Name: Sai Anish Garapati
UIN: 650208577

1.a) The program submitted along, processes the Citeseer UMD corpus by tokenizing the text and grouping the words based on their frequencies. The following steps were done:

-> All the text from files is read using the **os** library from python and a string is created from the text.

-> **word_tokenize** module from **nltk.tokenize** is used to tokenize the text on whitespace to form a list of words.
Function: word_tokenize(string).

-> All the punctuations from the words are removed based on the punctuations from the list **string.punctuation** `(!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~)` from the string library**.**

-> From the resultant list of words which is also the Collection, a dictionary is created with the word as key and its frequency in the collection as value, which forms the Vocabulary.

-> Stop words are removed from the vocabulary based on the **stopwords** module from **nltk.corpus** library which are:
{'all', "haven't", 'doesn', 'by', 'below', 'we', 'than', 'because', 'haven', 'it', 'down', 'further', 'don', 'what', "that'll", 'its', 'not', "aren't", 'weren', 'my', 'how', 'ours', 's', "doesn't", 'needn', 'having', 'when', 'him', 'at', 'once', 'wasn', 'the', 'll', 'has', 'until', "shouldn't", 'hasn', 'myself', 'are', 'you', 'didn', 'won', 'whom', 'most', 'so', "mightn't", 'be', 'been', 'ourselves', 'a', 'no', 'same', 'who', 'yours', 'nor', 'themselves', 'me', 'between', 'to', 'mightn', "don't", 'above', 'this', 'm', 'she', "you've", 'about', 'or', 'being', 'shouldn', 'yourself', "wasn't", 'in', 'where', 'those', "you'd", 'off', 'just', 'ma', 'he', "isn't", "she's", 'their', 'isn', "wouldn't", "needn't", 'our', "hadn't", 'were', 'for', 'but', 'o', 'i', 'such', 'them', "you'll", 'other', 'yourselves', 'own', 'during', 'these', 'out', 'herself', 't', 'through', 'very', 'both', 'wouldn', 'why', 'did', 'as', "it's", 'there', 'itself', 'before', "won't", 'only', 'have', 'ain', 'any', 'of', 'shan', 'couldn', 'mustn', 'that', 'over', "should've", "couldn't", 'd', 'now', "shan't", 'hers', "mustn't", 'each', 'then', 'which', 're', 'do', 'from', 'here', 'if', 'her', 'few', 'am', 'up', 'hadn', 'after', 'some', 'an', 'does', 'again', 'while', 'theirs', 'they', "you're", 'is', 'too', 'and', "weren't", 'y', 'aren', "didn't", 'was', 'doing', 'with', 'can', 'will', 'against', 'his', 'should', 'your', 'himself', 'into', 'more', "hasn't", 've', 'under', 'had', 'on'}

-> **PorterStemmer** from **nltk.stem** is used to stem the words and regroup the words in the dictionary based on the stemmed words.
Function: PorterStemmer().stem(word)

1.b) The uploaded code can be run from the terminal using the '**python file.py**' command from the directory containing the citeseer folder. If the file were to run from any other directory, the path variable in the code should be updated accordingly.

2.a) The total number of words in collection are 477989

2.b) The vocabulary size is 19630

2.c) Top 20 words in the ranking and their respective frequencies in the collection:
{'the': 25667, 'of': 18643, 'and': 14134, 'a': 13372, 'to': 11539, 'in': 10069, 'for': 7382, 'is': 6580, 'we': 5147, 'that': 4821, 'this': 4447, 'are': 3738, 'on': 3653, 'an': 3281, 'with': 3200, 'as': 3060, 'by': 2767, 'data': 2694, 'be': 2500, 'information': 2326}

2.d) Stop words from the top 20 words:
['the', 'of', 'and', 'a', 'to', 'in', 'for', 'is', 'we', 'that', 'this', 'are', 'on', 'an', 'with', 'as', 'by', 'be']

2.e) The count of unique words accounting for 15% of the total words in the collection is 0.

3.a) The total number of words in the new collection are 294927

3.b) The new vocabulary size is 13625

3.c) Top 20 words in the ranking and their respective frequencies in the new collection:
{'system': 3745, 'use': 3741, 'agent': 2695, 'data': 2694, 'inform': 2402, 'model': 2314, 'paper': 2247, 'queri': 1905, 'user': 1758, 'learn': 1742, 'algorithm': 1584, '1': 1569, 'problem': 1545, 'approach': 1544, 'applic': 1524, 'present': 1507, 'base': 1499, 'web': 1440, 'databas': 1425, 'comput': 1414}

3.d) There are no stop words from the Top 20 words from new collection

3.e) The count of unique words accounting for 15% of the total words in the collection is 0.