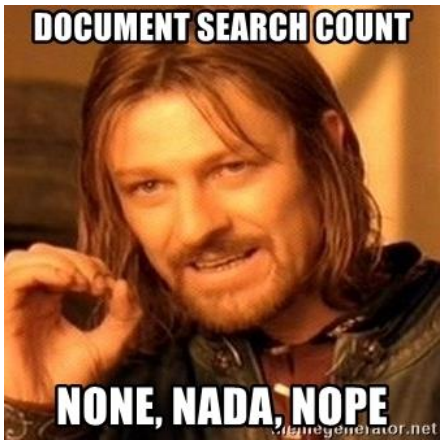
Assignment 2 – Dataset preparation Part 2

Objective: Build a text search engine dataset.



Step 1: **Download** at least 50 documents belong to different formats such as txt, doc, pdf, html etc and keep it in a single folder.

Step 2: **Read** the documents using different parsing mechanisms and keep them in an array.

Step 3: **Standardize** the dataset using following pre-processing techniques.

- Remove all the special characters, smileys, and keep only alpha numeric values in the txt.
- Convert upper case letters into lower case letters.
- Remove the words which are not in this English dictionary with half million words https://raw.githubusercontent.com/dwyl/english-words/master/words.txt

Step 4: **Receive** query from the user in terms of single keyword. Apply the same standardization technique in the previous step 3 for the query from user. Search whether any document in our dataset contains keyword from the user. Do ranking for the documents in our dataset based on number of times the word is found in a document.

Step 5: **Display** the documents in sorted order with highest matching document in the top.

Step 6: **Perform** the step 4 and 5 for any query with more than one keyword.

Suggested Platform: Python: Azure Notebook/Google Colab Notebook, packages such as numpy, nltk, regular expression package re.

Marking: Marking is based on both **performance during the lab hours** as well as **complete submission**.