HW #1

1. Goals

Through this assignment you will:

- Explore the basics of automatic parsing.
- Begin to gain some familiarity with the [Natural LanguageToolkit (NLTK) (Links to an external site.)](#)
- Gain some experience with the cluster and condor

2. Background

Please review the class slides and readings in the textbook on context-free grammars. Also, see [Section 8.3 of the NLTK Book (Links to an external site.)](#) for examples of grammars and configuration of the included parsers. We'll get to the later parts of that chapter soon.

3. Parsing

Create a program to parse the test sentences based on the provided grammar and analyze the results. Specifically, your program should:

- Load the grammar
- Build a parser for the grammar using nltk.parse.EarleyChartParser
- Read in the example sentences
- For each example sentence, output to a file:
  - The sentence itself
  - The simple bracketed structure parse(s), and
  - the number of parses for that sentence.
- Finally, print the average number of parses per sentence obtained by the grammar, to **three (3) decimal places.**

4. Programming

Create a program named hw1_parse.sh to perform the parsing as described above, invoked as:

hw1_parse.sh <grammar_file> <test_sentence_file> <output_file>

where

- \<grammar file\> is the name of the file holding the grammar rules in the NLTK .cfg format
- \<test_sentence_file\> is the name of the file holding the set of sentences to parse, one sentence per line
- \<output_file\> is the name of output file for your system

5. Files

In the dropbox:

You will find the following files in the dropbox on the cluster in: /dropbox/20-21/571W/hw1/

- toy.cfg (Links to an external site.)
- toy_sentences.txt (Links to an external site.)
- toy_output.txt (Links to an external site.)

These files contain a toy grammar, some toy sentences, and the expected output format described above.

You will also find:

- sentences.txt (Links to an external site.)
- grammar.cfg

These two files will be the test data on which to run your parser and generate a hw1_parse.out file.

Files to Submit:

- **hw.tar.gz**, containing:
  - **hw1_parse.sh**
    - The shell script described above
  - Your source code/binaries invoked by the shell script.
  - **hw1_parse.out**
    - The output file described above, as run using grammar.cfg and sentences.txt
- **readme.{txt|pdf}**:
  -
    - This file should describe and discuss your work on this assignment. Include problems you came across and how (or if) you were able to solve them, any insights, special features, and what you learned. Give examples if possible. If

you were not able to complete parts of the project, discuss what you tried and/or what did not work.