ADVANCED ALGORITHMS MINI PROJECT

COURSE CODE : UE16CS311

TEAM ID : 008

Team Member	Name	SRN
1.	Abhijith Venugopal	01FB16ECS008
2.	Abhishek Narayanan	01FB16ECS016
3.	Rachana Aithal K R	01FB16ECS483

PROJECT TITLE : Implementing a library for Natural Language Processing for English Language

PROJECT ABSTRACT :

This mini project is aimed at creating a user friendly library in C Programming Language for performing standard operations used often in real word problems in the field of natural language processing. The project library essentially consists of the following modules, with each module targeted towards different state-of-the-art NLP tasks, which have been elaborated below :

1. Text Pre-processing Module :

a. Noise Removal : Raw text consists of lot of noise such as special characters or punctuation to name a few(Example : hashtags in tweets, which are often irrelevant to the task at hand of analysing text for any type of classification. Also, there exist various words such as a, the, you, me ,my etc which are irrelevant with respect to various classical problems in text analytics.
Hence input text is first tokenised based on a specified delimiter and noise and stop words are eliminated to clean raw text in this module.

b. Lexicon Normalization : In English language, it is often observed that sentences have different structure of words but convey similar meaning. For example, two sentences might use words like have or having, but yet the context and semantics of the sentence is same in both. However, such inconsistency in raw text in natural languages cannot be learnt by NLP algorithms and hence the text data is to be made consistent before further processing is done.
In this module, we implement the state-of-the-art Porter's Stemming Algorithm to solve this problem of inconsistency, wherein, similar words are reduced to a canonical root word (example have and having can be reduced to hav). Though the root word may be meaningless, yet it ensures consistency of words in text data.

2. Feature Extraction Module :

With the rampant explosion of text on online social media, text processing, classification and analytics has become an inevitable task. Since machine learning algorithms or most data analytical frameworks are incapable to process and understand raw text, the text has to be transformed into vector space and represented as numeric vectors having a fixed dimension.
This library module achieves feature extraction from an input text corpus using the following techniques:

a. Entity parsing based techniques : Bag-of-Words (BoW) , Bag of N-grams

b. Statistical techniques : Term Frequency- Inverse Document Frequency (TF-IDF)
The above algorithms are capable of generating high dimensional sparse feature vectors to represent text in vector space.

3. Text Similarity Module :

One of the important areas of NLP is the matching of text objects to find similarities. Important applications of text matching includes automatic spelling correction, data de-duplication and genome analysis etc.

This module implements Levenshtein Distance and cosine similarity to find similarity or the degree of closeness one chunk of text possesses with another.