IT412: Natural Language Processing

Assignment 2: Stemming

Instructor: Prasenjit Majumder

Learning Outcome: At the end of this assignment you will learn about hierarchical clustering, distance measures for measuring the closeness of words and also implement a stemmer from scratch.

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 task use the nltk brown corpus words of "news" category
- For steps to extract words from the brown corpus use this link: https://www.nltk.org/book/ch02.html
- For retrieval task use the dataset used in the previous assignment
- For Word2Vec embeddings download the embeddings using the code in the link: https://radimrehurek.com/gensim/auto_examples/tutorials/run_word2vec.html

2.2 Exercise

1. You will first implement YASS stemmer.
2. For given list of word compute their string distance using D1, D2, D3, D4 and Levenstein distance. (Information about the YASS stemmer can be found in the paper)
3. For computed distance perform hierarchical clustering (use any of the measurement single linkage, average linkage and complete linkage) and find center word which would be stem word for all other words in cluster.
4. Generate a graph of no. of clusters vs threshold at which cluster merging is stopped.
5. Perform stemming on the FIRE dataset using YASS stemmer using D1, D2, D3, D4 and Levenstein distance
6. Use Terrier to index the files and then perform the retrieval task. Report the MAP scores for D1, D2, D3, D4 and Levenstein distance.
7. Use Word2Vec to represent the words and perform Agglomerative clustering based on Word2Vec to find the root word. Perform stemming on FIRE dataset and then report the MAP score for stemming using Word2Vec.
8. Find the root word of the words in FIRE dataset using Porter stemmer. Then report the MAP score for stemming using Word2Vec.

3 References

- YASS yet another suffix stemmer
- https://dzone.com/articles/the-levenshtein-algorithm-1
- https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html
- http://terrier.org/
- https://github.com/terrier-org/terrier-core
- https://radimrehurek.com/gensim/auto_examples/tutorials/run_word2vec.html
- https://www.geeksforgeeks.org/python-stemming-words-with-nltk/
- https://newbedev.com/how-to-speed-up-gensim-word2vec-model-load-time

4 Submission

- Your notebook should contain the graph of no. of clusters vs threshold
- Show the MAP scores for retrieval using YASS stemmer for D1, D2, D3, D4 and Levenstein distance, Porter stemmer and Word2Vec stemmer.
- The submission deadline for this assignment in **30th August 2021 at 11 PM**