Name – Dhruv Singhal **B.Tech. (C.S.E.) AI&ML**

SAP ID – 500075346 **Roll No. - R177219074**

Lab Experiment – 08

Source of Corpus:

The IIT Bombay English-Hindi corpus contains parallel corpus for English-Hindi as well as monolingual Hindi corpus collected from a variety of existing sources and corpora developed at the center for Indian Language Technology, IIT Bombay over the years. This corpora can also be obtained from Kaggle:
https://www.kaggle.com/vaibhavkumar11/hindi-english-parallel-corpus

Importance of Corpus:

A corpus is a large and structured set of machine-readable texts that have been produced in a natural communicative setting. While most available corpora are text only, there are growing numbers of multimodal corpora. A corpus can help us in basic analysis, like counting number of words, frequency of words.
Corpus is important in particular study of spoken language as the written language can be studied by various examining methods but we need corpora to study interactive conversation.

Statistics of Corpus:

Number of unique sentences of English: 1015949
Number of unique sentences of Hindi: 983941

Preprocessing in Corpus:

1. Lower casing:

   The lowercasing is an important text pre-processing step in which we convert the text into the same casing, preferably all in lowercase so that the words “INDIA”, “INDia”, and “India” can be treated in the same way as “india”. It prevents duplication of same words having different casing.

2. Remove extra white space and punctuations:

   A text may contain extra whitespace which is not desired as they increase the text size and not add any value to the data. Hence removing extra whitespace is a trivial but important text pre-processing step. Removing punctuation is an important text pre-processing step as it also does not add any value to the information. This is a text standardization process that will help to treat words like ‘some.’, ‘some,’, and ‘some’ in the same way.

3. Tokenization:

Tokenization is the process of splitting text into pieces called tokens. A corpus of text can be converted into tokens of sentences, words, or even characters. Tokenization breaks the raw text into words, sentences called tokens. These tokens help in understanding the context or developing the model for NLP. The tokenization helps in interpreting the meaning of the text by analyzing the sequence of words.

4. Stop word Filtering:

   Stop words are trivial words like “I”, “the”, “you”, etc. that appear so frequently in the text that they may distort many NLP operations without adding much valuable information. So almost always you will have to remove stop words from the corpus as part of your pre-processing.

5. Lemmatization:

   Lemmatization is converting the word to its base form or lemma by removing affixes from the inflected words. It is a technique in which we try to find the base word of a inflectional word. It helps to create better features for machine learning and NLP models hence it is an important pre-processing step.

   Lemmatization reduces the word forms to linguistic valid lemmas, like ‘great’ will have a lemma i.e. ‘good’.

6. Stemming:

   Stemming also reduces the words to their root forms but unlike lemmatization, the stem itself may not be a valid word in the language. It is a technique where a set of words are converted to a sequence to shorten its lookup.

   A stemmer operates on a single word without having any knowledge of its context, it reduces word forms to stems in order to reduce the size.