OSY 2018-2019

Fondements de la Recherche d'Information-WEB

Hugo Martinet

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

CACM

Question 1

La collection CACM contient $T = 131\ 915$ tokens. Ici, le choix a été fait de retirer les caractères spéciaux, i.e. *user-oriented* devient un seul mot *useroriented* (et non pas *user* et *oriented*) etc.

Question 2

La taille du vocabulaire de la collection CACM est $M = 10\ 662$. Mêmes hypothèses que pour la question précédente.

Question 3

Pour la moitié de la collection, il y a $T_{\frac{1}{2}} = 41\ 264$ tokens, et la taille du vocabulaire est $M_{\frac{1}{2}} = 5\ 942$.
Or

$$M = kT^b$$

$$M_{\frac{1}{2}} = kT_{\frac{1}{2}}^b$$

Alors

$$b = \frac{log(10\ 662) - log(5\ 942)}{log(131\ 915) - log(41\ 264)}$$

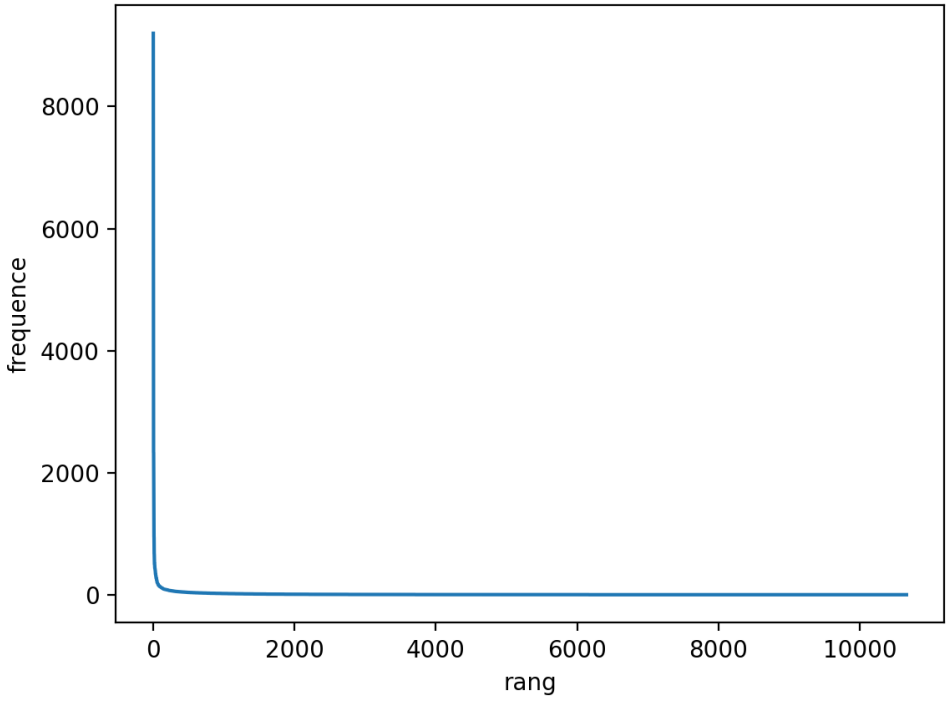
Ainsi $b = 0,5031$ et donc $k = 28,3$.

Question 4

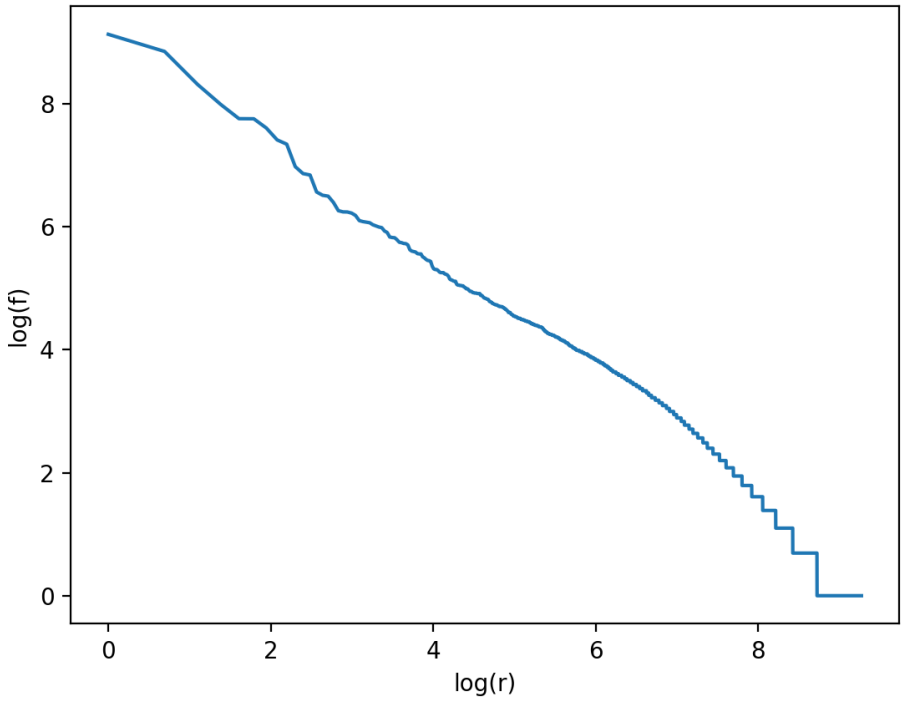
À partir de la question précédente, si $T = 1\ 000\ 000$, alors $M = 29\ 538$.

Question 5

Graphe de la fréquence *(f)* en fonction du rang *(r)*



Graphe de *log(f)* en fonction de *log(r)*



Cs276

Question 1

La collection Cs276 contient $T = 22\ 648\ 710$ tokens.

Question 2

La taille du vocabulaire de la collection Cs276 est $M = 354\ 383$.

Question 3

Pour la moitié de la collection, il y a $T_{\frac{1}{2}} = 10\ 531\ 367$ tokens, et la taille du vocabulaire est $M_{\frac{1}{2}} = 207\ 962$.
Or

$$M = kT^b$$

$$M_{\frac{1}{2}} = kT_{\frac{1}{2}}^b$$

Alors

$$b = \frac{log(354\ 383) - log(207\ 962)}{log(22\ 648\ 710) - log(10\ 531\ 367)}$$

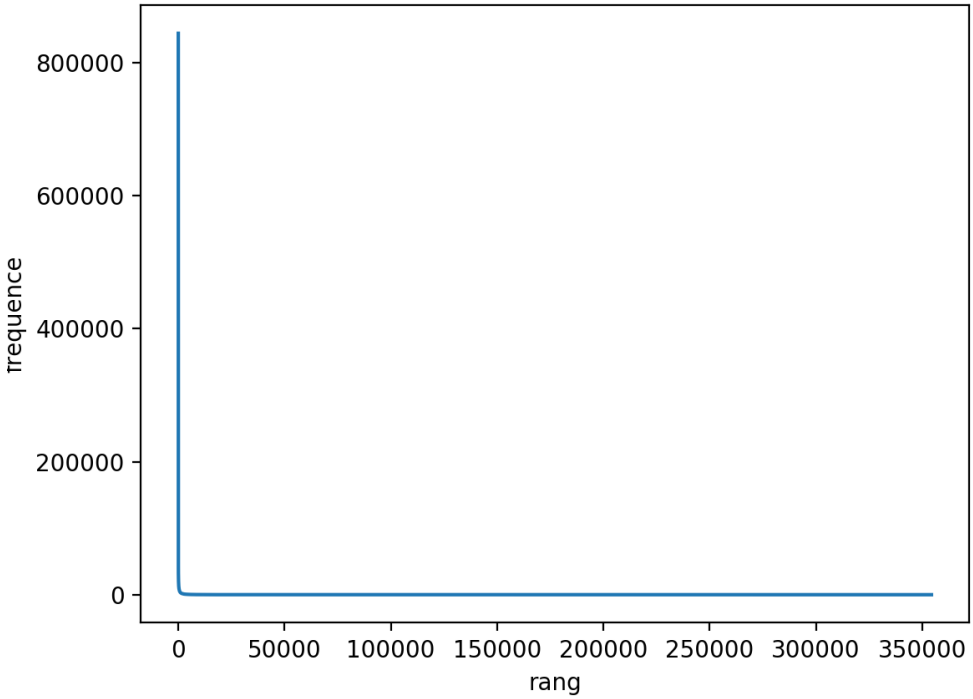
Ainsi $b = 0,6960$ et donc $k = 2,68$.

Question 4

À partir de la question précédente, si $T = 1\ 000\ 000$, alors $M = 40\ 387$.

Question 5

Graphe de la fréquence *(f)* en fonction du rang *(r)*



Graphe de *log(f)* en fonction de *log(r)*