UNIVERSITÉ
CÔTE D'AZUR

TP 13-(14-15): Conception et implémentation d'un algorithme de recherche d'informations

La recherche d'information (RI) étudie la manière de retrouver des informations dans un corpus. Celui-ci est composé de documents d'une ou plusieurs bases de données, qui sont décrits par un contenu ou les métadonnées associées.

Le but de ce TP est de concevoir et implementer un algorithme de recherche d'information. Nous détaillons dans la suite les étapes a suivre pour l'implementation de cet algorithme.

Etape 1 — Construction d'un corpus de documents sur le Covid-19

Vous devez sauvegarder le contenu textuel present dans les pages web suivantes (un document txt pour chaque page web) :

1. https://www.nature.com/articles/d41586-020-00502-w
2. https://www.nejm.org/doi/full/10.1056/NEJMoa2033700?query-featured_coronavirus=
3. https://www.nejm.org/doi/full/10.1056/NEJMoa2030340?query=featured_coronavirus
4. https://www.nejm.org/doi/full/10.1056/NEJMoa2035002?query-featured_coronavirus=
5. https://www.nejm.org/doi/full/10.1056/NEJMoa2029849?query=featured_coronavirus
6. https://www.nejm.org/doi/full/10.1056/NEJMpv2035416?query=featured_coronavirus
7. https://www.thelancet.com/journals/lanrhe/article/PIIS2665-9913(21)00007-2/fulltext
8. https://www.thelancet.com/journals/lanres/article/PIIS2213-2600(21)00025-4/fulltext
9. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)32656-8/fulltext
10. https://science.sciencemag.org/content/early/2021/01/11/science.abe6522

Les légendes des images ne doivent pas être sauvegardées.

Le résultat de cette premiere étape est un corpus de 10 documents en anglais sur le sujet de la Covid-19.

Etape 2 — Pretraitement des données et construction de la matrice d'incidence

L'étape 2 est composée de 5 sous-taches séparées. Vous devez :

1. segmenter chaque document a l'aide d'un tokenizer et extraire les mots (c'est-a-dire couper la séquence de caractères en tokens/ segments et les extraire).

2. éliminer les mots-vides (c'est-a-dire enlever les mots très courants, non porteurs de sens).

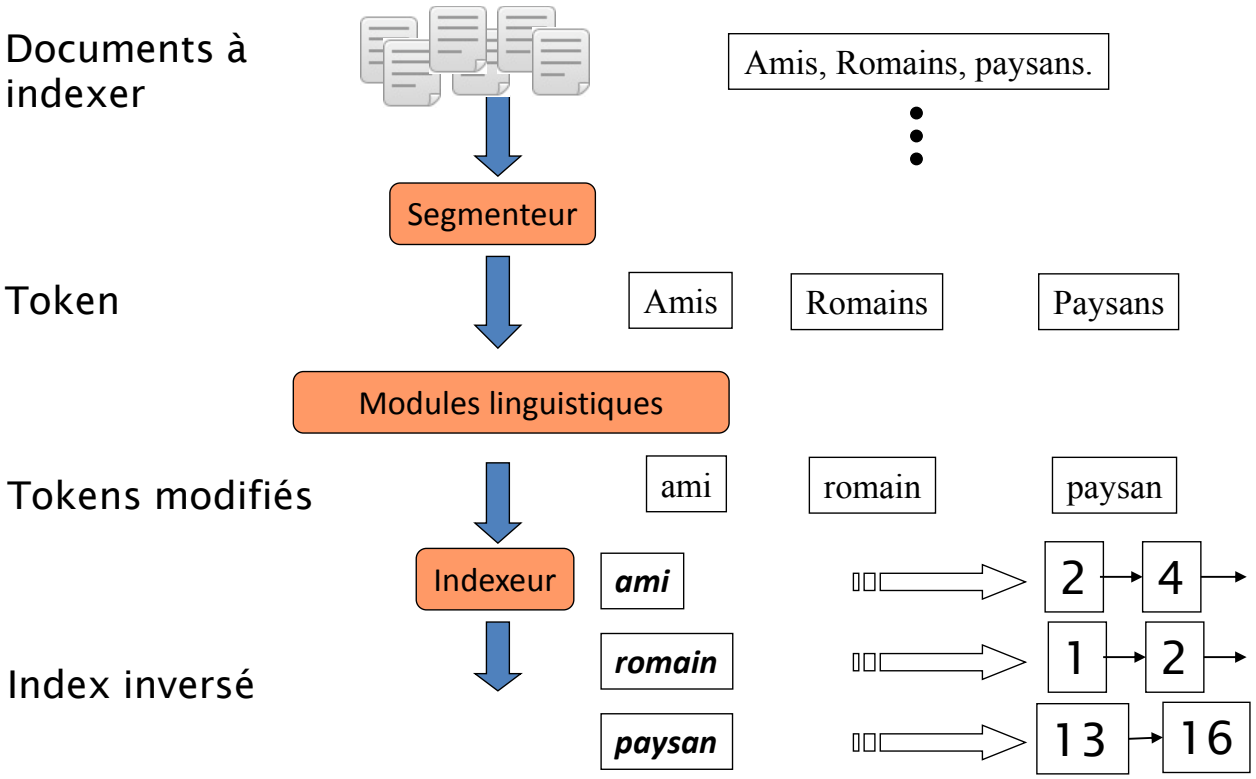
3. normaliser le vocabulaire des termes a travers la racinisation (c'est-a-dire faire en sorte que les mots d'une même racine soient retrouvés) et la lemmatisation (c'est-a-dire mapper les mots sur les forme du dictionnaire).

4. définir le dictionnaire relatif à cette collection de documents.

5. En fin, vous construisez la matrice d'incidence documents-termes.

Etape 3 — Construction de l'index inversé

Apres la construction de la matrice d'incidence documents-termes, vous devez construire l'index inversé (liste variable). Vous retrouvez une visualisation de toutes les étapes jusqu'ici dans l'image suivante :



Etape 4 — Implementation d'un algorithme de recherche d'information

La quatrième étape consiste dans l'implementation de l'algorithme de recherche d'information et elle passe a travers 2 sous-étapes a difficulté croissante, notamment la recherche d'information avec de requêtes booléennes et la recherche d'informations avec des requêtes textuelles plus complexes.

Petit rappel : si on considère une requête qui ne contient qu'un seul mot quel est l'ordre le plus pertinent pour retourner les documents correspondants?

Solution la plus simple : compter le nombre d'occurrences du mot dans chaque document et retourner d'abord les documents ayant le plus grand nombre d'occurrences de ce mot.
Solution un peu plus complexe : compter les occurrences dans les documents et les normaliser en accord avec la longueur du document (# occurrences/# terms)

Sous-étape 4.1 — Requêtes booléennes

Utilisez l'indexation que vous avez faite pour implementer un algorithme qui vous permet d'executer et trouver une réponse pour les diverses requêtes booléennes ci-dessous.

Requêtes booléennes :
1. desease AND severe AND pneumonia
2. antibody AND plasma AND (cells OR receptors)
3. antimalarial drugs OR antiviral agents OR immunomodulators
4. NOT plasma AND risk of infection AND restrictions
5. (older adults AND antibodies) AND NOT (genomes OR variant)

Petit rappel : meme la requête doit passer a travers un processus de pretraitement similaire a celui fait sur la collection des données.

Sous-étape 4.2 — Requêtes textuelles plus complexes

Utilisez l'indexation que vous avez faite pour faire en sorte que votre algorithme puisse executer et trouver une réponse pour les diverses requêtes textuelles complexes ci-dessous.

Requêtes complexes :
1. antibody treatments
2. efficacy and safety of the treatments
3. family access to hospitals
4. contact tracing results
5. genomic analysis of SARS-CoV-2 disease

Etant donné les réponses obtenues pour chaque requête (booléenne ou complexe), vous devez en fin classer les documents récupérés par ordre de pertinence par rapport à la requête en utilisant TF*IDF.

Pour le rendu du TP (possibilité de le faire en binôme):
- les documents de la collection, le code pour executer votre algorithme de RI, la matrice d'incidence, l'index inversé et les résultats aux requêtes (sans et avec ordre de pertinence).
- un cout rapport (2 pages maximum) décrivant la conception de l'algorithme et les principales étapes de l'implementation.