Université Paris-Sud Master AIC

Recherche d’ information

TP3

Étudiants: Yang Mo

Superviseur: Anne-Laure Ligozat

Date de rapport : 21/10/2017

Exercice 1: étude des lexiques

Comptages lexicaux

Générez la liste des 1000 mots les plus fréquents du corpus.

- **Quels sont les mots les plus fréquents ? Pourquoi ?**

Après la recherche, on a trouvé que les mots les plus fréquent sont les mots simples et basics (prépositions, déterminants...)

- **Constituent-ils de bons descripteurs des textes ?**

Absolument non. Les mots simples ne sont pas les bons descripteurs des textes. Pour les plus part de ces mots il n'y a pas de sens particuliére.

- **Comment pourrait-on choisir de meilleurs descripteurs ?**

C'est milleur à chosir les mots essentials (les noms, les verbes, les adjectifs……)

Générez la liste des 1000 lemmes ou racines les plus fréquents.

- **Quelles différences constatez-vous ? Donnez des exemples.**

Aprés lister des 1000 racines les plus fréquents, on ne observe plus les différents formes des mots, donc la frequence de chaque mot augmente un peu. C'est à dire la frequence des mots sont un peu plus concentré. Par exemple, le mot 'disease' apparaît 178 fois dans la liste de racines (comparé avec 167 fois dans la liste de mots)

- **Quel problème a-t-il été traité avec les lemmes ?**

C'est une maniére de normaliser des mots à uniform qui partage la même racine.ça permet de chercher le mot à n'import quel forme.

Générez les 15 catégories les plus fréquentes.

- **Quelle est la catégorie la plus fréquente ? Est-ce une bonne chose ? Pourquoi ?**

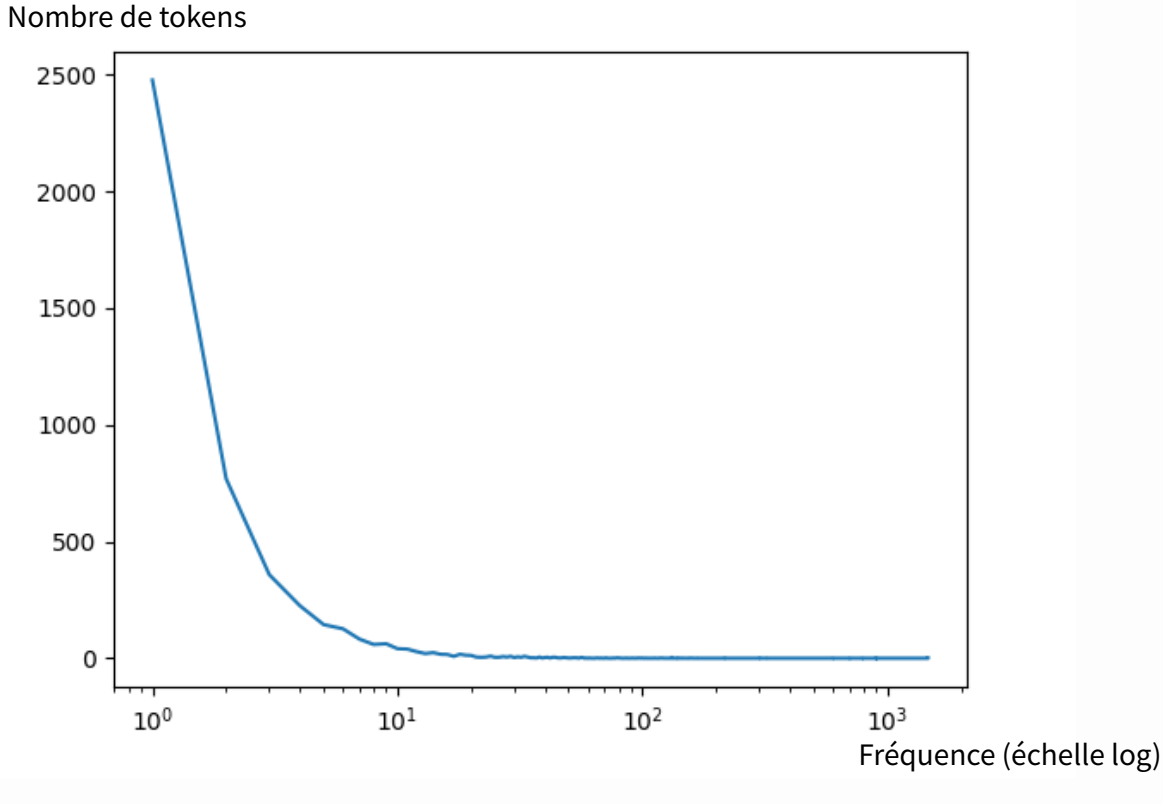
La catégorie la pus fréquente est ‘NN’ (Noun Normal). Je pense que la notion de ‘Noun’ est trop large, y inclus beaucoup de mots qui ne sont pas du sens particuliéres. Mais quand même il y a beaucoup de ‘Noun’ qui sont les mots essentials et on doit trouver une certaine maniéres de les rétirer.

- **Quelle(s) catégorie(s) garderiez-vous pour obtenir de bons descripteurs ? Pourquoi ?**

Les mieulleurs descripteurs selon moi sont parmi ‘NNP’ (proper noun), ‘PRP’ (personnal pronom), ‘NN’ (Noun). Parce que souvents ils correspondent aux ‘concept’ lorsque on veut trouver tous les documents ou enregistrements contenant ce concept. Le decripteur peut être un term, une personne, un object, il peut aussi être plus large que ça (un catégorie, eg. Animale, fruit…...)

Représentation graphique

- **Créez une représentation graphique avec plot prenant en abscisse une fréquence et en ordonnée le nombre de tokens et de lemmes ayant cette fréquence (échelle log).**



Exercice 2: calcul de similarité et pertinence de la recherche

Le but est de sélectionner des textes proches d'une requête. Afin de calculer leur similarité, ils seront représentés chacun par l'ensemble des termes que vous aurez retenus comme descripteurs auxquels on associera un poids, au moins côté documents.

Pondération des termes

- **Comparer la fréquence des termes dans un texte et leur tfidf. Que constatez-vous en terme de poids par rapport à la fréquence ? Donnez des exemples. À quoi est-ce dû ?**

On a choisi le texte avec index ('rd_16_psg_38_40.txt'), dans ce texte on a selectionné deux mots aléatoirement 'way' et 'healthcare'. Ils sont apparaîtent tous 1 fois dans ce texte (la fréquence est donc 0.0139). Le tfidf de 'healthcare' est 0.131 alors que le tfidf de 'way' est 0.117. Parce que le mot 'healthcare' apparaît moins dans autre textes que 'way', c'est à dire il a l'air un mot plus essential pour ce texte. Donc son poid dans la matrix thidf est plus que 'way'.

Calcul de similarités

- **Calculez le MRR pour le script courant.**

Le MRR pour le script courant est 0.520

- **Les textes les plus pertinents le sont-ils réellement selon la référence ? Si ce n'est pas le cas, quels sont les descripteurs responsables ?**

Il faut aussi considérer à mon avis esc-ce que les mots essentials (les concepts, les personnages, les terms) sont compris dans le texte? La similarité n'est pas la seule régle et il ne peut pas fonctionner tout seule.

Amélioration de la similarité

- **Au vu des causes d'erreurs, améliorez la recherche en faisant varier les mesures de similarité, et en trouvant des critères et des méthodes pour enlever automatiquement les mauvais descripteurs si il y en a dans les requêtes.**

Mon amelioration proposé : pour chaque texte, je calcule tfidf de touts les mots, et enelver tous les mots qui ne ont pas valeur tfidf ou ont valeur tfidf moins de 0.1. Par example, le texte initial de doc('rd_16_psg_38_40.txt') :

> <sent id="38"> This effort to focus on the preserved creative functions , instead of deficits of the patient , will improve their quality of life and is a rewarding way for caregivers to communicate with them . </sent> <sent id="39"> The findings of this scientific case study are published in the June issue of European Journal of Neurology . 800 Seniors.com is a leading referral system in the Elderly Healthcare industry . </sent> <sent id="40"> We are located on 5400 Atlantis Court , Moorpark , California 93021 . </sent>

aprés traitement, il deviens :

> effort focus preserved creative instead patient improve quality life rewarding way communicate scientific case study published june issue european journal neurology 800 seniorscom leading referral elderly healthcare industry located 5400 atlantis court moorpark california 93021

donc il reste que les mots de bons descripteurs. À partir de ça, on recalcule la simularité entre la question et les texte, enrigistrer les résultats dans le ficher csv, recalcule le MRR.

Malgré le taux de correction augmente (il y a beaucoup plus de cas qui retour rank 1, c'est à dire le request retour la reponse correct dans un prémier temps), mais le MRR n'a pas augmenté(0.486), parce que dans certains cas la request rendre rank 80-90 (c'est à dire il n'y a pas du tout simularité). La raison peut-être est le filtre a filtré les bons descripteurs également avec les mauvais descripteur dans ces cas.