IT412: Natural Language Processing

Assignment 3: Stemming

Instructor: Prasenjit Majumder

Lab Outcome: After performing this assignment you will be able to perform Stemming on your own

1 Problem description

In linguistic morphology and information retrieval, stemming is the process of reducing inflected words to their word stem, base or root form—generally a written word form. Stemming algorithms are commonly called stemmers. A stemming algorithm reduces words 'likely', 'likes', 'liked', 'liking' to its root form 'like'.

2 Implementation

2.1 Dataset

For this assignment you are provided a file which contains list of words.

2.2 Exercise

1. For given list of word compute their string distance using Levenshtein distance. For computed distance perform hierarchical clustering and find center word which would be stem word for all other words in cluster.

2. For given words find its stem word using NLTK PorterStemmer and compare its results with stemming done using clustering.

3 References

- YASS

- Levenshtein distance

- https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html