1 Text Cleaning and PreProcessing

Following tasks are to be performed with Python packages numpy, pandas, scipy, re, string.

1.1 Google Search Results

Delivery 1: Feb 17, 2022 on CV of python file and pdf file(optional).

1.1.1 Text Collection

1. Generate a corpus C of documents d_i by using google search results for any topic.
2. Each document d_i contains special characters, numbers, commas, spaces and Nan values.
3. Each documents has term count of tokens w_i greater than 10.
4. Create the document d_i in one of the following form:
   - Search results
   - Question Answers
   - Email
   - Book chapter page
   - Blog post
   - Product overview
5. Read the documents from corpus C.

1.1.2 Text Preprocessing

1. View the data and specify the data preprocessing requirements.
2. Check for lower and upper case.
3. Remove punctuations, commas and special characters etc.
4. Remove the frequent words, rare words, stopwords and Nan values if any.
5. Clean up the data for other requirements.