Task3

Nawar Saeed

Artificial Intelligence for the Web, VT21
DT506A

16 april 2021

As the table below shows, the words are very common and it is natural that they will be the most frequent words. In the previous task, these words led to many numbers of hits, so it is no wonder that they are considered the most frequent words. To find out α, the formula of zip'f low can be used.

$$Fn = c + n^{-\alpha}$$

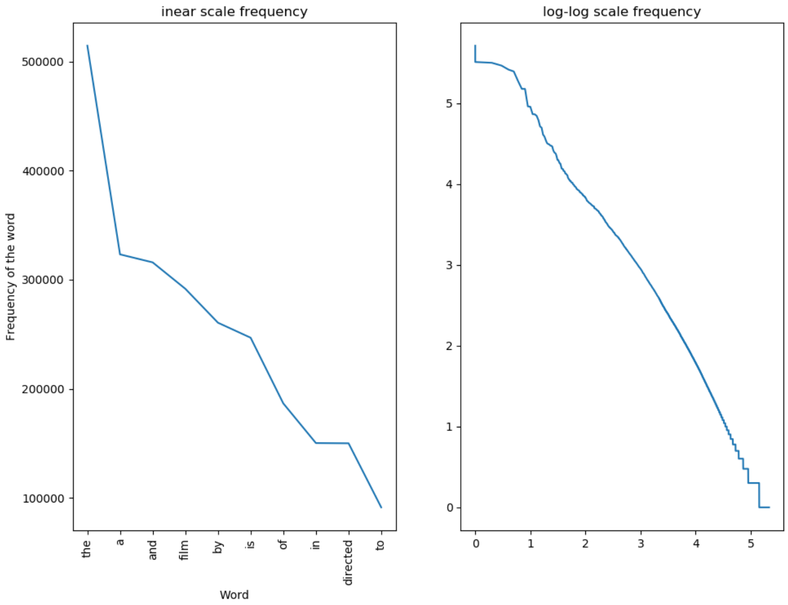
where Fn is the total of frequency of the words, n number of words and $c = \frac{1}{\ln m}$. From this formula, α can be factorized by multiplying by log on both sides and get α, as line 81 in Task3.py shows.

The ten most frequent word									
the	a	and	film	by	is	or	in	directed	to
514433	323219	315882	291694	260896	246857	186705	150305	150099	91405

The estimeated valueof the factor α:

$$\alpha = 1.287545187472393$$

zip'f law says that the curve should be or converge towards a linear appearance, which is almost true above according to the right graph below. In addition, alpha is obtained at 1,287, which may support our assumption.



Figur 1: Plot of words frequency linear scale vs. log-log scale