Task15

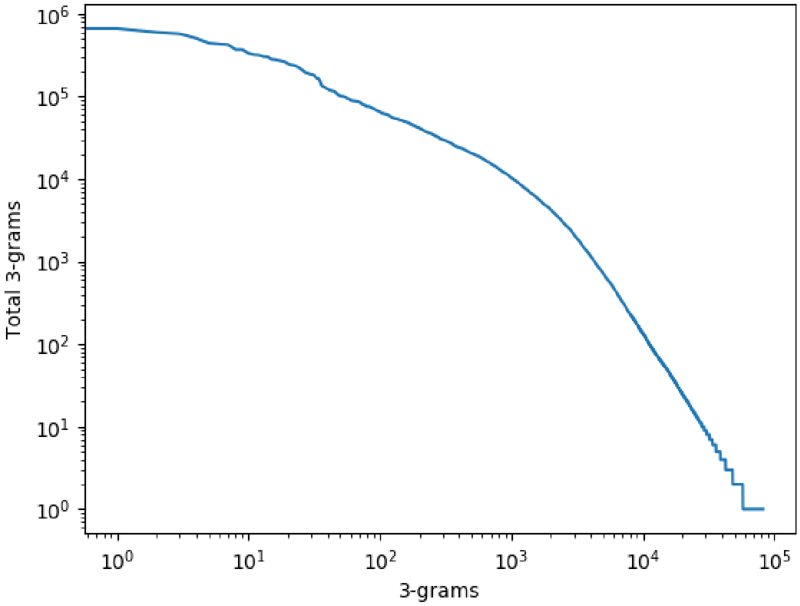
Nawar Saeed

Artificial Intelligence for the Web, VT21
DT506A

16 april 2021

This task is about implementing a q-gram index, which is implemented according to line 70 in Task15.py. The code is documented, so I guess it is possible to follow up what is happening where the ego to only read the comments. It is colored according to how many 3-grams it has been created (including the padding) and it has been obtained according to my implementation 52391342 3-grams. The most frequent 3-grams are shown according to the table below. In addition, it asks for the number of words beginning with the letter K, which I got it to 188267 words beginning with this letter. Finally, it is asked to plot the log-log scale of the frequency of 3-grams, which can also be seen below. As it turns out, the curve does not have a linear appearance, thus it contradicts what Zipf's law says.

The top-10 3-grams									
n$$	e$$	s$$	a$$	$$s	$$m	$$c	$$a	r$$	$$b
680884	665124	597939	572969	506367	443555	430980	422520	369866	369435



Figur 1: log-log plot of the frequency of 3-grams