1. What is the difference between text mining and text analytics?
2. What is a corpus?
3. List the text cleaning tasks?
4. Differentiate between stemming and lemmatization?
5. Name of the Dictionary which lemmetizer of nltk uses?
6. list different types of tokenizers
7. Suppose you are trying to summarize a text document which of the below techniques would you use?
   1. Stemming
   2. Lemmatization
8. Which of the below techniques helps to reduce the number of distinct words in a document suitable for plotting graphs based on frequencies
   1. Stemming
   2. Lemmatization
9. what is the output of the below code, assume all\_terms is a list of words in the entire corpus

from nltk.probability import FreqDist

fdist = FreqDist(all\_terms)

fdist

1. Customize the stop words such that words "no, nor, not" are not removed from the corpus
2. Write a UDF which returns a dictionary with frequency count of each of the stop words used in the corpus.