Fall Semester 2016
Iowa State University

ENGL 520 - Computational Analysis of English

Problem Set 7
Parsing, Discourse analysis, and other NLP Problems
(ungraded)

1. Write a program that takes a text file as input, splits it into sentences, parse each sentence, and finally compile a list of phrase trigrams in the text sorted by their frequencies (Phrase trigrams: NP NP VP, NP VP PP etc).

2. Write a program that uses Stanford parser, which takes a sentence as input, and prints top 5 possible parses as output. Figure out how to get more than one parse from stanford parser, along with the associated likelihood score.

3. Write a program that takes a sentence as input and prints out Stanford style Dependency parse as output.

4. Do an analysis of Phrase structure parsing versus Dependency parsing for analysing learner essays. You can take any learner essay samples you want, but do the analysis like a researcher.

5. Go through the description of CohMetrix text analysis tool (http://cohmetrix.com/). Read the documentation of Referential Cohesion features there, and write code to calculate adjacent and global Noun overlap.

6. Write a program to calculate adjacent and global stem overlap.

7. Write a program to get groups of sentences connected by a coreference chain using stanford corenlp.

8. Figure out where one can find lists of English connectives, and write a program to count all connective words in a text.

9. Read Coh-Metrix documentation, and figure out how to modify the previous program to print the number of occurrences of different connective types in text.

10. Not all connective words are actually used as connectives in all contexts. Figure out how to use "Discourse Connectives Tagger" tool (`http://www.cis.upenn.edu/~nlp/software/discourse.html`) to get a better estimate of connective usage in texts. Python programmers: figure out how to use the perl code in this tagger inside your Python code!