Use the following dataset - [https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews](https://www.google.com/url?q=https%3A%2F%2Fwww.kaggle.com%2Flakshmi25npathi%2Fimdb-dataset-of-50k-movie-reviews)

**# Problem 1:**

# Apply all the preprocessing techniques that you think are necessary

**# Problem 2**

# Find out the number of words in the entire corpus and also the total number of unique words(vocabulary) using just python

**# Problem 3**

# Apply One Hot Encoding

**# Problem 4**

# Apply bag words and find the vocabulary also find the times each word has occurred

**# Problem 5**

# Apply bag of bi-gram and bag of tri-gram and write down your observation about the dimensionality of the vocabulary

**# Problem 6**

# Apply tf-idf and find out the idf scores of words, also find out the vocabulary.