**Find the % of overlap with standard list of stop words and your list of stop words – 10/08/2021**

- Take a suitable corpus (~ 1 GB to begin with)

- Find the frequency of all distinct tokens **(you have done it in last lab)**

- It is expected the stop words will have more frequency

- Plot the histogram of frequencies (in sorted order)

- Choose a cut-off frequency (say 0.6)

- All tokens with a frequency above cut-off are likely to be stop words and those are below the cut-off are not likely to be stop words.

- Thus, based on the cut-off, you can prepare your own stop word list

- Now download  [NLTK's list of English stopwords · GitHub](https://gist.github.com/sebleier/554280)

<https://gist.github.com/sebleier/554280>

- Compare your list with the standard list and find the % of overlap.

- Write the analysis based on the comparison