Natural Language Processing - Assignment 1

Deepan Chakravarthi Padmanabhan

Assignment submission date: April 08, 2020

1 Questions on Chapter: Regular Expressions, Text Normalization, and Edit Distance

1. What is coreference resolution?
2. Are all examples of lemma, are examples of different wordforms (inflected words)?
3. V is the set of words in the vocabulary. |V| is the number of distinct words in the vocabulary. N is the total number of running words. Explain the difference between V, |V| and N by providing an example.
4. Explain the differences and implications of Herdan's law and Heap's law with a corpora example?
5. How can number of lemmas in a language is a measure of words in a language instead of wordform types? Give examples.
6. In lookahead assertions, what is meant by match cursor movement?
7. What is named entity extraction, named entity detection and named entity coreference?
8. The following tokenization (word segmentation) methods are discussed in the chapter - Unix tools for crude tokenization, Penn Treebank Tokenization, Byte-Pair Encoding (BPE) for tokenization, wordpiece tokenization, MaxMatch, SentencePiece. What are the other tokenization algorithms/methods?
9. What is the difference between morphemes and subwords? Give examples.
10. What is the time complexity of BPE? It looks tiresome for a large corpus? Comment.
11. What is the difference between information extraction and information retrieval?
12. Is a particular tokenization method efficient than others? How to choose the best method for our task? Can you provide some example case?
13. Is there any restrictions and assumptions in applying minimum edit distance algorithm?

2 Motivation to take up the course

1. Natural Language Processing (NLP) includes processing of more unstructured data such as text and speech compared to images. The process of analyzing and modelling unstructured data is more challenging and an imperative technical skill to learn.
2. NLP has a wide range of applications namely, machine translation, speech recognition, sentiment analysis, question answering, automatic summarization, chatbots, market intelligence, text classification, character recognition, image captioning, and spell checking.
3. NLP is widely integrated with the large number of educational contexts such as research, science, linguistics, e-learning, evaluation/grading system, and contributes resulting positive outcomes in other educational settings such as schools, higher education system, and universities.

3 Expectations about the course

1. The course would help to understand the basic NLP methods.

2. Design and build applications and systems that enable interaction between machines and natural languages used by humans.

3. Provide hands-on experience by implementing an existing state-of-the-art work in NLP apart from the assignments and classes (similar to the project work). This can be extended to accommodate some healthy competition scenario among students by bring in some metrics and competing against student groups (similar to the competitions conducted for Computer Vision classes but the competition should include the concepts taught in the course work and assignments provided).

4. An ultimate expectation would be provide scope (project title, time and resources) to take ownership and successfully complete a project work that contributes to the NLP open source community using Deep Learning by solving a problem in NLP. A major barrier is selecting the project title. However, if a list of possible titles/problems are provided, this can be achieved in the stipulated time frame.