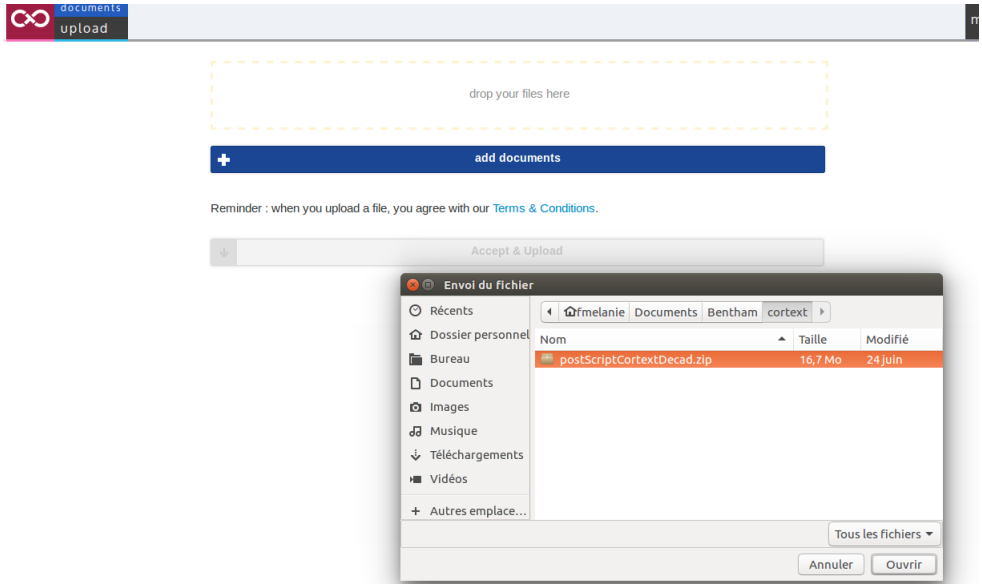


Figure 10: Télécharger un document (importer un dossier)

👆 Quand l'ensemble des fichiers à analyser est glissé ou téléchargé dans l'interface de l'outil, il suffit de cliquer sur *accepter et importer* (*accept and dowload*). Le processus peut être plus ou moins long.

1.3 Parser les données

Quand les données sont téléchargées, un nouveau formulaire s'ouvre. Il invite l'utilisateur à lancer le script *Data Parsing*.

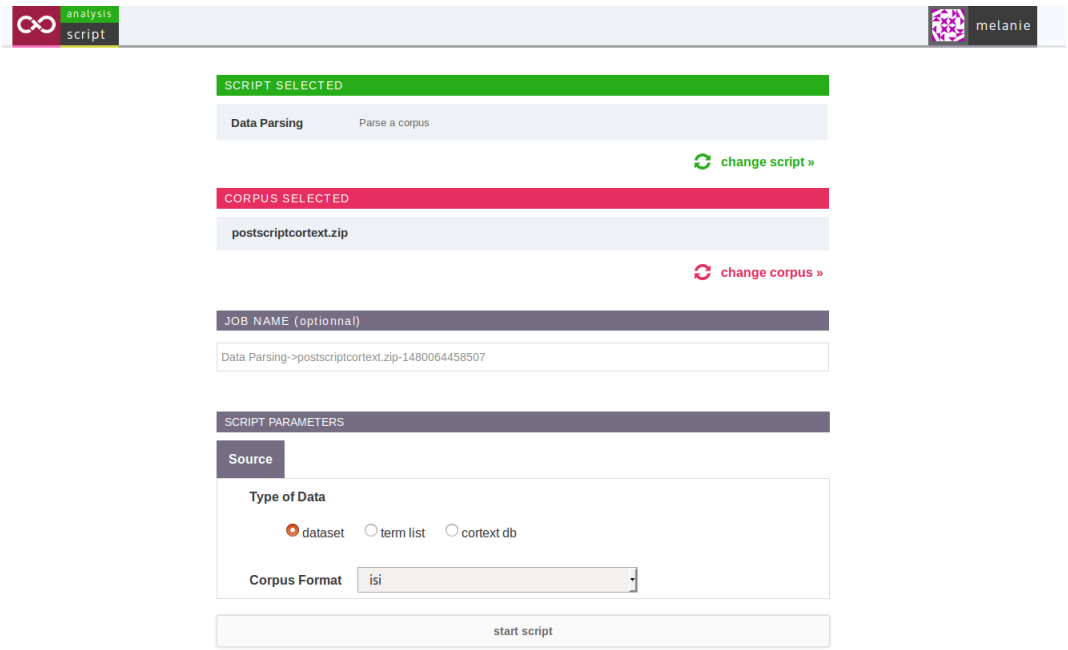


Figure 11: Data parsing

- Script sélectionné (*Script selected*): parser les données (*Data parsing*)
  Dans la gestion d'un projet, ce script est le premier des scripts qu'il faut obligatoirement lancer.
- Corpus sélectionné (*Corpus selected*):
  Il s'agit du dossier zippé utilisé pour l'exemple.
- Nom de la tâche (*Job Name*) (optionnel)
  Vous pouvez nommer la tâche que vous allez effectuer. C'est optionnel. Il est préférable de nommer de manière significative l'analyse demandée à Cortext et de ne pas conserver le nom proposé par défaut.
- Paramètres du script (*script parameters*)
  L'ensemble des paramètres fait référence au format des données. Le format est lié à la source des données. Elles peuvent être issues de LexisNexis ou Factiva, d'un fichier excel, ou même de Cortext.

Selon les données importées, il faut modifier leur type et leur format :

(a) **Type de données** (*Type of data*): le type de données importées (texte, liste de termes, base de données CorText).

(b) **Format du Corpus** (*Corpus Format*) [4]

- **txt**: Cette option est à sélectionner quand il s'agit de fichiers *.txt* bruts, sans métadonnées. A un fichier correspond un texte. L'import du fichier dans Cortext consistera à parser le titre du fichier et son contenu. Il est possible si besoin par la suite d'ajouter des métadonnées et de les lier aux fichiers *.txt*. [5]

4. Chacun des formats est décrit dans la documentation en ligne à l'adresse suivante : https://docs.cortext.net/upload-corpus/

5. Il faut alors utiliser le script *corpus_ list_indexer* et utiliser comme clef le titre du fichier. Nous ne décrivons plus précisément ce type de travail dans ce tutoriel

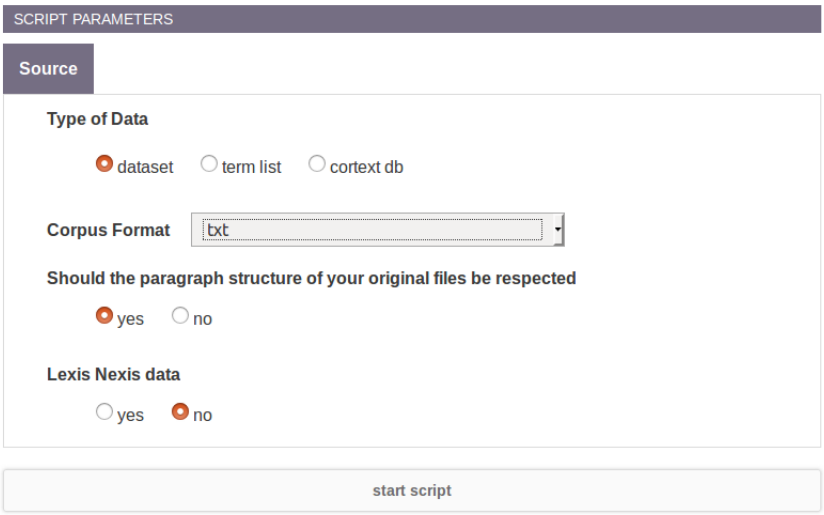


Figure 12: Les paramètres d'import du format txt

– *CSV* est un type particulier de fichier texte[6]. La Figure 7 est un exemple de fichier csv accepté par Cortext (le fichier *008_157.txt*).

Les caractéristiques du fichier csv décrites précédemment (1.2) sont à renseigner dans le formulaire au moment de l'importation du fichier (Figure 13). A l'utilisateur de spécifier le séparateur de colonnes ainsi que ce qui dans le texte désigne un changement de paragraphe.

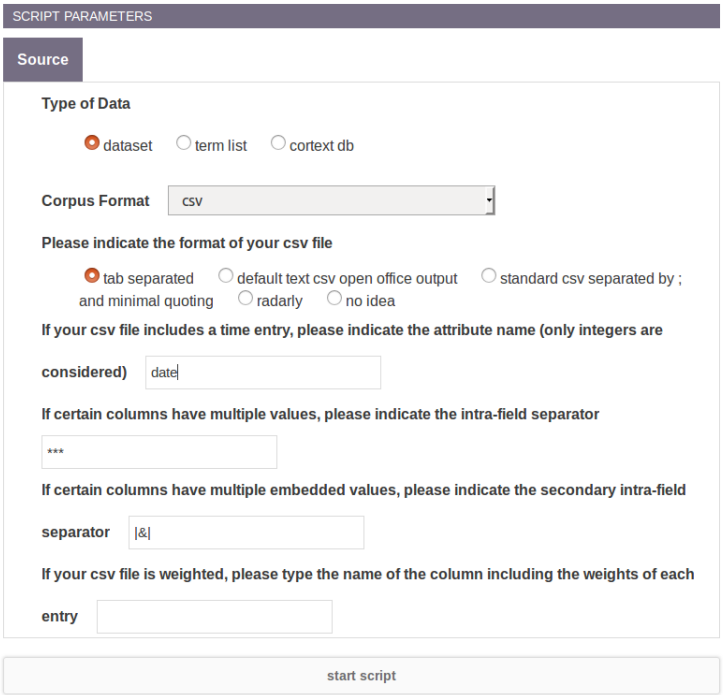


Figure 13: Les paramètres d'import du format csv

- les colonnes sont séparées par des tabulations : **tab-separated**. Notre fichier se compose de 3

6. Un fichier peut avoir une extension *txt* et être importé comme fichier *csv*. C'est la structuration interne du fichier qui importe et non son extension.

colonnes : `date`, `text`, `decennie`.

- les délimitations de paragraphe (ou passage à la ligne) sont les 3 étoiles : **paragraph delimiter**. 3 étoiles sont insérées à chaque changement de paragraphe. Comme nous l'expliquions ultérieurement (section 1.2), il est nécessaire pour Cortext que le corpus soit égal à une ligne. En effet, pour Cortext chaque ligne de données est un enregistrement de la base de données qu'il constitue.

☝ Une fois l'ensemble des paramètres du script sélectionné, cliquer sur *start script*.

Pendant le traitement, il est possible de cliquer sur (running) pour obtenir des informations sur le processus en cours. L'ensemble du processus est retranscrit dans un fichier *log*. Nous donnons ici un aperçu de ce fichier : le début du traitement (cf.Figure 14) et la fin du traitement (cf. Figure 15). Dans cette dernière, un message indique que le processus s'est déroulé correctement : *Parsing ended successfully*.

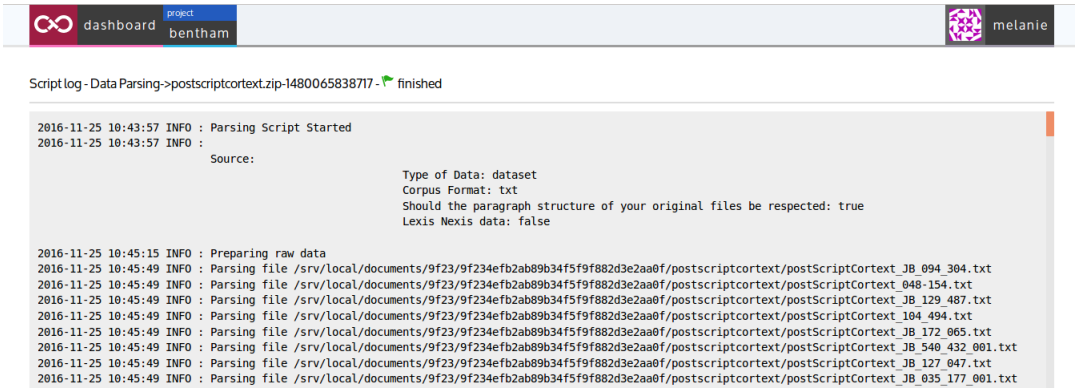


Figure 14: Fichier log (début du processus)

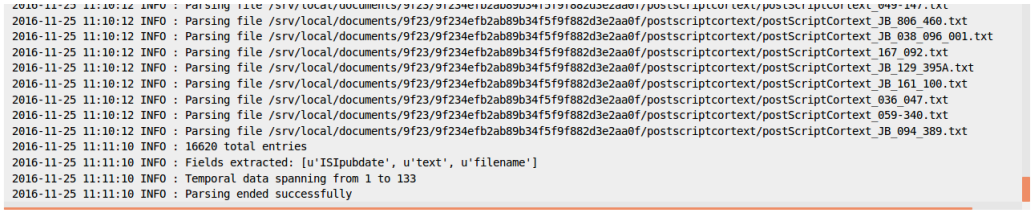
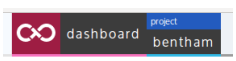


Figure 15: Fichier log (fin du processus)

⚠ Le traitement peut être long, il dépend de la masse de données à analyser ainsi que du débit web disponible. Pour notre test, le traitement a pris environ 25 minutes.

1.4 Les icônes présentes sur la plateforme

Pour retourner sur la page d'accueil, cliquer sur le nom de votre projet en haut de l'écran.



Quand un traitement est terminé - et ce quelque soit le script qui a été lancé, plusieurs types d'informations sont à la disposition de l'utilisateur comme l'illustre les copies d'écran pour les 2 résultats suivants : une représentation de réseau (*network mapping* Figure 16), une exploration de corpus (*corpus explorer* Figure 17). Nous reviendrons ultérieurement plus précisément sur le contenu des dossiers résultats.

Figure 16: Résultat d'une tâche (*Network mapping*)

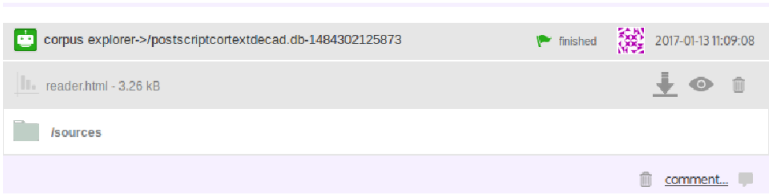


Figure 17: Résultat d'une tâche (*Corpus explorer*)

Les informations disponibles sont :

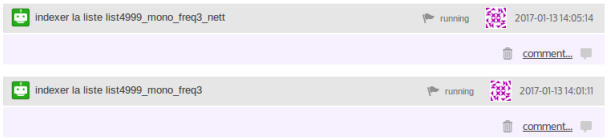
- (a) le nom de la tâche.
- (b) le statut du traitement.

  Il s'agit de savoir si le traitement a abouti ou non. Plusieurs icônes sont utilisées :

  – en cours de traitement (*running*)
  – traitement terminé avec succès (*finished*)
  – traitement avorté (*error*)

  Dans les 2 copies d'écran (Figure 16 et Figure 17), le processus s'est terminé avec succès. En cliquant sur le drapeau vert (l'icône *finished*), vous avez accès au fichier log. Quand le processus s'est terminé avec succès, retourner au fichier log peut s'avérer nécessaire et utile pour vérifier les paramètres choisis pour effectuer une tâche, comprendre précisément le processus. Quand le processus a été avorté, retourner au fichier log peut s'avérer utile pour déterminer les raisons de l'échec du traitement.

  ⚠ Il est possible de lancer plusieurs scripts en même temps.



- (c) **les données personnelles** .

Cliquer sur l'(*avatar*) donne accès au compte personnel, aux données personnelles, celles qui ont été entrées lors de la création du compte utilisateur.

- (d) **télécharger le réseau**(*Download network*)
- (e) **voir le fichier**(*View file*) Cliquez sur l'icône (l'oeil) pour avoir accès aux résultats.
- (f) **supprimer le résultat**(*Delete*)

2 Description globale des scripts de la plateforme

Dans les pages suivantes, nous décrivons les différents types de scripts qu'il est possible de lancer sur la plateforme Cortext [7]. Le menu **sélectionner un script**(*Select a script*) est divisé en 4 sections :

1. *Corpus*
2. *Text*
3. *Time*
4. *Analysis*

Quelques remarques générales à propos des tâches lancées sur la plateforme :

Scripts et données:
Chaque processus nécessite *le choix d'un script* et *le choix des données*. Un script s'applique à lot de données.

Effacer et commenter:
A tout moment, il est possible de *supprimer* ou *commenter* un élément du projet (respectivement repéré par (a) et (b) dans la Figure 18).
Il est possible de créer un ou plusieurs commentaires. Il s'agit d'une zone de texte libre. Une fois le commentaire créé et enregistré, il est possible de le modifier mais impossible de le supprimer.
Les commentaires sont très utiles. L'utilisateur peut y décrire les tâches effectuées, commenter succinctement les résultats, indiquer les paramètres choisis pour un processus et les raisons de ce choix, ...

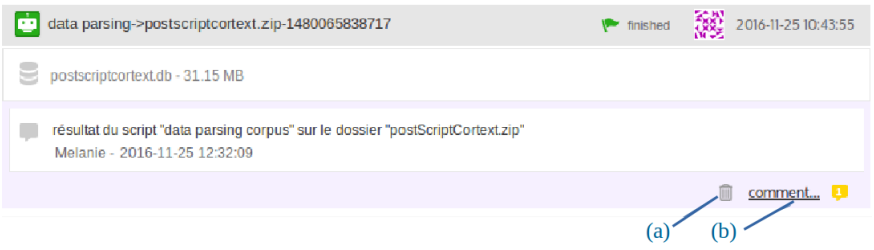


Figure 18: Fonctions *supprimer* et *commenter*

2.1 Corpus

La section *Corpus* contient 5 scripts.

7. Nous laissons dans nos titres de chapitre le nom anglais du script, tel qu'il apparaît sur la plateforme. Il est ainsi plus facile au lecteur de se repérer sur la plateforme. Les chapitres descriptifs pour chacun des scripts donne lieu non seulement à l'explication du script mais aussi à une traduction possible de sa dénomination.

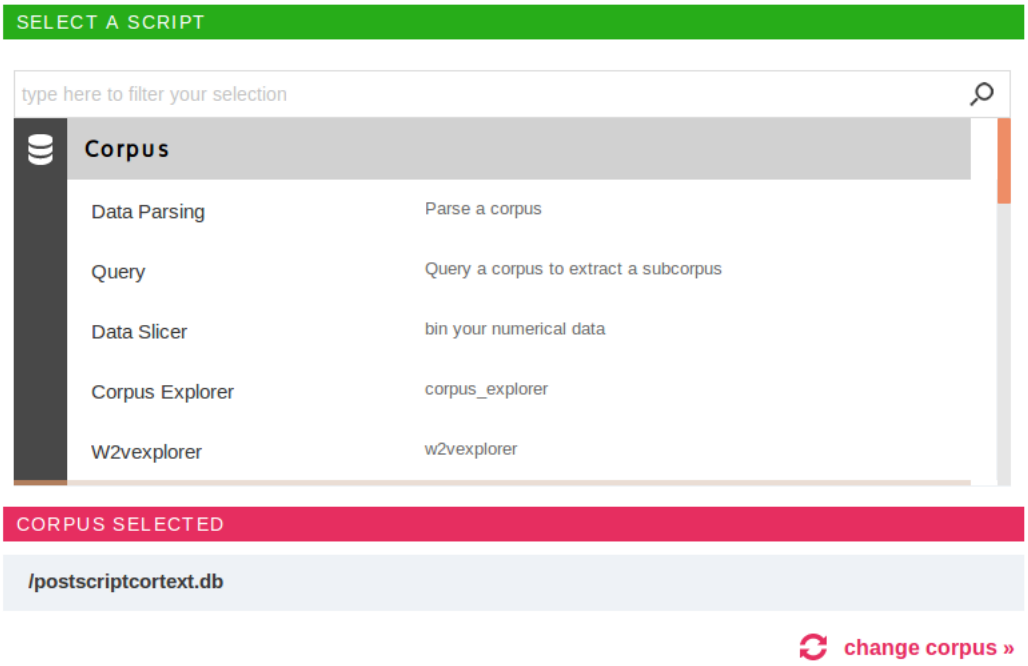


Figure 19: La section *Corpus*

Data parsing

Le script *analyse de données* (*data parsing*) est une **étape obligatoire et nécessaire au début de chaque projet**, c'est pourquoi vous êtes automatiquement invité à lancer ce script après avoir téléchargé un jeu de données sur la plateforme. Nous avons décrit ce processus en 1.3.
Lors de ce processus, Cortext structure les données en une base de données (extension *.db*). Dans notre exemple, chacune des colonnes du document *csv* devient un champ de la base, le titre de chaque colonne (contenu sur la première ligne) devient le titre de chacun des champs de la base, le contenu des colonnes (contenu sur la seconde ligne) devient un enregistrement. La base de données créée sera utilisée par Cortext pour tout ajout ou modification d'information, pour toute analyse. [8]

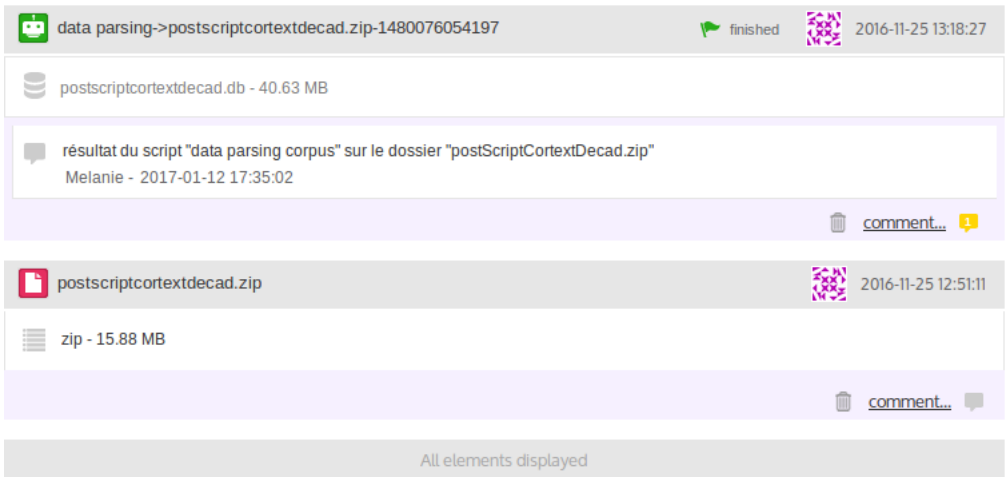


Figure 20: *Data parsing* (résultat)

8. Les copies d'écran - Figure 21 et Figure 23 - illustre le fait que les scripts sont lancés sur la base de données (fichier *.db*) du projet.

Ce n'est qu'après avoir lancé ce script, après avoir créé une base de données dans Cortext, qu'il est possible d'avancer plus avant dans le projet, d'utiliser d'autres scripts de la plateforme.

Query

[...] L'utilisation de ce script n'est pas abordée dans ce tutoriel.

Data slicer

Le script *Data slicer* permet - comme son nom l'indique - de *découper les données* selon un certain nombre de critères. La Figure 21 montre les critères proposés à l'utilisateur.

Sous *data*, l'utilisateur retrouve le nom des champs de sa base de données. Elle est constituée - dans le cas d'un import de type csv - des titres des colonnes du fichier [9], des champs résultants du lancement de scripts Cortext [10].

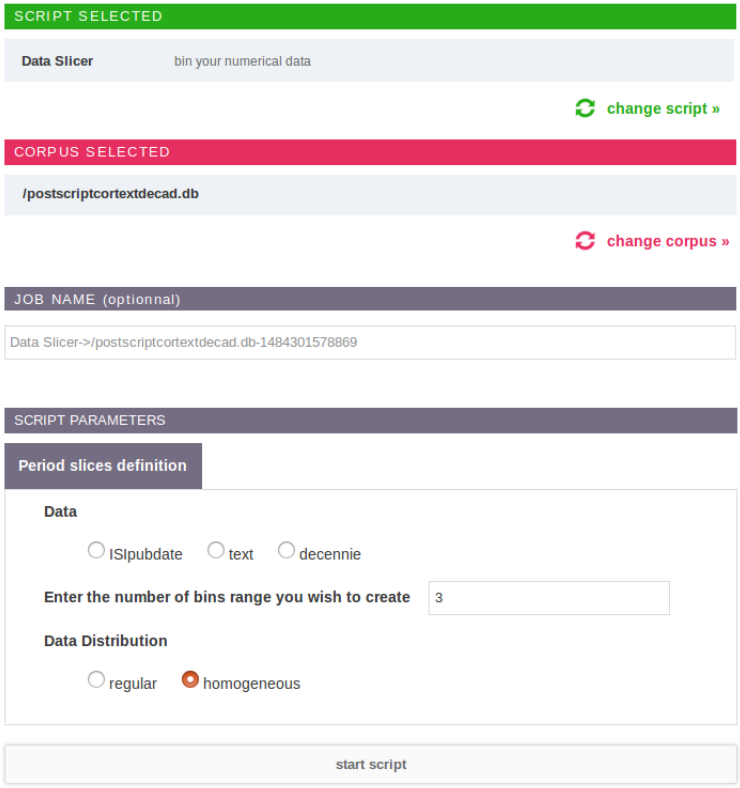


Figure 21: Script : data slicer

Ainsi dans l'exemple de la Figure 21, trois champs peuvent être utilisés pour scinder les données : *ISIpubdate*, *text*, *decennie*. Dans notre exemple, nous avons sélectionné le champ *decennie* et avons choisi de constituer à partir des valeurs de ce champs *3 ensembles* de distribution *homogène* [11].

9. Si l'utilisateur avait importé un fichier texte, seul le champ *text* serait disponible.
10. Les champs de type ISI* sont produits lors de l'indexation d'un lexique sur le corpus. La mention ISI réfère à un format particulier de la base de données de Cortext.
11. Ce type de distribution - homogène ou régulière - est utilisée dans plusieurs scripts, voire sous la section *Analysing* les paramètres du script *Network mapping* - 2.4

La distribution est :

- *régulière* si chaque période est composée d'un même nombre d'année.
- *homogène* si chaque période est composée d'un même nombre de document.

Le script a pour effet de créer un nouveau champ dans le base de données. Ainsi, dans notre exemple, est créé le champ `decennie-hom-3` comme le montre la Figure 22.[12] Ce champ est dorénavant disponible et peut être utilisé pour les futures analyses.

Figure 22: Création du champ data slicer

Corpus explorer

Ce script permet de visualiser le corpus. Il permet de visualiser les données et métadonnées disponibles sur la plateforme, données sur lesquelles reposent les visualisations. Ce script peut être utile pour vérifier et comprendre les imports de données dans Cortext, les champs constitués.

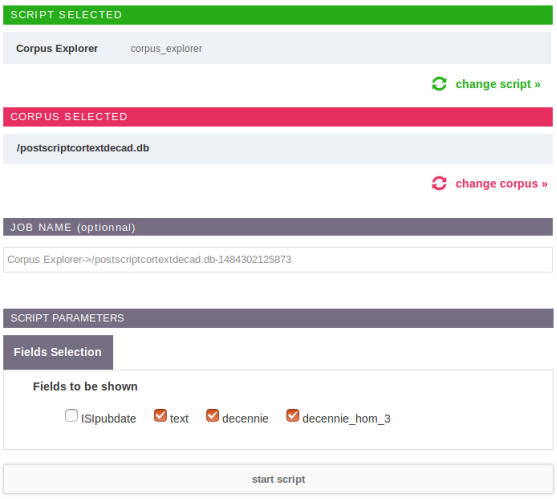


Figure 23: *Corpus explorer*(formulaire)

En lançant le script *Corpus explorer*, vous pouvez visualiser l'ensemble des champs disponibles : ceux que vous avez importé dans Cortext (issu de votre corpus initial) et ceux qui ont été créés dans Cortext - par exemple ici Figure 25 le champ `decennie-hom-3`. Ce champ a été nommé ainsi par l'utilisateur. Un tel intitulé semble indiquer qu'il s'agit d'un découpage homogène (*hom*) en 3 périodes (*3*) du champ *decennie*. Il est le résultat du script *Data slicer*, lancé avec les paramètres de la Figure 21.

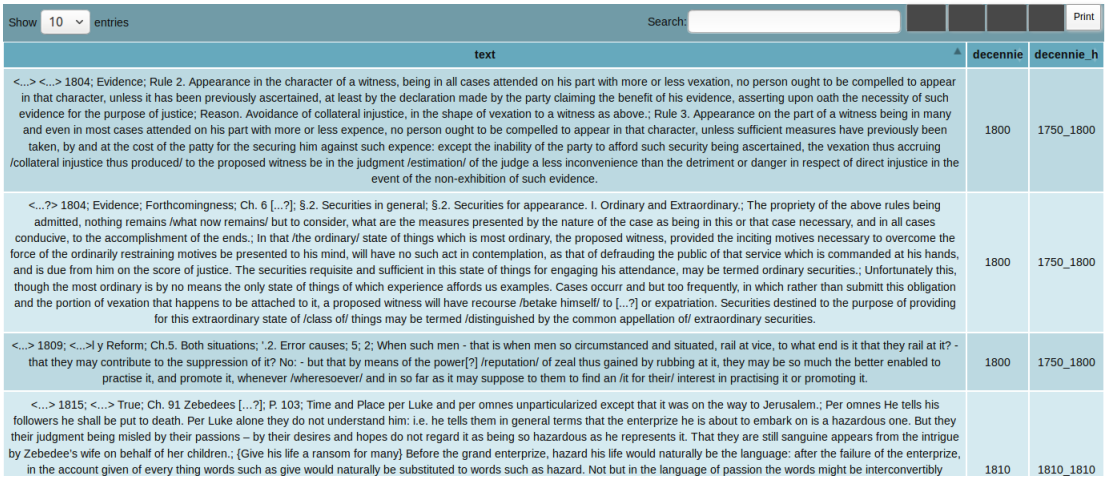
⚠ Nous attirons l'attention de l'utilisateur sur le fait qu'il est impossible de modifier le nom d'un champ ou de supprimer un champ. Supprimer sur la page d'un projet un dossier contenant les résultats d'un script n'entraîne pas la suppression du nom du champ créé lors de l'utilisation du script. Le nom du

12. si l'utilisateur n'a pas choisi de nom, un nom donné par défaut, créé par Cortext

champ est conservé dans la base. Dans notre exemple, si l'on supprime le dossier contenant le résultat du script *Data slicer* (cf. Figure 21), le champ `decennie-hom-3` n'est pas supprimé de la base, il reste disponible dans les formulaires.

Figure 24: *Corpus explorer* (dossier résultat)

En cliquant sur *reader.html* (cf. Figure 24), vous accédez au tableau contenant le corpus (Figure 25). Il est possible d'ordonner par ordre alphabétique ou numérique le contenu de chacune des colonnes en cliquant sur le titre de celle-ci.


Show 10 entries Search: Print

text	decennie	decennie_h
<...> <...> 1804; Evidence; Rule 2. Appearance in the character of a witness, being in all cases attended on his part with more or less vexation, no person ought to be compelled to appear in that character, unless it has been previously ascertained, at least by the declaration made by the party claiming the benefit of his evidence, asserting upon oath the necessity of such evidence for the purpose of justice; Reason. Avoidance of collateral injustice, in the shape of vexation to a witness as above.; Rule 3. Appearance on the part of a witness being in many and even in most cases attended on his part with more or less expence, no person ought to be compelled to appear in that character, unless sufficient measures have previously been taken, by and at the cost of the patty for the securing him against such expence: except the inability of the party to afford such security being ascertained, the vexation thus accruing /collateral injustice thus produced/ to the proposed witness be in the judgment /estimation/ of the judge a less inconvenience than the detriment or danger in respect of direct injustice in the event of the non-exhibition of such evidence.	1800	1750_1800
<...?> 1804; Evidence; Forthcomingness; Ch. 6 [...?]; §.2. Securities in general; §.2. Securities for appearance. I. Ordinary and Extraordinary.; The propriety of the above rules being admitted, nothing remains /what now remains/ but to consider, what are the measures presented by the nature of the case as being in this or that case necessary, and in all cases conducive, to the accomplishment of the ends.; In that /the ordinary/ state of things which is most ordinary, the proposed witness, provided the inciting motives necessary to overcome the force of the ordinarily restraining motives be presented to his mind, will have no such act in contemplation, as that of defrauding the public of that service which is commanded at his hands, and is due from him on the score of justice. The securities requisite and sufficient in this state of things for engaging his attendance, may be termed ordinary securities.; Unfortunately this, though the most ordinary is by no means the only state of things of which experience affords us examples. Cases occurr and but too frequently, in which rather than submitt this obligation and the portion of vexation that happens to be attached to it, a proposed witness will have recourse /betake himself/ to [...?] or expatriation. Securities destined to the purpose of providing for this extraordinary state of /class of/ things may be termed /distinguished by the common appellation of/ extraordinary securities.	1800	1750_1800
<...> 1809; <...>l y Reform; Ch.5. Both situations; '.2. Error causes; 5; 2; When such men - that is when men so circumstanced and situated, rail at vice, to what end is it that they rail at it? - that they may contribute to the suppression of it? No: - but that by means of the power[?] /reputation/ of zeal thus gained by rubbing at it, they may be so much the better enabled to practise it, and promote it, whenever /wheresoever/ and in so far as it may suppose to them to find an /it for their/ interest in practising it or promoting it.	1800	1750_1800
<...> 1815; <...> True; Ch. 91 Zebedees [...?]; P. 103; Time and Place per Luke and per omnes unparticularized except that it was on the way to Jerusalem.; Per omnes He tells his followers he shall be put to death. Per Luke alone they do not understand him: i.e. he tells them in general terms that the enterprize he is about to embark on is a hazardous one. But they their judgment being misled by their passions – by their desires and hopes do not regard it as being so hazardous as he represents it. That they are still sanguine appears from the intrigue by Zebedee's wife on behalf of her children.; {Give his life a ransom for many} Before the grand enterprize, hazard his life would naturally be the language: after the failure of the enterprize, in the account given of every thing words such as give would naturally be substituted to words such as hazard. Not but in the language of passion the words might be interconvertibly	1810	1810_1810

Figure 25: *Corpus explorer* (résultat du script)

w2vexplorer

[...] L'utilisation de ce script n'est pas abordée dans ce tutoriel..

2.2 Text

Cette section regroupe des scripts qui permettent à l'utilisateur d'agir sur ses données textuelles. La première des étapes est l'*extraction de termes*, la seconde l'*indexation de termes*. De la première découle la seconde. Cependant...

- la première étape n'est pas obligatoire. L'utilisateur peut en effet disposer de sa propre liste de termes et ne pas vouloir passer par Cortext pour constituer un lexique. Nous abordons ce point ultérieurement (page 25 dans le paragraphe 2.2).
- la seconde étape n'est pas obligatoire. Mais pourquoi alors utiliser Cortext? Le principe même de l'outil repose sur l'indexation d'un lexique sur un texte.

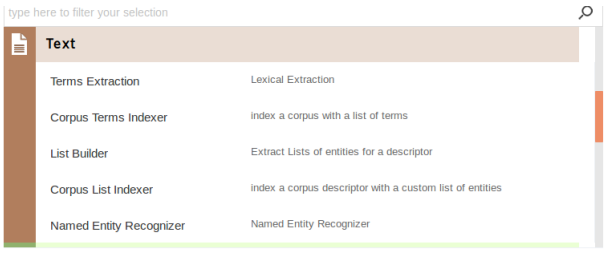


Figure 26: La section *Text*

Terms Extraction

CorText extrait les termes d'un texte. Il utilise un certain nombre de paramètres (Figure 27). Pour une première utilisation de Cortext, il est possible de laisser l'ensemble des paramètres par défaut. L'ensemble des paramètres est disposé sous 2 onglets : *lexical extraction parameters* et *dynamics*.

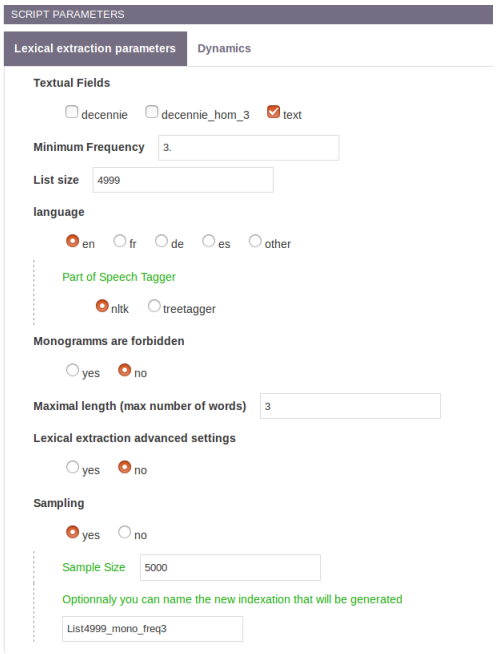


Figure 27: Script terms extraction

Observons tout d'abord les paramètres de l'*extraction lexicale parameters* disposés sous le premier onglet :

- Champs textuels (*Textual fields*)
  Sélectionner le (ou les) champ(s) qui contien(ne)t le texte dont il faut extraire les termes.
- Fréquence minimum (*Minimum Frequency*)
  Définir la fréquence minimum du terme dans le corpus. Cortext utilise la *c-value*, et non la fréquence brute. La *c-value* est une approche statistique lexicale qui prend en compte les informations statistiques associées aux termes, mesure l'indépendance des termes et privilégie les termes longs qui ne sont pas des composants d'autres termes. La c-value favorise les termes candidats n'apparaissant pas dans des termes plus longs.
- Taille de la liste (*List size*)

- La langue (*Langage*)
  Sélectionner la langue de vos données.
  Par défaut le texte est étiqueté morpho-syntaxiquement - utilsation d'étiquette POS (Part of Speech) - avec ***Treetagger***. Si l'utilisateur est davantage familier à *nltk* et Python, il peut sélectionner ce parseur.

- Exclure les monogrammes (*Monogramm are forbidden*)

  Si *yes* est sélectionné : seuls les multi termes sont pris en compte. *Yes* est le critère sélectionné par défaut.
  Si *no* est sélectionné : les monogrammes sont conservés pour l'analyse.

- Paramètres avancés de l'extraction lexicale (*Lexical extraction advanced setting*)

  Par défaut, les calculs statistiques se font au niveau de la phrase, selon le calcul du chi2, sur un corpus brut (sans repérage de chunk ou groupes de mots) et en incluant les phrases nominales (c'est-à-dire les phrases averbales) comme l'illustre la Figure 28.

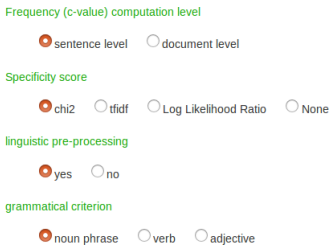


Figure 28: *Terms extraction* (extraction lexicale et paramètres par défaut)

  En cliquant sur *yes*, il est possible de modifier ces paramètres.

- Échantillonnage (*Sampling*)

  Si vous avez une grande masse de données, vous êtes invité à échantillonner l'ensemble de vos données. L'extraction des termes sera basée sur les sous-corpus d'échantillonnage (selon un nombre de documents tirés au hasard parmi votre corpus d'origine). Néanmoins, les termes détectés seront indexés sur l'ensemble du corpus quelle que soit la stratégie adoptée.

- Nommer l'indexation (*Optionally you can name the new indexation*)

  Nous vous recommandons de nommer votre extraction de manière explicite. Cela est important et nécessaire pour les raisons suivantes :

  – si vous faites différentes extractions, selon des paramètres variés, il est important que le nom de votre extraction soit claire et explicite pour que vous puissiez à tout moment savoir le contenu de cette extraction en vue de son utilisation.

  – si pour diverses raisons vous avez à reprendre ce travail après une durée plus ou moins longue.

  Ce nom sera le nom d'un nouveau champ de votre base.

  Si vous ne renseignez pas ce champ, Cortext nommera automatiquement le champ *Term*, puis *Term2*, *Term3* ... pour les extractions successives.

Exemple: List4999-mono-freq3 (peut être le nom d'un champ retenant 4999 termes de fréquence minimum 3 et qui inclut les monogrammes)

Sous l'onglet *Dynamics* les paramètres par défaut sont ceux de la Figure 29. Nous choisissons de ne pas les modifier.

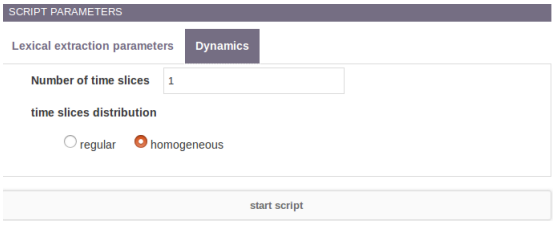


Figure 29: *Terms extraction* (onglet *Dynamics*)

☝ Quand l'ensemble des paramètres du script est sélectionné, cliquer sur *start script*.

Le résultat du script est placé dans un dossier sur la plateforme (Figure 30). *eXtracTterm-List4999-mono-freq3* est le nom de la liste que nous avons constituée et le nom du nouveau champ inséré dans la base de données.

Le dossier résultat contient :

- un fichier csv
- deux dossiers : *indexed list* et *lexical analysis*
  ... sur lesquels nous ne nous attarderons pas pour l'instant.

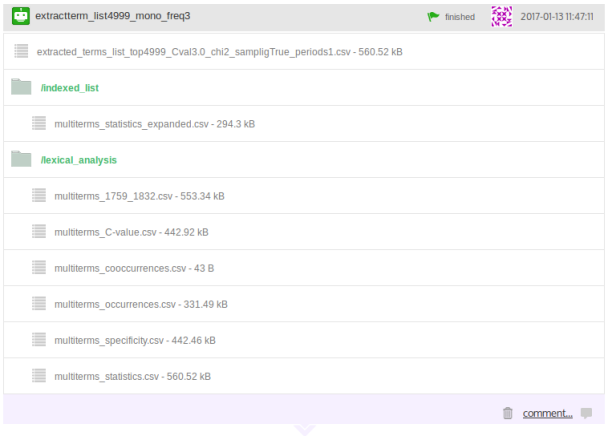


Figure 30: Terms extraction (Dossiers résultats)

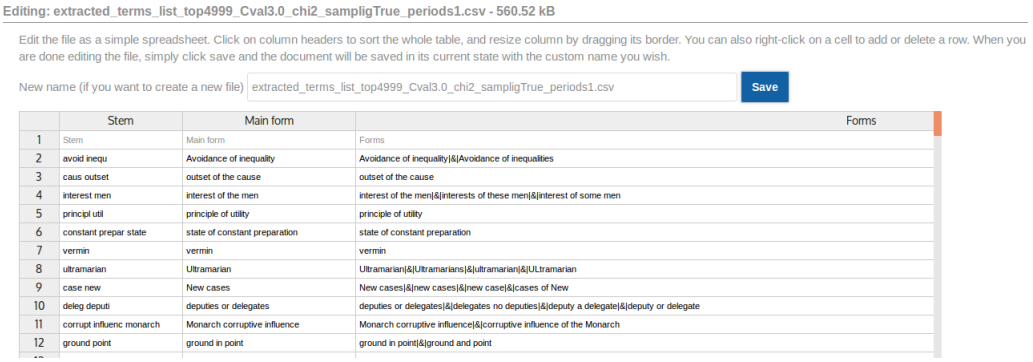
⚠ Pour étendre le menu, il suffit de cliquer sur l'icône dossier , pour réduire à nouveau le contenu du dossier, cliquer sur la flèche mauve au centre de l'image, en bas .

Il est possible d'éditer le fichier csv en cliquant sur *extracted-terms-list-top4999-mono-freq3.csv* .

Les trois premiers champs de cette table sont le lexique, les champs suivants étant les diverses valeurs statistiques calculées par Cortext pour chacun des mots du lexique.

Observons les 3 premiers champs de cette table : *Stem*, *Main Form*, *Form*. Si le titre des colonnes n'est pas important (ce sont les noms donnés par défaut par Cortext, ils peuvent être modifiés sans que cela nuise au bon fonctionnement des traitements), notons en revanche que l'ordre des colonnes est important. C'est sur cette ordre que s'appuie Cortext quand il est appelé à réutiliser la liste :

- colonne 1 (*Stem*) :
  Stem est mis pour *racine du mot*. Chacun des mots présents dans cette colonne doit être unique. La colonne est la clef de la table dans la base de données.
- colonne 2 (*Main Form*) :
  Les mots de cette colonne seront les étiquettes utilisées dans les représentations.
- colonne 3 (*Form*) :
  Les mots de cette colonne sont les mots du texte. Le signe /&/ permet de séparer les variations d'un même terme. A ces variations correspond un seul stem et une seule forme principale.


Editing: extracted_terms_list_top4999_Cval3.0_chi2_sampligTrue_periods1.csv - 560.52 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_top4999_Cval3.0_chi2_sampligTrue_periods1.csv Save

	Stem	Main form	Forms
1	Stem	Main form	Forms
2	avoid inequ	Avoidance of inequality	Avoidance of inequality|&|Avoidance of inequalities
3	caus outset	outset of the cause	outset of the cause
4	interest men	interest of the men	interest of the men|&|interests of these men|&|interest of some men
5	princip util	principle of utility	principle of utility
6	constant prepar state	state of constant preparation	state of constant preparation
7	vermin	vermin	vermin
8	ultramarian	Ultramarian	Ultramarian|&|Ultramarians|&|ultramarian|&|ULtramarian
9	case new	New cases	New cases|&|new cases|&|new case|&|cases of New
10	deleg deputi	deputies or delegates	deputies or delegates|&|delegates no deputies|&|deputy a delegate|&|deputy or delegate
11	corrupt influenc monarch	Monarch corruptive influence	Monarch corruptive influence|&|corruptive influence of the Monarch
12	ground point	ground in point	ground in point|&|ground and point

Figure 31: Liste csv des termes extraits

Les listes sont utilisées par le script d'indexation des termes sur le corpus (*Corpus Terms Indexer*) : les mots à indexer sont les mots de la colonne 3, les étiquettes sont les termes de la colonne 2.

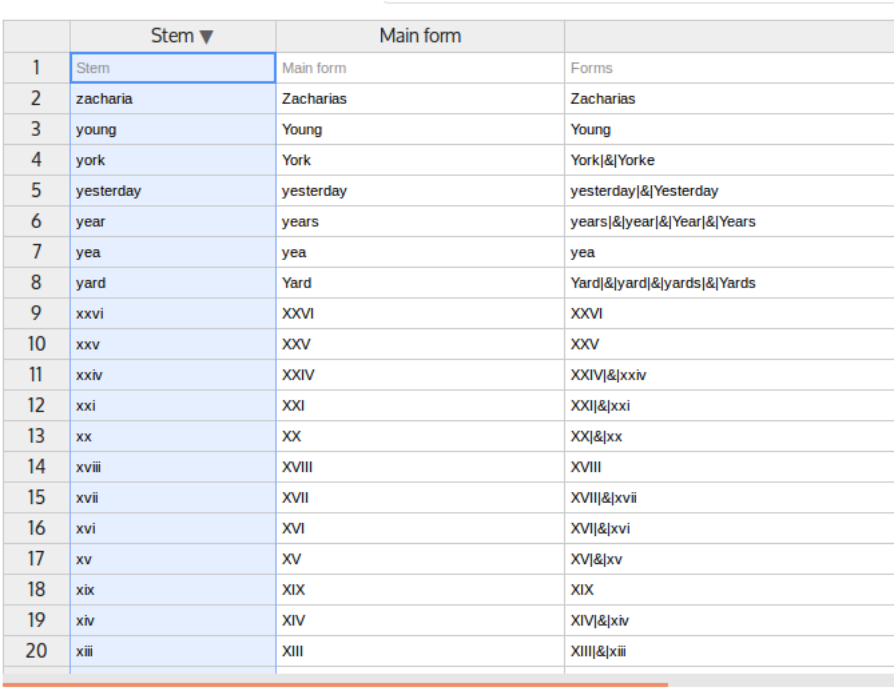
Une liste est maintenant disponible dans notre projet. Plusieurs actions sont possibles :

1. Éditer la liste.

   Il s'agit alors simplement d'observer le résultat. L'utilisateur peut lire les informations, ordonner les colonnes (comme en Figure 32 où les données sont classées par ordre alphabétique inverse des *stem*), ... mais il ne peut rien modifier.
   La liste obtenue répond-elle à nos attentes? Certains termes présents sont-ils surprenants? Faut-il modifier certains des paramètres, relancer le script et la constitution d'une nouvelle liste? Éditer une liste permet de se poser ce type de questions.

2. Modifier la liste.
   Il est possible de supprimer une ligne, modifier un mot (par exemple : modifier une forme principale - *Main form*)

3. Dupliquer la liste.
   Si l'utilisateur ne veut pas faire les modifications directement sur la liste obtenue, il peut faire une copie de la liste, et apporter les modifications souhaitées sur la copie.



	Stem ▼	Main form	
1	Stem	Main form	Forms
2	zacharia	Zacharias	Zacharias
3	young	Young	Young
4	york	York	York|&|Yorke
5	yesterday	yesterday	yesterday|&|Yesterday
6	year	years	years|&|year|&|Year|&|Years
7	yea	yea	yea
8	yard	Yard	Yard|&|yard|&|yards|&|Yards
9	xxvi	XXVI	XXVI
10	xxv	XXV	XXV
11	xxiv	XXIV	XXIV|&|xxiv
12	xxi	XXI	XXI|&|xxi
13	xx	XX	XX|&|xx
14	xviii	XVIII	XVIII
15	xvii	XVII	XVII|&|xvii
16	xvi	XVI	XVI|&|xvi
17	xv	XV	XV|&|xv
18	xix	XIX	XIX
19	xiv	XIV	XIV|&|xiv
20	xiii	XIII	XIII|&|xiii

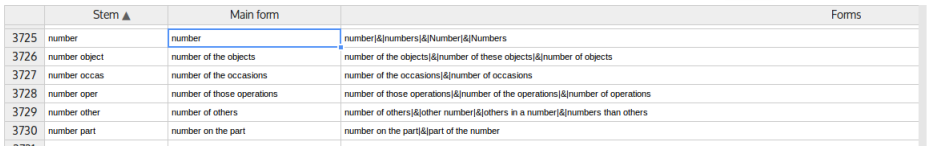
Figure 32: Liste csv ordonnée selon les *stem*

Prenons par exemple la liste *extracted-terms-list-top4999-mono-freq3.csv*. Comment faire pour supprimer certains éléments tels que :

- les chiffres romains (xii, xiv, ...)
- les mots qui ne sont pas des termes (number, ...)

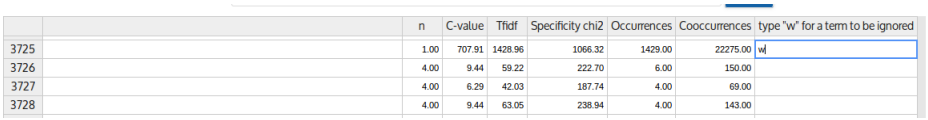
2 solutions existent :

- supprimer les lignes concernées (comme vous le feriez dans votre éditeur *Excel* ou autre tableur : sélectionner la ligne, cliquer sur *suppr*).
- mettre *w* dans la dixième colonne, comme indiqué en Figure 34.



	Stem ▲	Main form	Forms
3725	number	number	number|&|numbers|&|Number|&|Numbers
3726	number object	number of the objects	number of the objects|&|number of these objects|&|number of objects
3727	number occas	number of the occasions	number of the occasions|&|number of occasions
3728	number oper	number of those operations	number of those operations|&|number of the operations|&|number of operations
3729	number other	number of others	number of others|&|other number|&|others in a number|&|numbers than others
3730	number part	number on the part	number on the part|&|part of the number

Figure 33: Fichier csv



		n	C-value	Tfidf	Specificity chi2	Occurrences	Cooccurrences	type "w" for a term to be ignored
3725		1.00	707.91	1428.96	1066.32	1429.00	22275.00	w
3726		4.00	9.44	59.22	222.70	6.00	150.00	
3727		4.00	6.29	42.03	187.74	4.00	69.00	
3728		4.00	9.44	63.05	238.94	4.00	143.00	

Figure 34: Fichier csv (suppression du terme *number*)

	Stem ▲	Main form		n	C-value	Tfidf	Specificity chi2	Occurrences	Cooccurrences	type "w" for a term to be ignored
4977	wrong	wrongs	wrongs|&|Wrongs	1.00	14.89	87.65	219.00	21.00	332.00	
4978	wrongdoer	wrongdoer	wrongdoer|&|wrongdoers	1.00	35.74	180.81	372.33	41.00	888.00	
4979	x2014	x2014	x2014	1.00	5.96	42.52	211.75	4.00	67.00	w
4980	xi	XI	XI|&|xi	1.00	56.59	254.61	1478.60	73.00	361.00	w
4981	xii	XII	XII	1.00	36.74	181.51	919.24	50.00	371.00	w
4982	xiii	XIII	XIII|&|xiii	1.00	20.85	117.47	671.94	27.00	138.00	w
4983	xiv	XIV	XIV|&|xiv	1.00	19.86	113.01	755.12	30.00	195.00	w
4984	xix	XIX	XIX	1.00	8.94	56.54	364.57	26.00	152.00	w
4985	xv	XV	XV|&|xv	1.00	24.82	132.70	323.32	31.00	168.00	w
4986	xvi	XVI	XVI|&|xvi	1.00	30.78	156.67	1941.38	51.00	348.00	w
4987	xvii	XVII	XVII|&|xvii	1.00	9.93	61.77	263.89	14.00	49.00	w
4988	xviii	XVIII	XVIII	1.00	9.93	61.77	278.54	19.00	154.00	w
4989	xx	XX	XX|&|xx	1.00	9.93	61.77	201.82	18.00	125.00	w
4990	xxi	XXI	XXI|&|xxi	1.00	12.91	76.92	294.18	15.00	82.00	w
4991	xxiv	XXIV	XXIV|&|xxiv	1.00	5.96	40.11	380.26	10.00	51.00	w
4992	xxv	XXV	XXV	1.00	4.96	34.33	928.70	5.00	9.00	w
4993	xxvi	XXVI	XXVI	1.00	6.95	46.79	262.01	10.00	53.00	w
4994	yard	Yard	Yard|&|yard|&|yards|&|Yards	1.00	19.86	111.87	365.31	35.00	243.00	
4995	yea	yea	yea	1.00	9.93	61.77	185.26	10.00	107.00	
4996	year	years	years|&|year|&|Year|&|Years	1.00	575.86	1267.92	609.19	725.00	8572.00	

Figure 35: Fichier csv (suppression des chiffres romains)

En utilisant la méthode décrite ci-dessus, il est possible de regrouper des formes de surface (*Forms*) rencontrées dans le texte. Dans l'exemple de la Figure 36, *account* n'est pas la forme principale d'un terme, mais une forme de surface à ajouter aux formes de surface listée par ailleurs et reliées au terme *account*. Ainsi dans la Figure 37 est ajouté *account* à la liste des formes de surface référant au terme *account*, la ligne initiale contenant la forme de surface *account* est commentée (Figure 38).

153	account	account	account|&|accounts|&|Account|&|Accountant|&|Accounts|&|Accountants
154	/ account	account /	account /|&|/ account

Figure 36: Fichier csv (initial)

153	account	account	account|&|accounts|&|Account|&|Accountant|&|Accounts|&|Accountants|&|account /|&|/ account
154	/ account	account /	account /|&|/ account

Figure 37: Fichier csv (ajouter une forme de surface)

153		1.00	624.00	624.00	7909.00	192.47	518.27	1192.95	
154		2.00	4.00	4.00	37.00	406.80	6.29	42.03	w

Figure 38: Fichier csv (commenter une ligne)

⚠ Il est préférable - selon notre expérience - de conserver la liste initiale inchangée et de faire les modifications dans une nouvelle liste, copie de la liste initiale. Attention à ne pas oublier de mettre l'extension dans la nouvelle liste créée : *.csv* (Figure 39).

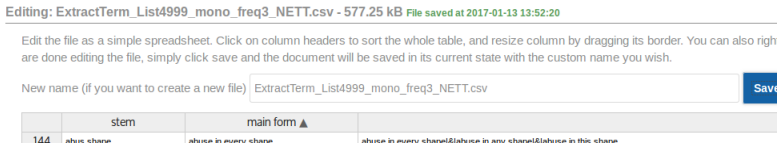


Figure 39: Créer une nouvelle liste .csv

Sur la page d'accueil est maintenant disponible un nouveau fichier *csv* : *ExtractTerm-List4999-mono-freq3-NETT.csv*. Ce fichier est quasiment identique au fichier *ExtractTerm-List4999-mono-freq3.csv*,

à la différence prêt que certains termes ne seront pas indexés, ceux qui ont la mention *w* en colonne 10.

Figure 40: Dossier contenant la liste nettoyée

Il est bien-sûr possible d'utiliser cette nouvelle liste dans les étapes suivantes du traitement. Les lignes contenant w ne seront pas utilisées pour l'indexation. C'est exactement comme si elles étaient supprimées. [13]

⚠ Les modifications ne sont pas rétroactives. Elles ne sont pas prises en compte dans les processus antérieurs. Si une liste a été indexée sur un corpus puis modifiée, l'indexation - antérieure aux modifications - ne sera pas actualisée. Il convient dans ce cas de lancer une nouvelle indexation, de relancer le script *corpus terms indexer*).

⚠ Une liste doit être indexée pour pouvoir être utilisée dans les scripts de visualisation et d'analyse (cf. section 2.2 p.24 Corpus Terms Indexer).

Corpus Terms Indexer

Pour indexer des termes à un corpus, il faut choisir une base de données et un script (cf. Figure 41).

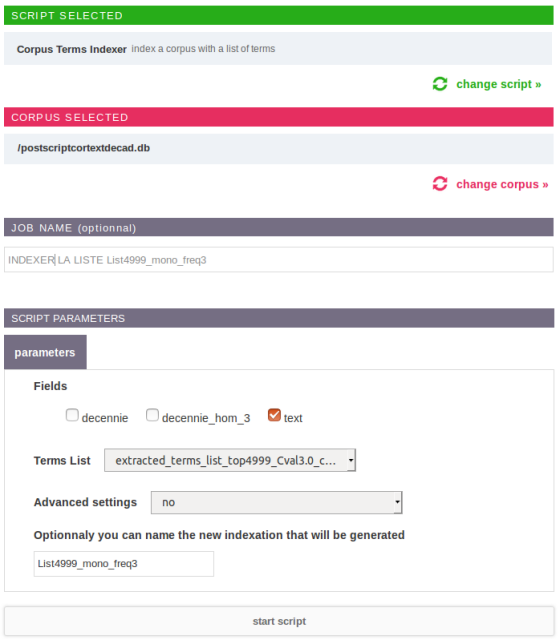


Figure 41: *Corpus Terms Indexer* (paramètres)

13. Dans notre utilisation de l'outil, nous préférons supprimer les lignes plutôt que d'utiliser cette configuration.

Les paramètres du script sont :

- le champ qui contient le texte.
- la liste.
  Dans l'exemple choisi, les termes de la liste *extracted-terms-list-top4999-mono-freq3* seront indéxés sur le texte contenu dans le champ *text*.
- les paramètres d'analyse.
  Ici les paramètres par défaut sont conservés, ce sont ceux de la Figure 42.

- nommer le résultat de la tâche (*choose a name for the indexation*).

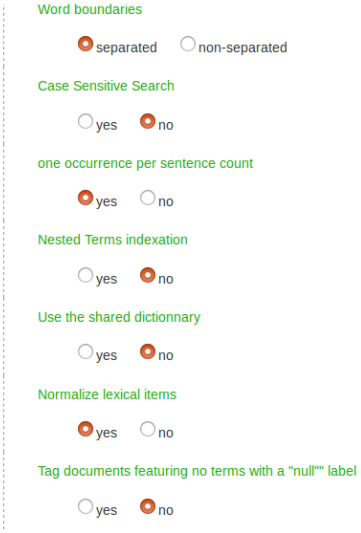


Figure 42: *Corpus Terms Indexer* (advanced setting)

Comment indexer une liste une liste qui n'a pas été constituée dans Cortext ?

1. Mettre dans le format adéquat la liste

   Le format de la liste est celui décrit dans la partie 2.2. La liste doit être constituée de 3 colonnes [14].

2. Télécharger la liste.

   Il faut déposer les données sur la plateforme Cortext. Pour ce faire - comme mentionné en section 1.2 p.7 - sélectionné :

   corpus > add documents > Accept & Upload

14. L'utilisateur peut entrer un tableau constitué de 10 colonnes, les trois premières contenant le lexique à indexer, la dixième pouvant indiquer si le mot est à exclure ou non lors de l'indexation - mettre *w* pour chacun des mots à exclure.

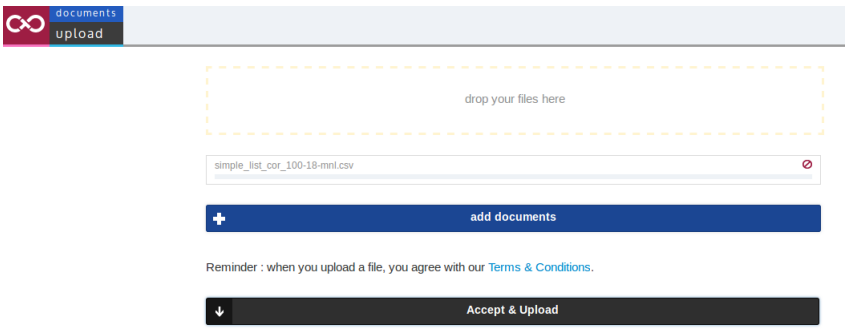


Figure 43: *add documents* (importer une liste)

Sous l'onglet *source* choisir comme type de données *term list*.

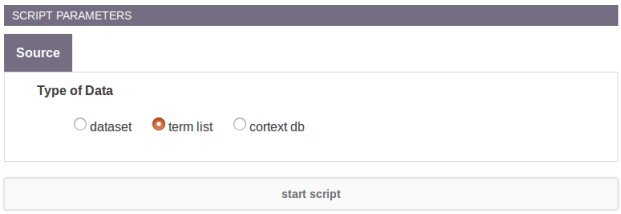


Figure 44: *add documents* (type de données *term list*)

3. Indexer les termes sur le corpus

Choisir le script *term indexer*, sélectionner :

- le corpus
- les paramètres du script

List Builder

[...] L'utilisation de ce script n'est pas abordée dans ce tutoriel.

Notons simplement que selon la documentation en ligne (`https://docs.cortext.net/list-indexer`), ce type de liste peut potentiellement avoir des termes dupliqués, présents plusieurs fois dans la liste.

Corpus List Indexer

[...] L'utilisation de ce script n'est pas abordée dans ce tutoriel.

Named Entity Recognizer

[...] L'utilisation de ce script n'est pas abordée dans ce tutoriel.

2.3 Time

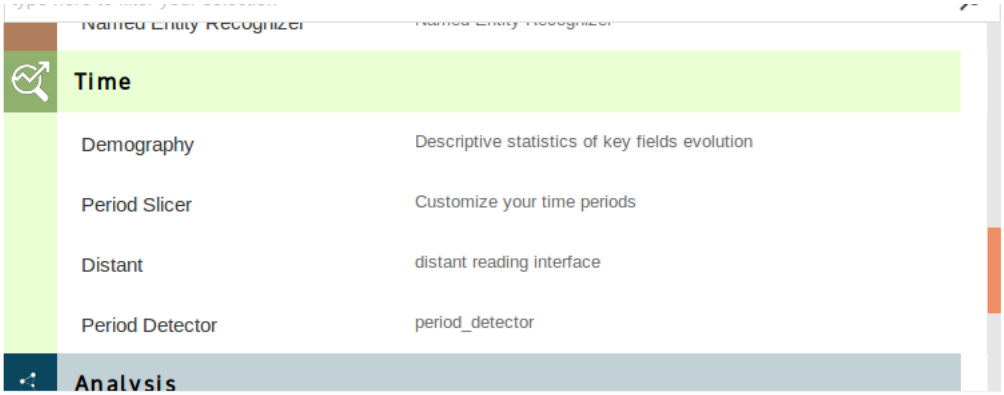


Figure 45: La section *Time*

Demography

[...] L'utilisation de ce script n'est pas abordée dans ce tutoriel.

Period Slicer

Ce script permet de définir des plages temporelles. Pour ce faire la base de données doit contenir un champ avec des données temporelles. Par exemple, la base de données *postscriptCortextDecad* a un champ qui contient des dates. Elles sont comprises entre 1759 et 1832. Ainsi il est possible de déterminer des périodes comme en Figure 47.

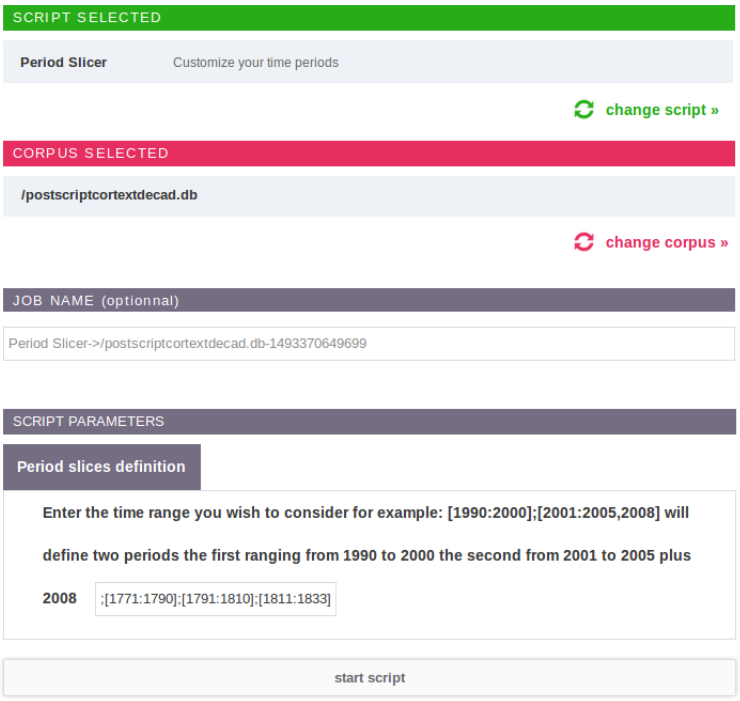
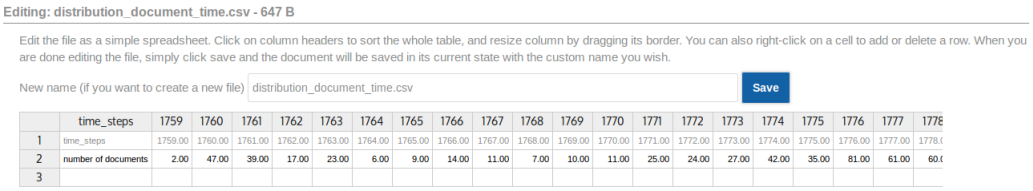


Figure 46: Period slicer (paramètres)

Nous avons défini 4 périodes : la première s'étend de 1758 à 1770 (inclus), la seconde de 1771 à 1790, la troisième de 1791 à 1810 et enfin de 1811 à 1833. Chaque période est placée entre crochets, les périodes sont séparées les unes des autres par un "; " et les dates sont séparées par " : " pour indiquer une plage temporelle s'étendant d'un point à un autre. Le résultat du script est un fichier csv, un tableau.



	time_steps	1759	1760	1761	1762	1763	1764	1765	1766	1767	1768	1769	1770	1771	1772	1773	1774	1775	1776	1777	1778
1	time_steps	1759.00	1760.00	1761.00	1762.00	1763.00	1764.00	1765.00	1766.00	1767.00	1768.00	1769.00	1770.00	1771.00	1772.00	1773.00	1774.00	1775.00	1776.00	1777.00	1778.(
2	number of documents	2.00	47.00	39.00	17.00	23.00	6.00	9.00	14.00	11.00	7.00	10.00	11.00	25.00	24.00	27.00	42.00	35.00	81.00	61.00	60.(
3																					

Figure 47: Period slicer (fichier *csv* résultat)

Nous verrons ultérieurement comment projeter ce découpage temporel sur nos données (cf. section 2.4 p.34)

Distant

[...] L'utilisation de ce script n'est pas abordée dans ce tutoriel.

Period Detector

[...] L'utilisation de ce script n'est pas abordée dans ce tutoriel.

2.4 Analysis

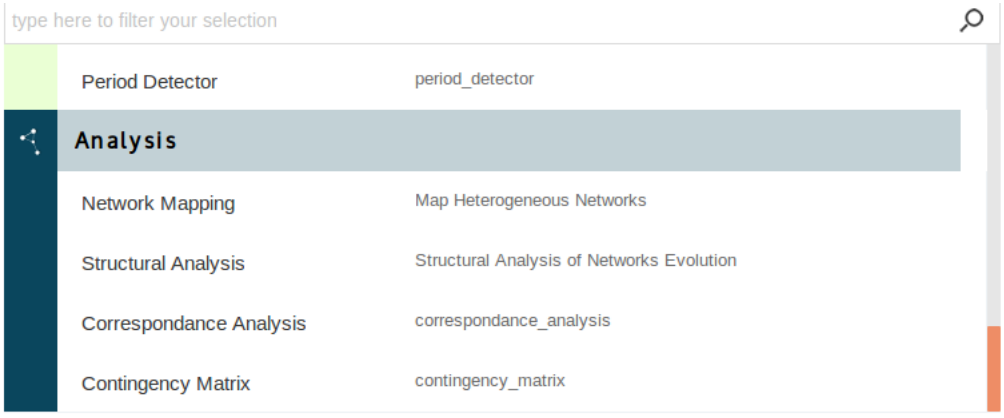


Figure 48: La section *Analysis*

Network mapping

Le formulaire de paramétrage du script est constitué de 4 onglets : les 2 premiers permettent de définir les éléments du graphes (type et nombre de nœuds et d'arcs, les relations entre eux), les 2 suivants

permettent de paramétrer les visualisations, de décider ou non d'avoir une représentation temporelle des données par exemple. Dans un soucis de clarté, nous nous arrêterons dans un premier temps sur les paramètres des 2 premiers onglets [15], puis nous nous intéresserons aux paramètres des 2 suivants.

ONGLETS : SÉLECTION DES NŒUDS ET DES ARCS

- Sélection des nœuds (*Nodes selection*)

  Le premier des onglets permet de définir les nœuds qui constitueront le réseau. Les nœuds sont choisis en fonction de leur fréquence.

  – si ils sont issus de champs différents, le réseau est dit **hétérogène**.
  La documentation en ligne donne des exemples de réseaux hétérogènes [16]. Il peut s'agir par exemple d'établir sur l'ensemble des publications d'un panel de laboratoires, la liste des items les plus fréquents et représentatifs par laboratoire.

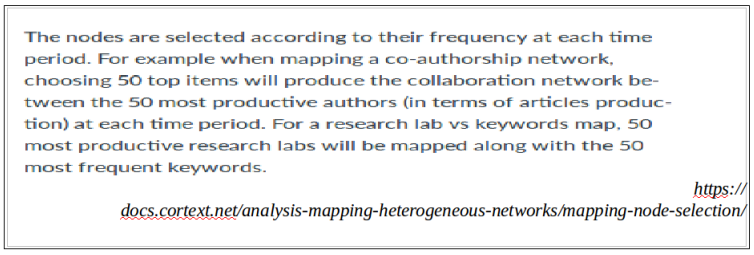


Figure 49: Nœuds hétérogènes (documentation en ligne)

  – si ils sont issus d'un même champs, le réseau est dit **homogène**
  Dans notre exemple (Figure 50), il s'agit d'un réseau homogène. Il s'agit alors d'observer parmi une liste de termes ceux qui s'attirent le plus, ont le plus de critères communs.

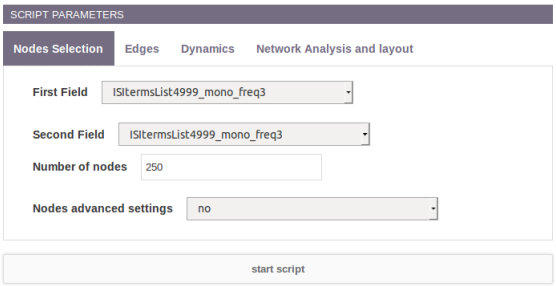


Figure 50: Network mapping (Sélection des nœuds

  Ainsi l'utilisateur a la possibilité de constituer des réseaux homogènes ou hétérogènes selon ce qu'il veut visualiser, selon ses données et les hypothèses établies.

- Les arcs (*Edges*)

15. Les résultats obtenus découleront de modifications des paramètres des 2 premiers onglets, les 2 onglets suivants conservant leur paramétrage par défaut.
16. https://docs.cortext.net/analysis-mapping-heterogeneous-networks/mapping-node-selection/

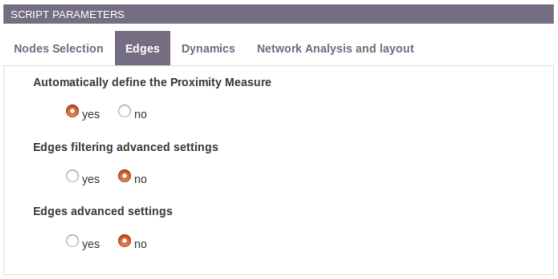


Figure 51: Network mapping (les arcs

Il est possible de paramétrer :

– la mesure de proximité.
Par défaut elle est distributionnelle. En cliquant sur *no* il est possible de modifier ce paramètre (sont disponibles : chi2, cramer, etc Figure 52).

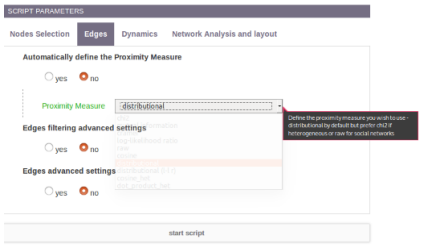


Figure 52: Arcs(mesure de proximité)

– la configuration avancée pour la sélection de arcs.
Les paramètres par défaut (cf. Figure 52) sont :

- optimisation de la valeur de proximité

- valeur statistique de la proximité à 0.1 (valeur conventionnelle en statistique). [17]

- nombre de nœuds considérés : 99999

- nombre de voisins considérés : 9999

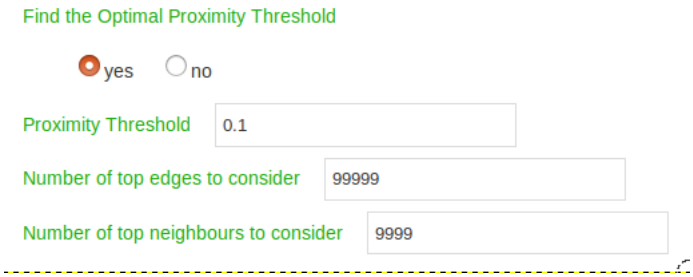


Figure 53: Arcs(sélection des arcs)

– configuration avancée des arcs.
Le paramètre *short range* est très utile pour contraindre les cooccurrences. Il peut s'avérer

17. Si dans le résultat, tous les nœuds sont connectés les uns aux autres, il est conseillé de modifier la valeur de proximité (value of proximity threshold), tester avec 0.2 ou 0.3.

très utile pour les textes longs, il permet de contraindre un contexte étroit. En sélectionnant *advanced settings*>*yes*, il est possible de préciser le contexte sur lequel peut s'étendre l'arc, dans l'exemple donné il est limité à 5, c'est-à-dire que l'analyse se fait sur une fenêtre de 5 phrases.

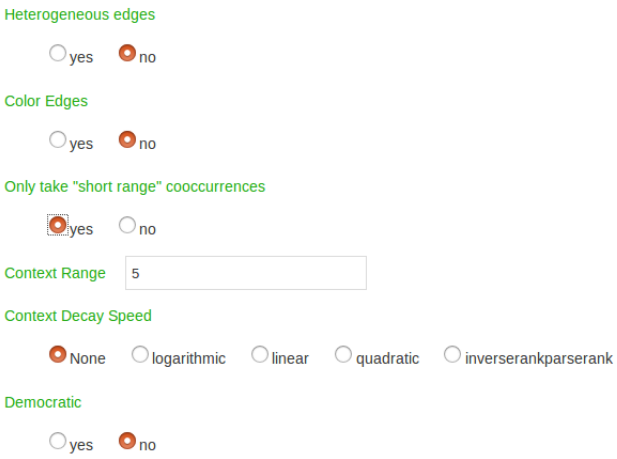


Figure 54: Arcs(configuration avancée par défaut des arcs)

⚠ Pour d'avantage d'information sur l'ensemble de ces mesures, nous vous invitons à consulter la page https://docs.cotext.net/metrics-definitions/.

☝ Pour l'instant nous ne modifions pas les 2 onglets suivants. Nous laissons les paramètres par défaut. Pour valider le choix, cliquez sur *start script*.

RÉSULTATS : SÉLECTION DES NŒUDS ET DES ARCS

Le résultat de ce script est rangé dans 2 dossiers : *mapexplorer* et *maps* (Figure 55).

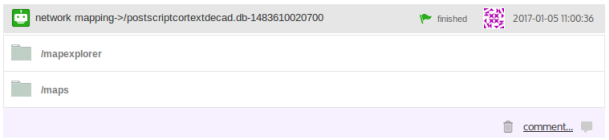


Figure 55: Network mapping(dossiers résultats)

- *maps*.

  Le dossier *maps* contient un document *pdf*. La représentation - le réseau, les nœuds, les liens - est une image fixe. Pour visualiser le document *pdf* il suffit de cliquer sur le nom du fichier (Figure 57).

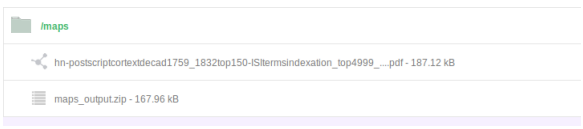


Figure 56: Network mapping(dossier maps)

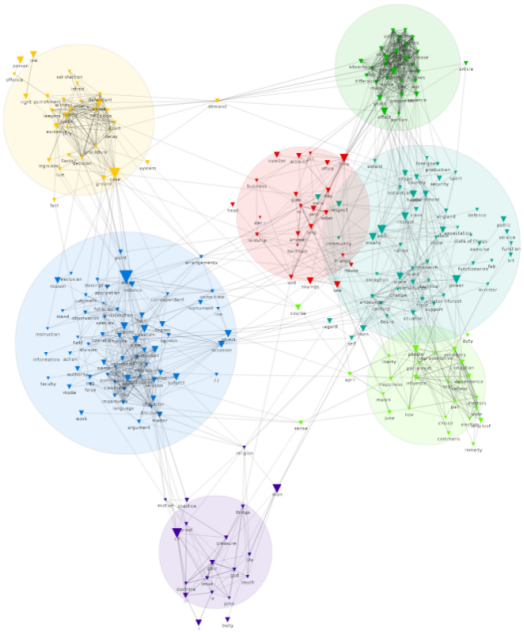


Figure 57: Network mapping(édition du fichier pdf)

- *mapexplorer*.

  Le contenu de la représentation est le même que sous *maps*, mais il est dynamique. Il s'agit d'un document *gexf*. Il est possible de changer quelques paramètres de visualisation, comme :

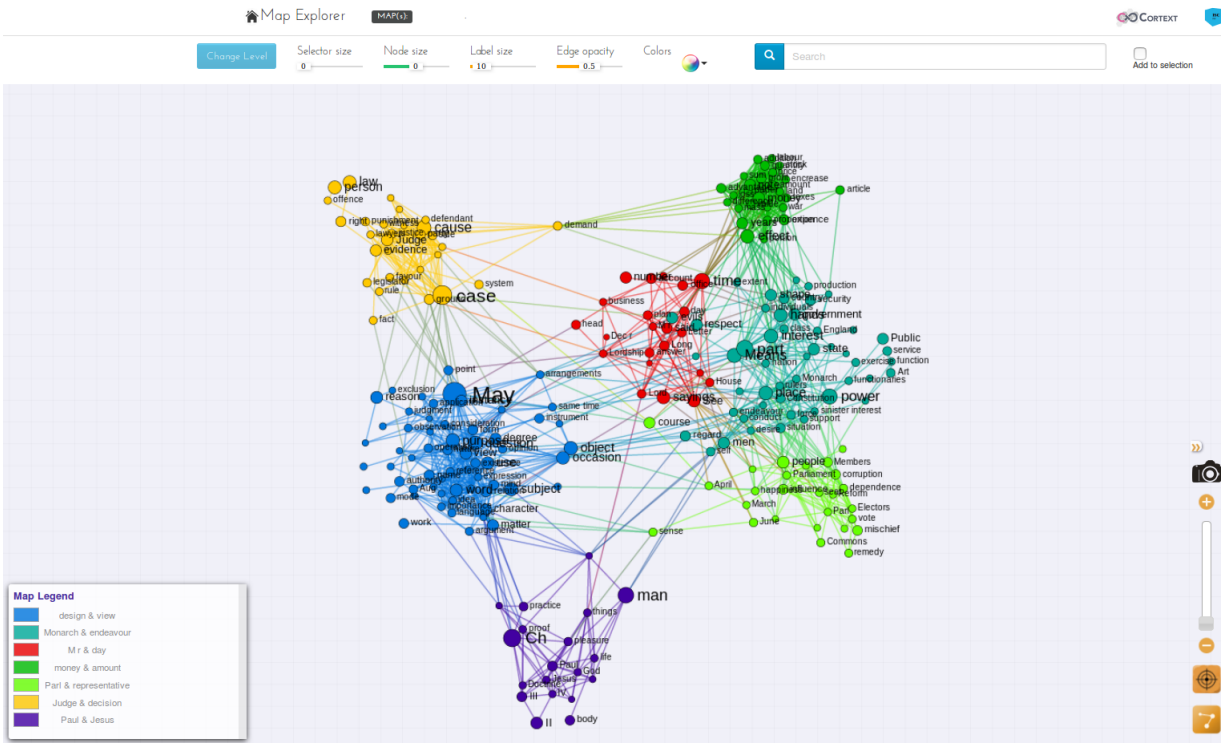


Figure 58: Network mapping(édition du fichier gexf)

- *selector size.*

Cette option permet de sélectionner une partie du réseau. Cela est très utile si vous avez de nombreux nœuds à un même endroit, si les nœuds sont amalgamés. Il est ainsi possible de sélectionner un seul nœud et visualiser l'ensemble des nœuds qui y sont reliés. Figure 59) est sélectionné le nœud *may*, dans la colonne de droite sont listés des nœuds avec lesquels *may* est en relation. Il est possible de sélectionner dans cette colonne un des nœuds de la liste et visualiser ainsi avec quels autres nœuds il entretient des liens, etc etc.

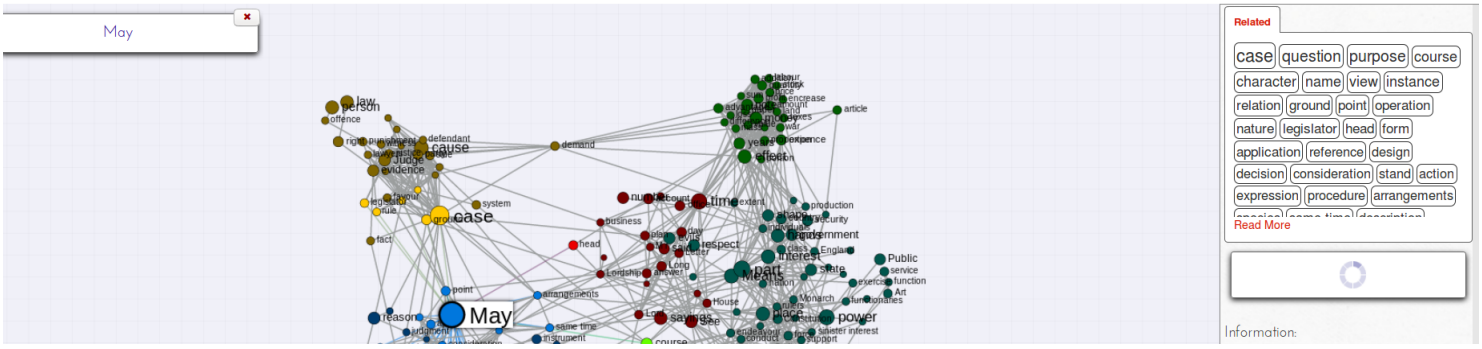


Figure 59: Network mapping (sélection d'un nœud)

Pour annuler la sélection, cliquer sur la croix.

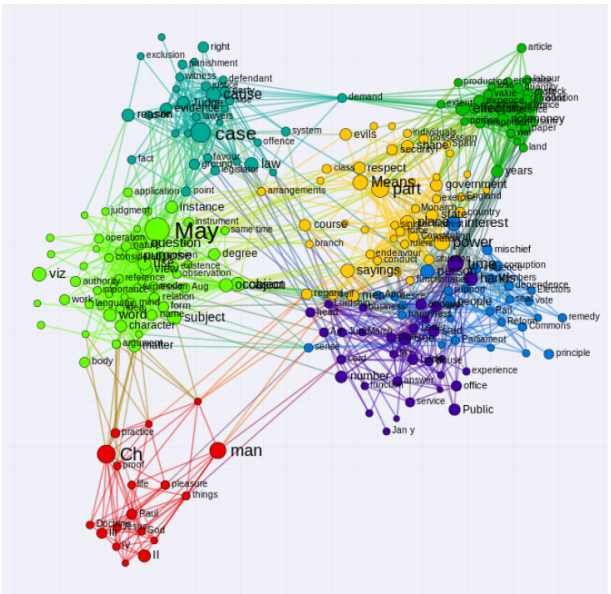
- *copie d'écran.*

En sélectionnant l'appareil photo sur la droite de l'écran il est possible d'enregistrer l'image au format *png* (c'est-à-dire avec la transparence Alpha).

COMPARAISON DE *maps*

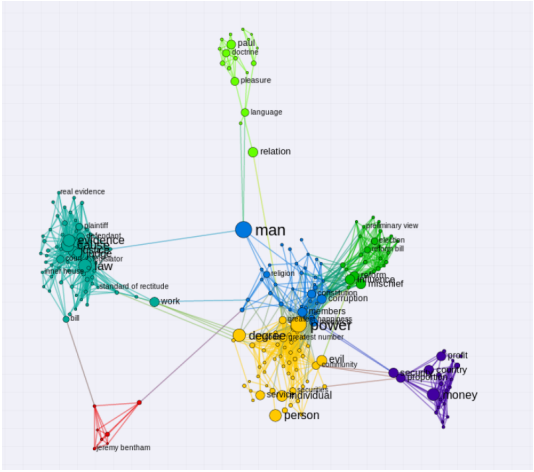
COMPARAISON 1

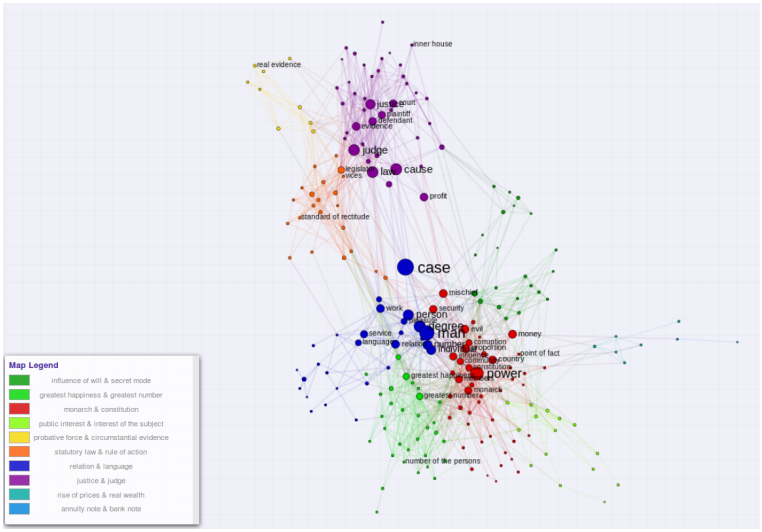
Les 2 réseaux présentés ci-dessous ont les mêmes paramétrages, seules changent les listes qui sont indexées.



COMPARAISON 2

Dans les 2 réseaux ci-dessous, la liste indexée est la même (*termYatea.csv*), mais dans la carte de droite, la cooccurrence est limitée à une fenêtre de 5 phrases, alors qu'elle n'est pas limitée dans la représentation de gauche. Les cartes réseaux sont différentes.





ONGLETS DYNAMICS ET NETWORK ANALYSIS AND LAYOUT

Les 2 onglets *Dynamics* et *Network analysis and layout* permettent de créer des réseaux temporels ou non.

- Dynamique (*Dynamics*)

  Pour obtenir une représentation temporelle, il s'agit de définir :

  – un nombre de périodes
  Il est possible de définir un nombre de périodes :
  soit celui défini ultérieurement, nombre de périodes défini dans le script *period slicer - custom period* (cf. section 2.3 p.28)
  soit selon un champ de type temporel - *standard period*

  – le nombre de périodes
  Si nous reprenons l'exemple utilisé dans l'explication du script *period slicer* (dans 2.3) nous indiquerons ici *4*. Si nous prenons le champ *decennnie* nous indiquerons ici le nombre de périodes que nous souhaitons constituer.

  – le type de répartition souhaitée
  régulière - si l'utilisateur décide que chaque période sera composée d'un même nombre d'année
  homogène - si l'utilisateur décide que chaque période sera composée d'un même nombre de document

  – le recouvrement de ces périodes
  Un recouvrement est-il possible ou non?

  – séquençage

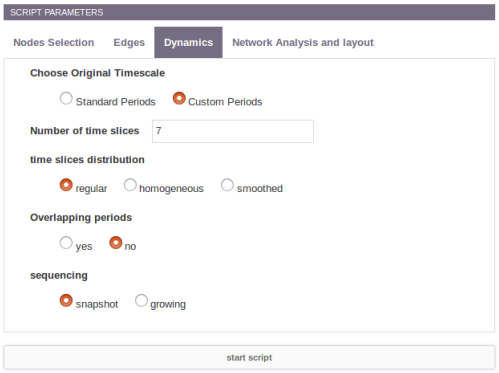


Figure 60: Dynamics (paramètres du découpage)

- Analyse et mise en page du réseau (*Network Anlalysis and layout*)

  Si et seulement si la base de données contient un champ temporel, il est possible de demander une analyse et représentation temporelle. Sélectionner *historical>yes*.

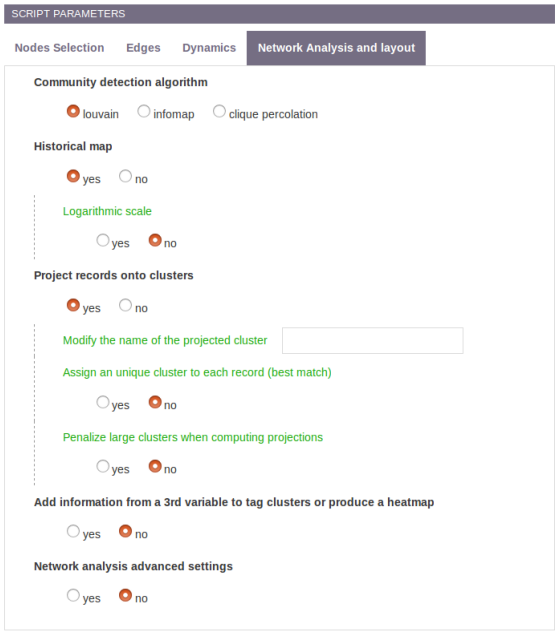


Figure 61: Network Anlalysis and layout (historical)

RÉSULTAT

Sur l'écran d'accueil du projet, l'élément *Network mapping* contient 5 dossiers (Figure 62). Le dossier *tubes* contient des représentions dynamiques temporelles des données (Figure 63).

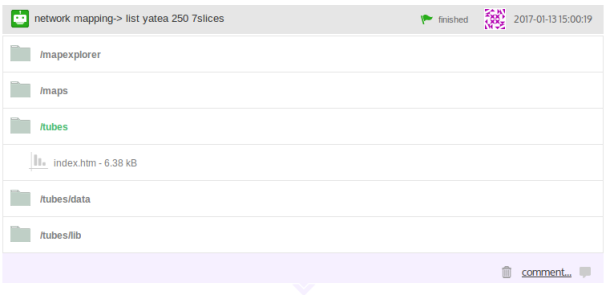


Figure 62: Network mapping (5 dossiers résultat)

Les représentations de type *tube* sont entièrement interactives. La largeur des tubes est proportionnelle aux nombres d'enregistrements. Les tubes plus foncés sont également plus robustes (plus de nœuds sont partagés entre deux périodes consécutives). Notez que les couleurs utilisées pour les *tubes* reprennent les couleurs utilisées pour les cartes réseaux.

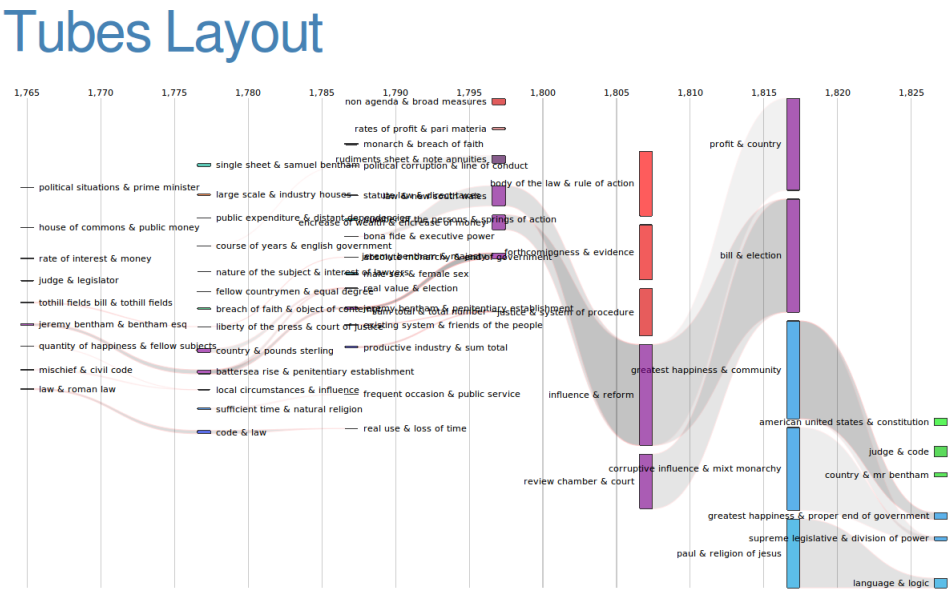


Figure 63: Network mapping (tubes)

La visualisation est dynamique. En passant la souris sur les différents éléments de l'image, des informations complémentaires sont disponibles, comme :

- les termes communs à 2 clusters

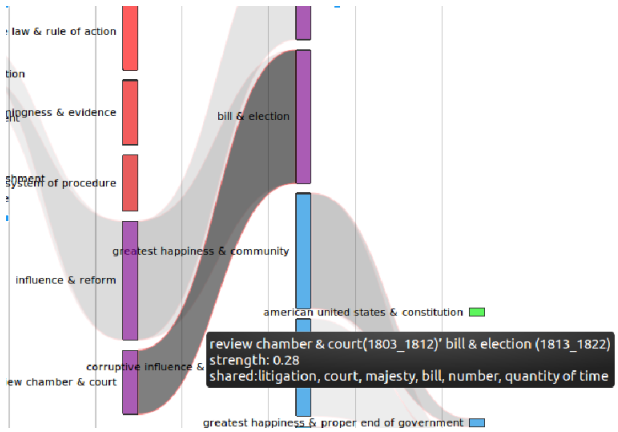


Figure 64: Network mapping (termes communs à 2 clusters)

- l'ensemble des termes d'un cluster:

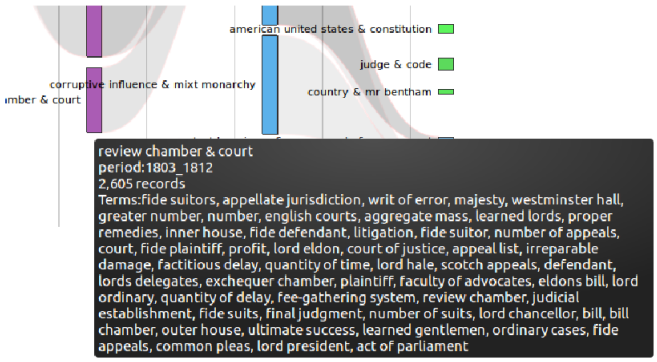


Figure 65: Network mapping (termes d'un cluster)

Que se passe-t-il si sous l'onglet *Dynamics* un seul et unique slice est demandé et que sous l'onglet *Network analysis* est demandé une représentation temporelle (*historical map*) ?

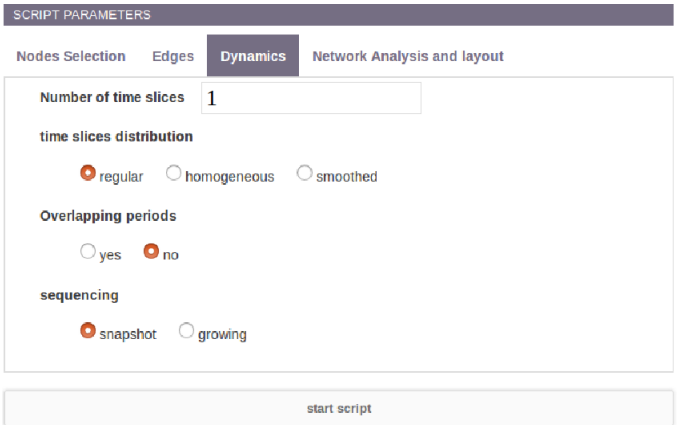


Figure 66: Network mapping (1 slice)

Dans les représentations de réseau, les nœuds sont spatialisés selon les tensions qui existent entre eux, s'ils sont déterminés comme proches par les calculs statistiques. L'espace de représentation se fait sur l'axe des x et y. Mais dans ce type de représentation, le position du nœud est relative à sa date. La

mise en page du réseau est uniquement optimisée selon l'axe des y. Cette option produit des cartes historiques telles que celle de la Figure 67.

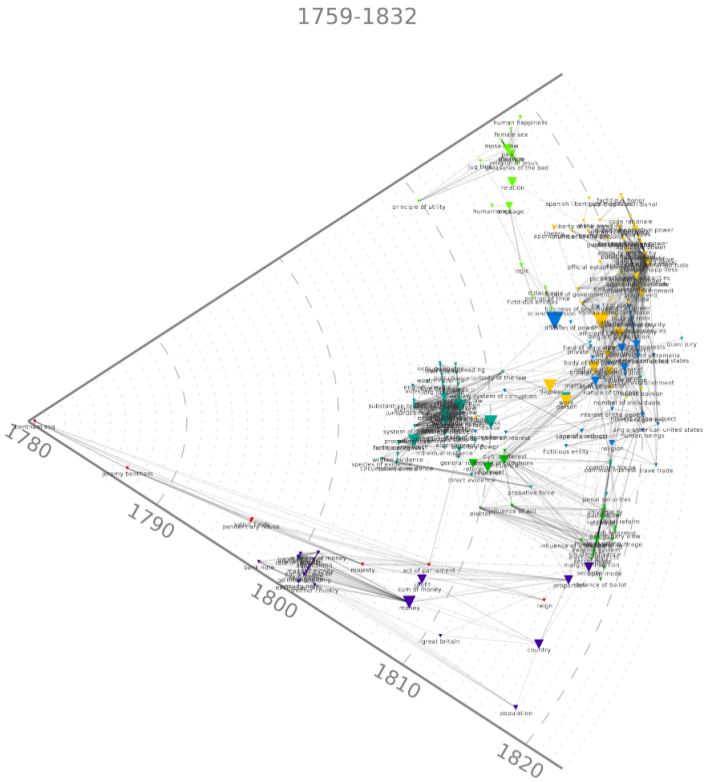


Figure 67: Network mapping (représentation selon l' *axe des y*)

Série temporelle de réseaux

Il est possible de faire une série de représentations au fil du temps.

Dans les représentations de réseaux, les cercles sont utilisés pour représenter les nœuds. Il est possible d'avoir en place et lieu des cercles - dont la taille est proportionnelle au nombre d'occurrences du terme dans un cluster - des zones alpha (*alpha-shapes*). Elles sont dessinées pour obtenir un résultat plus *organique*[18].

Pour optenir ce type de résultat, il convient d'activer les options suivantes dans le script *Networks Analysis and Layout* :

- Sous l'onglet *Dynamics* - quelque soit le type de périodes choisi - laisser à *1* le nombre de découpage (*slice*)
- Sous l'onglet *Network Analysis*
  - choisir un champ de type temporel comme troisième champ. Le champ choisi doit être différent du (ou des) champs choisis sous le premier onglet (*Nodes selection*).
  - inclure l'ensemble des noeuds dans la représentation. Il s'agit d'inclure la valeur *exhaustive* (*node to project exhaustive*)

18. Nous reprenons ici la terminologie utilisée dans la documentation en ligne : https://docs.cortext.net/analysis-mapping-heterogeneous-networks/mapping/

– cocher dans les paramètres avancé les zones alpha (*replace circle with alpha-shapes > yes*)

La Figure 68 donne un aperçu des paramètres entrés dans chacun des 4 onglets.

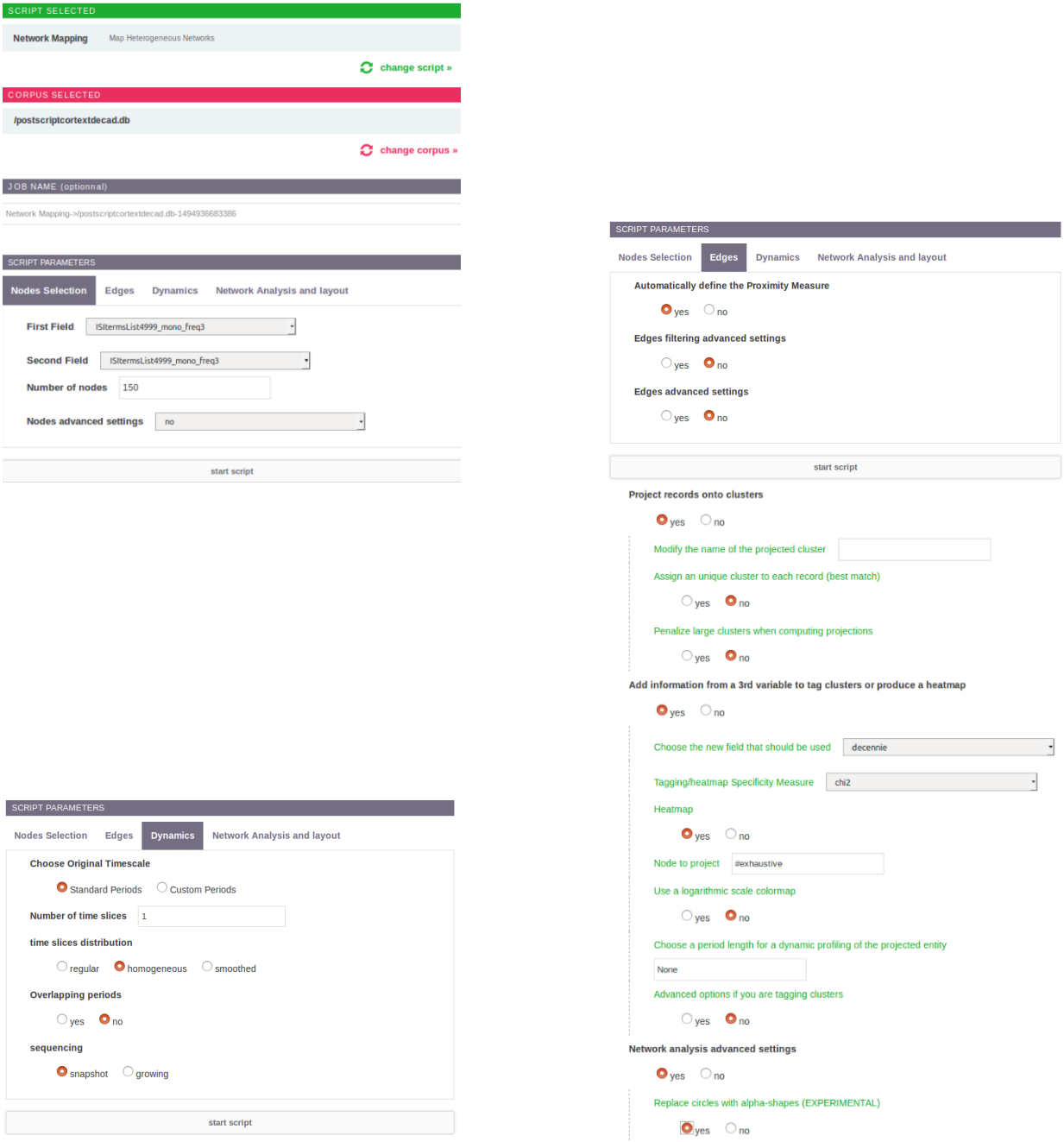


Figure 68: Network mapping(zone alpha et paramètrage des 4 onglets)

Dans notre exemple, le champ temporel utilisé est le champ *décennie*. Ainsi le dossier résultat contenant les réseaux contient 9 fichiers, soit un réseau par décennie (Figure 69).

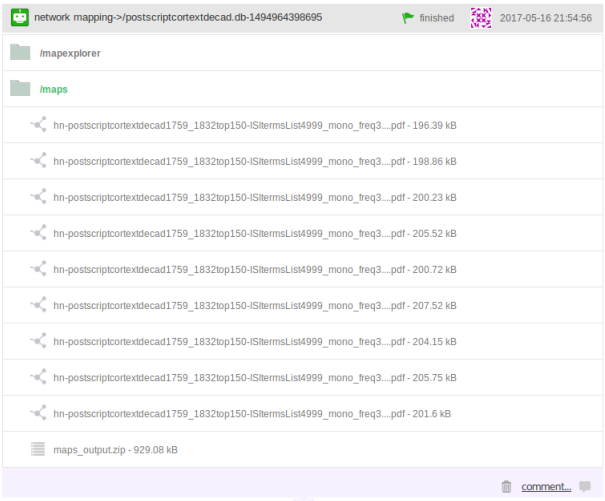


Figure 69: Network mapping(zone alpha et dossier résultat)

Un seul et même réseau est créé. Une zone alpha indique pour chacune des périodes représentée - dans notre exemple ce sont les décennies - les termes significatifs. La Figure 70 est un aperçu de ces 9 fichiers : le réseau est inchangé pour chacune des périodes, les zones alpha illuminent les clusters significatifs pour une période donnée. Plus la zone est rouge, plus elle est représentative de la période.

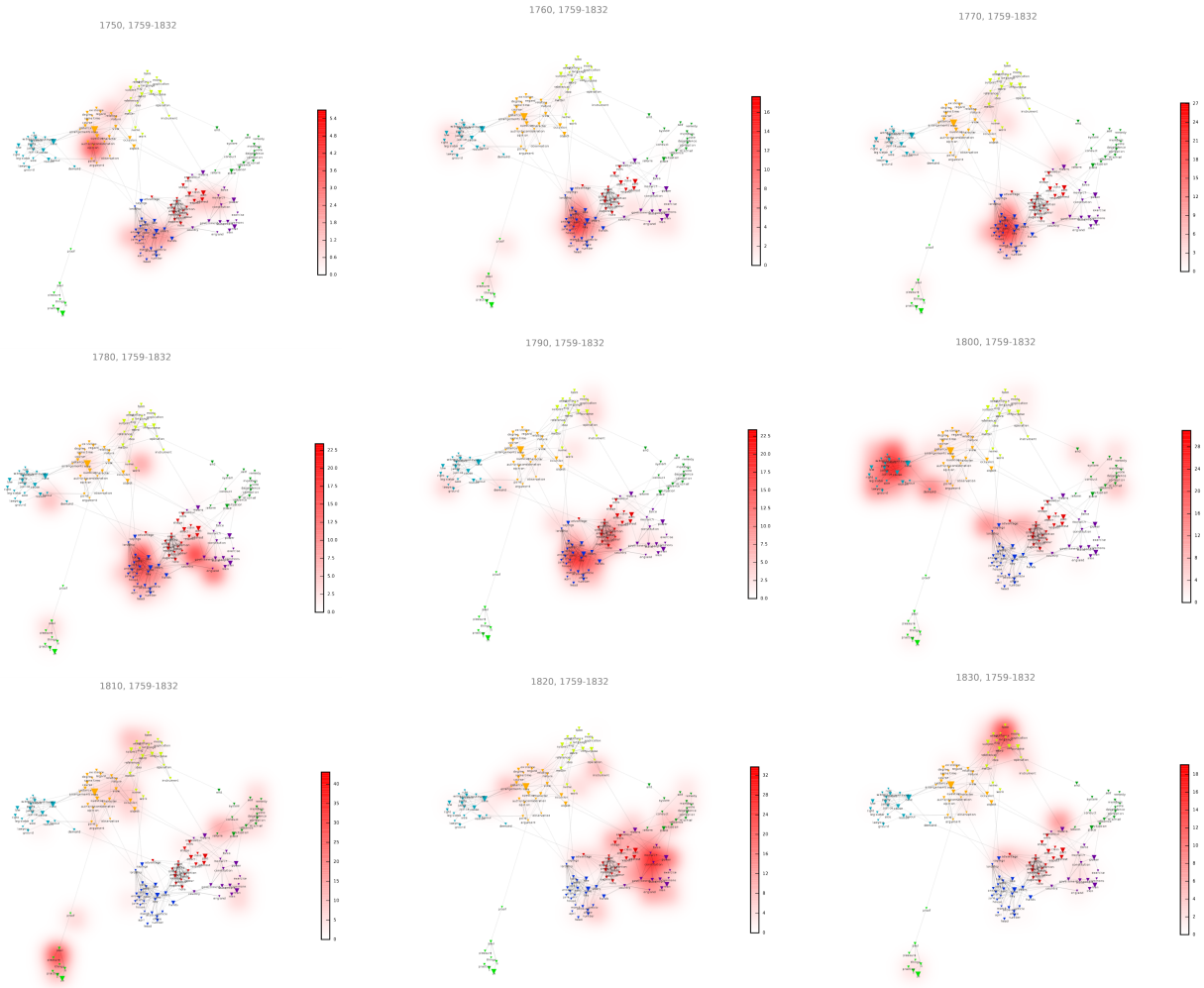


Figure 70: Network mapping (zone alpha et réseau par décennie)

Structural Analysis

[...] L'utilisation de ce script n'est pas abordée dans ce tutoriel.

Correspondance Analysis

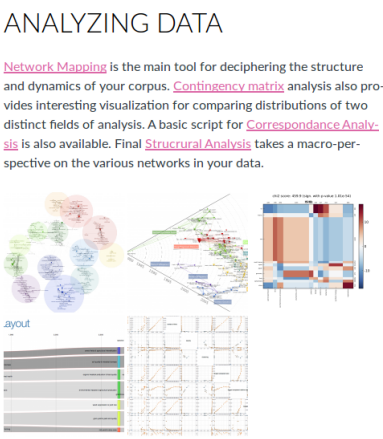
[...] L'utilisation de ce script n'est pas abordée dans ce tutoriel.

Contingency Matrix

[...] L'utilisation de ce script n'est pas abordée dans ce tutoriel.

3 Conclusion

De nombreuses représentations sont possibles sous Cortext, toutes n'ont pas été abordées dans ce tutoriel. Nous espérons cependant vous avoir aidé à appréhender l'outil, et vous invitons à améliorer et poursuivre votre utilisation de l'outil en consultant la documentation en ligne. A vous de jouer : importer vos données, appliquer les scripts selon vos propres hypothèses scientifiques et intuitions.



4 Bibliography

Sites web utiles :

http://apps.lattice.cnrs.fr/bentham/

https://docs.cortext.net/

Référence d'articles qui utilisent Cortext pour analyser les données :

Climate-Risk Disclosure Analysis. (2016). Accessed July 8. http://www.medialab.sciences-po.fr/publications/climateriskdisclosure/index.php

Poibeau, Thierry and Ruiz, Pablo. (2015). Generating Navigable Semantic Maps from Social Sciences Corpora. In Digital Humanities Conference (DH 2015). Sydney, Australia. https://arxiv.org/pdf/1507.02020.pdf

Rule, Alix, Jean-Philippe Cointet, and Peter S. Bearman. (2015). Lexical Shifts, Substantive Changes, and Continuity in State of the Union Discourse, 1790–2014. Proceedings of the National Academy of Sciences 112 (35): 10837–44. doi:10.1073/pnas.1512221112. http://www.pnas.org/content/11/235/10837.full.pdf

Tancoigne, Elise, Barbier, Marc, Cointet, Jean-Philippe, Richard, Guy. The place of agricultural sciences in the literature on ecosystem services. In the literature on ecosystem services. Ecosystem Services, 2014, 10, pp.35-48. https://hal.archives-ouvertes.fr/hal-01157244/document

Venturini, T., N. Baya Laffite, J.-P. Cointet, I. Gray, V. Zabban, and K. De Pryck. (2014). Three Maps and Three Misunderstandings: A Digital Mapping of Climate Diplomacy. Big Data & Society 1 (2). doi:10.1177/2053951714543804. http://journals.sagepub.com/doi/full/10.1177/2053951714543804

Figures