CS 565: Assignment-1

Basic pre-processing: Segmentation, N-Gram Analysis, Collocation

Due Date: 2^{nd} **February,** 2016

An *n-gram* is a sequence of n items from a given sequence of word or text. The items can be letters, words or akshara according to the application, for our case it would be words only. For instance lets say corpus is *"His relationship with many western nations was troubled during his tenure as chief minister..."* then the list of unigrams would be { *His, relationship, with, many, western...*}, bigrams would be { *His relationship, relationship with, with many,...*} and similarly trigrams would be {*His relationship with, relationship with many, with many western...* }. The given example of *n-grams* are contiguous, whereas we can also create a list of non-contiguous *n-grams* as we discussed in the class. While non-contiguous *n-grams* can be useful for finding collocations, contiguous *n-gram* analysis is more common for downstream applications such as language modeling, machine translation, text categorization etc.

Objective of the assignment is to get started with basic NLP tasks, exploring available tools and choosing the one which you like the most. Some of the famous tools are available on the course website. For example, NLTK is famous nlp toolkit for python language; Stanford corenlp tool is from the famous Stanford NLP group, Apache openNLP etc. For "Getting Started", Sunil has chosen the reference material for tokenization and *n-gram* calculation by Dragomir's notes.

Finer points of assignments are as follows:

- Work in a group of 2 students.
- For all other assignments, the same set of students should be working in a group, i.e., once you form a group for assignment, you are not allowed to change the group.
- Submit the joint report and code.
- Any assignment related queries and discussions will not be entertained on my personal email-id. You should use Canvas for the same.
- No late submission will be entertained.
- You will be informed about submission procedure of your report and code by a separate post on Piazza or Canvas.

1 Getting Started: Part I

1. Download any freely available corpus or take any accessible corpus from your chosen tool. Explore available sentence segmenter from the tool and use that for sentence segmentation.
2. For each sentence, detect all words in it. Create a dictionary out of it.
3. Find all possible unigrams. For each unigram, calculate its frequency of occurring in the given corpus. Then arrange the unigrams in monotonically decreasing order of frequency.
4. Find all possible bigrams and calculate their frequencies. Then arrange them in monotonically decreasing order of frequency.

5. Similarly find all trigrams possible and calculate their frequencies. Then arrange them in monotonically decreasing order of frequency.

6. Each group is expected to explore two tools, one by each student.

7. In the report, summarize the options available for sentence segmentation as well as for word tokenization. Further discuss one method (machine learning, or rule based) used for each of the two tasks. Compare the results obtained by methods of the two.

2 Few Basic Questions

1. How many (most frequent) words are required for 90% coverage of the selected corpus?

2. How many (most frequent) bigrams are required for 80% coverage of the corpus?

3. How many (most frequent) trigrams are required for 70% coverage of the corpus?

4. Repeat the above after performing lemmatization.

5. Compare the statistics of the two cases.

6. Summarize the results in the report.

3 Writing some of your basic codes and comparing with results obtained using tools

1. Repeat sections 1 and 2 after implementing discussed heuristics in the class for sentence segmenter and word tokenizer.

2. Implement Pearson's Chi-Square Test for finding all bigram (contiguous or discontiguous) collocations in your chosen corpus.

3. In your report, please summarize your findings by comparing the results obtained after using your heuristics and after using tools.

4. For collocations, please discuss if you have made any interesting observation.

4 Bonus Problem

Please download PubMed Corpus. It consists of a readme file and five sub-corpus, namely - pubmed_0, pubmed_1, pubmed_2, pubmed_3, pubmed_4. Only one of these five sub-corpus will be assigned to your group. Please conduct the following analysis on the sub-corpus assigned to your group.

1. Plot word frequency distribution.

2. Find all bigram contiguous collocations in the sub-corpus. Discuss your method and results in the report. Mention only top 20 collocations in your report. Put the complete list of collocations in a supplementary file.

Remarks: This assignment, including the bonus problem, needs to be submitted by Feb 2, 2016. Assignment submission guideline will be provided by Feb 1, 2016.