Lab Exercise 6 : Basic Vectorization

In this lab exercise, we will vectorize the corpus using one-hot encoding and bag of words.

1. For the **text** in the file given **LabE6.txt**(1000 reviews from IMDB dataset):
    a. Apply preprocessing.
    b. Create **one-hot encoded vectors** for the each tokens in the **vocabulary**.
2. Apply newline tokenization to the **text** (use split("\n"). Consider **each element** in the list as a **document**.
    a. Apply preprocessing.
    b. Create **BoWs** vectors for each of the documents
3. Read a **search text** from the user
    a. Using cosine similarity : List the top five **similar documents** based on the **search text**

Optional : Create your own module where all pre-processing functions. You may use them in your code by just importing. Refer here.