Spécificité des termes

26 novembre 2019

—Lexicologie, terminologie, dictionnairique—

Exercices.

1. Téléchargez et décompressez le corpus Reuters dans Lexicology2019/Data/Reuters/ .

2. Créez une liste de fréquence de termes (mots simples), et sélectionnez les 5.000 termes les plus fréquents (supprimez le reste).

3. Pour les 5000 mots, construisez une matrice terme-document M 7770 x 5000 où les lignes correspondent aux documents, les colonnes aux termes, et chaque case $M_{i,j}$ contient le nombre d'occurrences du mot j dans le document i.

4. Pour chaque terme t et chaque document d, calculez maintenant le poids **tf-idf** (term frequency - inverse document frequency). Le formule se compose de deux facteurs :

   — $tf_{t,d}$ (term frequency of t in d) est la fréquence du terme t dans document d,

   — et $idf_{t,D}$ (inverse document frequency of t in corpus D) est l'inverse de la proporition de documents qui contiennent t. Plus le mot est rare, plus son idf sera donc élevé. Il est spécifique à chaque terme, mais constant par rapport aux documents.

$$idf_{i,D} = log \frac{|D|}{|\{d \in D : t \in d\}|} \tag{1}$$

   où D est le corpus (collection de plusieurs documents), $|D|$ est le nombre de documents dans le corpus, $|\{d \in D : t \in d\}|$ est le nombre de documents d dans D qui contiennent le terme t.

   Finalement, le poids tf-idf se calcule comme suit :

$$tf - idf_{t,d} = tf_{t,d} \cdot idf_{t,D} \tag{2}$$