Please refer to the “***Sheet1***” worksheet of ***Data.xlsx*** file provided. The data contains text reports from an assignment from 37 students (S1 to S37).

A1. **Tokenization:** Load the documents for all the 37 students. Tokenize all the documents and store the tokens. Let this be a nested list called ***“Documents”***

A2. **Token Population:** Merging the tokens from all documents, create a master list of distinct tokens available across all documents. Let us call this as ***“token population”. Print the length of this list as V***

A3. **Bag-of-Words:**

Create a **“*bag-of-words”*** from the **“*token population”***  by removing the ***stop-words***

A4. **Document Vectorization:** For each document / instance, create 2 feature vectors as follows

* First vector attributes indicate the presence / absence of tokens from ***bag-of-words*** in that document. Denote this as 𝑉1𝑑𝑤where d denotes the document and w denotes the word in ***bag-of-words***; V1 denotes the vector of first kind. Assuming there are M tokens in **bag-of-words**, the vector should look like:



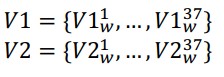
* Second vector attributes indicate the count of presence of each word from ***bag-of-words*** in the document. Denote this as 𝑉2𝑑𝑤

where d denotes the document and w denotes the word

in ***bag-of-words***; V2 denotes the vector of second kind. Assuming there are M tokens in **bagofwords**, the vector should look like:



Do this for all documents. After this is done, you should have 37 vectors each of first and second kind. This may be denoted as below.



A5. Calculate the Cosine similarity between the documents by using the second feature vector for each document

sklearn.metrics.pairwise.cosine\_similarity

