Vishwa Singh
PA10 (1032170273)
CCNLP1

# Cognitive Computing and Natural Language Processing

## Assignment No. 1

**Problem Statement:** Program to read a paragraph from a text file. Print the paragraph after removing the stop words. Identify POS ~~in~~ of each word in the paragraph.

**Objective:**

1) To study and explore NLTK for text processing
2) To learn concepts text processing in NLP

## Theory:

**Tokenization:** It is a way of saperating a piece of text into smaller units called tokens. Tokens can either be words, characters ~~and~~ or subwords.

**Stemming** : It is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of ~~wods~~ words know as a lemma.

**Lemmatization** : Usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove ~~inflect~~ inflectional endings only and to return the base form of the word.

**POS tagging** : POS tagging refers to the process of identifying the part of speech tag for a word.

**Stop words removal** : It is the process of removing the words that ~~can~~ and might not hold significant bearing to the sentence.

**Bag of Words** model : It is a method of extracting features from text for use in modeling, such as machine Learning algorithm.

Examples :

Bag : ["it" "was" "the" "best" "of" "times" "age" "worst"]

Sentence : "It was the best of times" : [1 1 1 1 1 0 0]
"It was the worst of times" : [1 1 1 0 1 0 1 0]
} Bag Entries

★ NLTK Modules :

1) Corpora : A package containing modules of example text
2) Tokenize : Seperate text strings
3) Stem : Stem words of text
4) Chunk : Identify short non-nested phrase
5) Cluster : Clustering algorithm
6) tag : Used for POS tagging

**Platform** : Jupyter Notebook
**Input** : Text / Paragraph in English
**output** : Token, Text after removing stop words, token with POS tag & Stem from text

## FAQ

### Question 1

=> **Stemming** algorithm works by cutting off the end or the begining of the word, taking into account a list of common ~~words~~ prefixes and suffixes that can be found in an inflected word.

**Lemmatization**, on the other hand, takes into consideration the morphological analysis of the words. To do so, it is necessary to have detailed dictionaries which the algorithm can look through to link the form back to its lemma.

Example :

* Stemming — * Lemmatization

chang ← { Change
changing
Changes
changed
changer } → change

Question 2

Semantic Analysis : It is a process of drawing meaning from the text. It allows computers to understand the language by analyzing their grammatial structure, and identifying relationships between individual words in a particular context

Syntactic Analysis : It is the process of analyzing natural language with the rules of formal grammar. Grammatical rules are applied to categories & groups of words.