CS F469, Information Retrieval: Assignment-1

Instructor: Abhishek (abhishek@pilani.bits-pilani.ac.in)
Topic: Text Pre-processing

Due Date: 7^{th} **February,** 2020

The objective of the assignment is to get started with the necessary tools and techniques required to work with unstructured text corpus. You are given a collection of unstructured documents obtained from English Wikipedia.

You are free to work on the assignment using any programming language and any open-source text processing libraries or toolkits (unless otherwise stated). A list of popular libraries is mentioned at the end of the assignment.

Corpus Statistics

Extract the text from the corpus. Note that the dataset contains Html $< a >$ tag, which needs to be parsed to extract the corresponding text. After text extraction, tokenize it and answer the following questions:

1. **Unigram analysis:**
   (a) Mention the total unique unigrams present in the corpus.
   (b) Plot the distribution of the unigram frequencies.
   (c) How many (most frequent) uni-grams are required to cover the 90% of the complete corpus.
2. **Bigram analysis:**
   (a) Mention the total unique bigrams present in the corpus.
   (b) Plot the distribution of the bigram frequencies.
   (c) How many (most frequent) bi-grams are required to cover the 80% of the complete corpus.
3. **Trigram analysis:**
   (a) Mention the total unique trigrams present in the corpus.
   (b) Plot the distribution of the trigram frequencies.
   (c) How many (most frequent) tri-grams are required to cover the 70% of the complete corpus.
4. Repeat (1), (2) and (3), after performing the stemming process on the tokens.
5. Repeat (1), (2) and (3), after performing the lemmatization process on the tokens.
6. Briefly summarize and discuss the frequency distributions obtained in (1) to (5). Do these distribution approximately follow the Zipf's law?
7. From the corpus, report three examples based on you observation where the tool used for tokenization did not tokenize the character sequence properly.

8. Which tool/library you used for tokenization, stemming and lemmatization? What are the underling algorithms this tool/library use for tokenization, stemming and lemmatization?

9. From the corpus, analyse and briefly summarize how the tool tokenizes dates and numeric values (especially related to currencies). (No need for an exhaustive analyse, an analysis consisting of 5 different examples would be sufficient.)

10. Find top 20 bi-gram collocations in the text corpus using Chi-square test. Do not use any libraries. References for Chi-square test for finding collocations:

    (a) Slide no. 25, 26 and 27 from `http://www.cse.unt.edu/~tarau/teaching/NLP/05Collocations.pptx`.

    (b) Section 5.3.3 of the book chapter `https://nlp.stanford.edu/fsnlp/promo/colloc.pdf`.

List of tools/libraries

- Python NLTK (`http://www.nltk.org/`)
- CoreNLP (`http://nlp.stanford.edu/software/corenlp.shtml`)
- Apache OpenNLP (`http://opennlp.apache.org/`)
- spaCy (`https://spacy.io/`)

Corpus: English Wikipedia in partially processed form and divided into approx 1500 files is available at `https://drive.google.com/drive/folders/1ZsnuEm7_N6aUwhjFpv-TZXFt4DiYex4t?usp=sharing`. For this assignment, you can randomly choose any one file from the sub-folders available at the mentioned link.

Assignment submission guidelines will be shared by 24th January 2020.