FRI-WEB

Michelle EVANS BOUCHET
Solen LE ROUX–COULOIGNER
Julien RASPAUD

Monday 17th December, 2018

1 Création d'un index inversé et moteur de recherche booléen et vectoriel

1.1 Traitements linguistiques pour la collection CACM

1.1.1 Combien y a-t-il de tokens dans la collection ?

Il y a $\boxed{192129 \text{ tokens}}$ dans le corpus.

1.1.2 Quelle est la taille du vocabulaire ?

Il y a 9851 mots dans le vocabulaire, ce qui après avoir retiré les mots de la stop-liste, revient à un total de $\boxed{9496 \text{ mots}}$.

1.1.3 Calculer le nombre total de tokens et la taille du vocabulaire pour la moitié de la collection et utiliser les résultats avec les deux précédents pour déterminer les paramètres k et b de la loi de Heap.

La loi de Heap est définie par :

$$V_R(n) = Kn^\beta$$

avec V_R le nombre de mots différents dans la collection, n la taille du texte, et K et β des paramètres arbitraires. On a donc, pour la collection entière,

$$9851 = K \times 192129^\beta$$

Pour obtenir une deuxième formule, nous procédons aux mêmes calculs sur la moitié seulement du corpus i.e. en s'arrêtant à 1602 textes au lieu de 3204. On a alors $\boxed{5613 \text{ mots}}$ différents et $\boxed{55263 \text{ tokens}}$. On a donc les deux formules suivantes :

$$\begin{cases} 5613 = K \times 55263^\beta \\ 9851 = K \times 192129^\beta \end{cases}$$

A partir de ces deux équations, on peut déduire β et K:

$$\beta = \frac{log(9851) - log(5613)}{log(192129) - log(55263)} = \boxed{0.451}$$

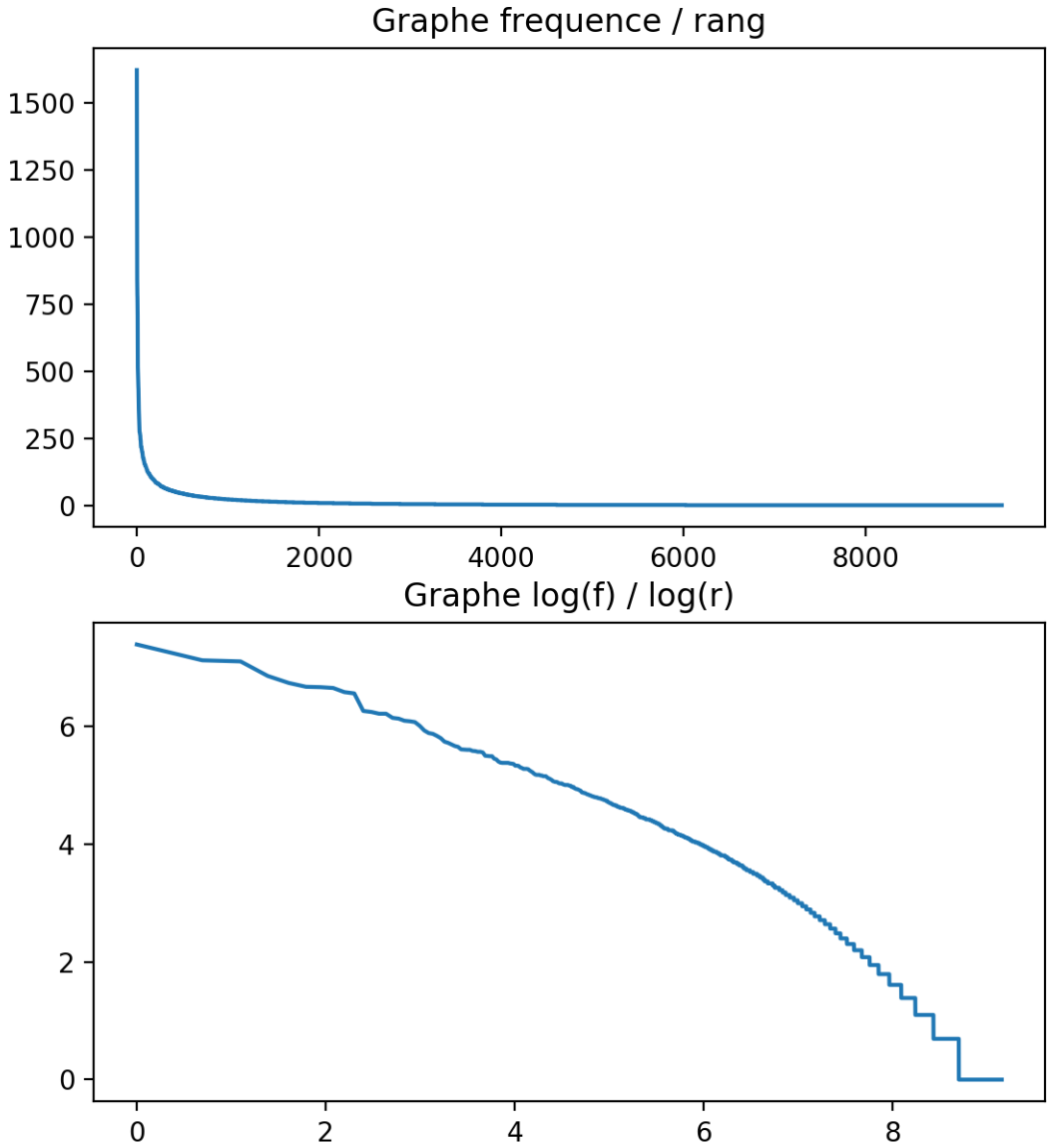
$$K = \frac{9851}{192129^\beta} = \boxed{40.6}$$

1.1.4 Estimer la taille du vocabulaire pour une collection de 1 million de tokens.

D'après la loi de Heap, on a donc pour la collection CACM:

$$V_R(10^6) = 40.6 \times 10^{6 \times 0.451} = \boxed{20743 \text{ mots}}$$

1.1.5 Tracer le graphe fréquence (f) vs rang (r) pour tous les tokens de la collection. Tracer aussi le graphe log(f) vs log(r).



1.2 Traitements linguistiques pour la collection CS276

1.2.1 Combien y a-t-il de tokens dans la collection ?

Il y a $\boxed{26.517.667 \text{ tokens}}$ dans le corpus.

1.2.2 Quelle est la taille du vocabulaire ?

Après avoir retiré les mots de la stop-liste, on obtient un total de $\boxed{333.828 \text{ mots}}$.

1.2.3 Calculer le nombre total de tokens et la taille du vocabulaire pour la moitié de la collection et utiliser les résultats avec les deux précédents pour déterminer les paramètres k et b de la loi de Heap.

Pour obtenir les deux formules, nous utilisons les valeurs obtenues précédemment et cherchons à obtenir les même sur une moitié seulement du corpus. On obtient par ce procédé les résultats suivants :

$$\begin{cases} 333.828 = K \times 26.517.667^{\beta} \\ 196.159 = K \times 12.259.014^{\beta} \end{cases}$$

A partir de ces deux équations, on peut déduire β et K:

$$\beta = \frac{log(333.828) - log(196.159)}{log(26.517.667) - log(12.259.014)} = \boxed{0.689}$$

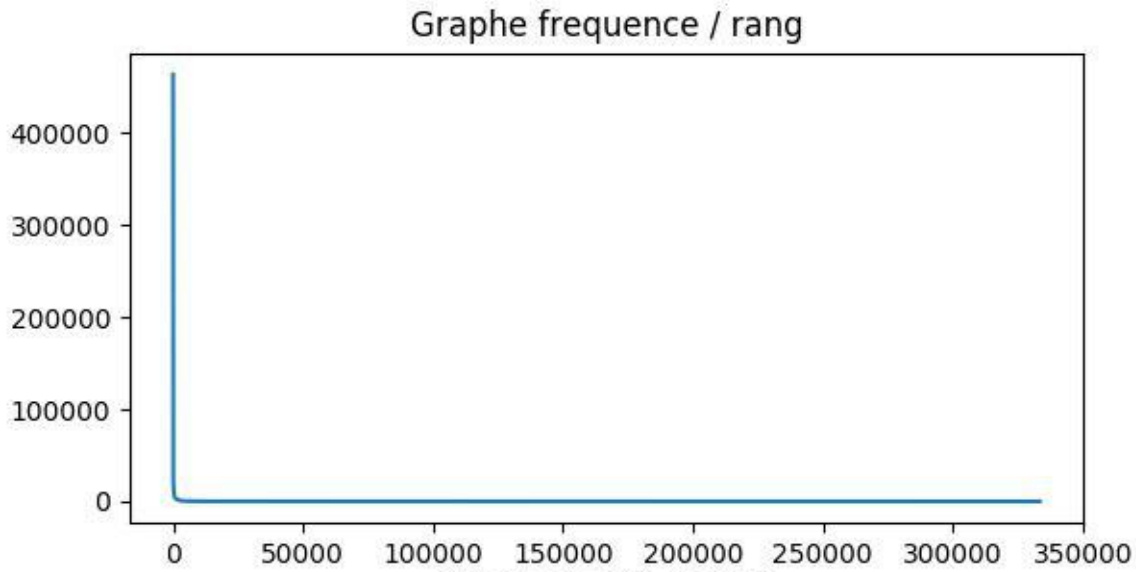
$$K = \frac{333.828}{26.517.667^{\beta}} = \boxed{2.56}$$

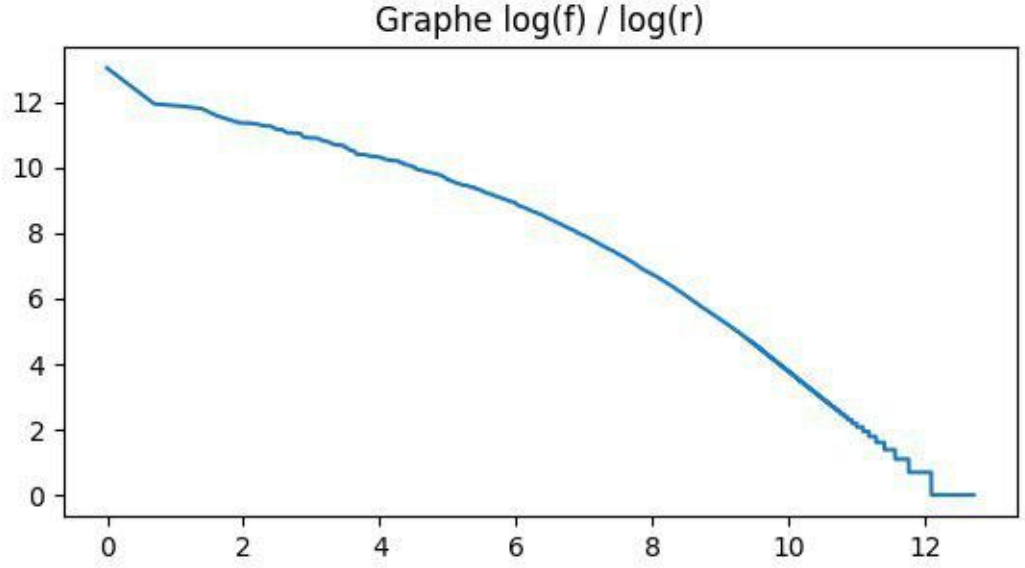
1.2.4 Estimer la taille du vocabulaire pour une collection de 1 million de tokens.

D'après la loi de Heap, on a donc pour la collection CACM:

$$V_R(10^6) = 2.56 \times 10^{6 \times 0.689} = \boxed{34853 \text{ mots}}$$

1.2.5 Tracer le graphe fréquence (f) vs rang (r) pour tous les tokens de la collection. Tracer aussi le graphe log(f) vs log(r).


Graphe frequence / rang
400000
300000
200000
100000
0
0
50000
100000
150000
200000
250000
300000
350000


Graphe log(f) / log(r)
12
10
8
6
4
2
0
0
2
4
6
8
10
12