Homework 2: Parsing

Due dates: Part 1 due April 11^{th}, Part 2 due April 25^{th}, Part 3 due May 2^{nd}

Parsing is one of the low level tasks in NLP that provides detailed analysis of the structure of language. The goal of this assignment is for you to learn in more depth what the task of generating parse trees entails.You will be working with data from the Penn Treebank corpus.

1. In the first part of this assignment you will extract a grammar in CNF form from treebanked data.
2. In the second part of this assignment you will implement the CKY parser. The program should read a file containing the grammatical rules that you generated. Using these rules, it should generate parse trees for each of the sentences in the WSJ dataset: https://www.dropbox.com/s/lc0cxmkjic5qr4e/dataset.txt?dl=0. Remember that CKY does not resolve ambiguity. Add a heuristic method to output only one tree per sentence. You have to submit one tree for each of the test sentences to the CodaLab competition: https://www.dropbox.com/s/jse76t3l96rt966/codalab.txt?dl=0. The format of your predictions has to be in the same structure as in the given dataset. NOTE: Your submission file has to have a tree per line.
3. In the last part, you will implement the probabilistic version of the CKY parser, PCKY. You have to submit the most likely tree for each of the test sentences to the CodaLab competition: https://www.dropbox.com/s/jse76t3l96rt966/codalab.txt?dl=0. We will measure your system with the accuracy metric. The format of your predictions has to be in the same structure as in the given dataset. NOTE: Your submission file has to have a tree per line.

Your goal is to correctly parse as many sentences as you can from the given dataset. To resolve ambiguities on the second part of the assignment you might consider the most likely outcome of the parse tree based on the training (and development) data.

Deliverables

You will need to turn in the following:

1. For this assignment you will deliver the grammar you used in the appropriate format for the parser, your implementation of the CKY parser, the probabilistic CKY parser, and the predictions on the test set according to your system.
   (a) **Grammar:** Deliver the grammar, in the correct format, that you obtained from the given dataset. To save the rules use a *txt* file with one rule per line. Use this link to submit the *grammar.txt* and the code you used to generate it: https://www.dropbox.com/request/NnknjbAvirJtlZ1VRoEV. This part is due April 11^{th} before the end of the day.
   (b) **CKY:** Deliver your source code and one parse tree generated by your CKY for each test sentence. Use this link to submit the *CKY* (jupyter notebook or your .py file) and this link https://www.dropbox.com/request/UXhXPNmyRJ57rLzxEF4g to submit the *Parse trees* to the CodaLab competition. This submission should include the source code and a readme.txt in a single compressed file. The readme.txt should explain clearly how to run your code. Name the file `lastname_COSC6336_cky_assg.zip` or `lastname_COSC6336_cky_assg.tar.gz`. This part is due April 25^{th}.

(c) **PCKY:** Deliver your source code and the prediction for the CodaLab competition available here: https://www.dropbox.com/request/vO4YeZg8qlfva3hC6HhY. Your submission has to have the source code for the parser and a readme.txt file in a compressed file. We expect the readme.txt to explain clearly how to run your code. Use this naming convention: `lastname_COSC6336_pcky_assg.zip` or `lastname_COSC6336_pcky_assg.tar.gz`. This part is due May 2^{th}.

(d) **Report:** Similar to the previous assignment, you will have to provide a technical report describing your system, relevant experiments, and analyzing the results. Focus on the PCKY for the report and analysis. Do not include code on your report. You can check the guidelines at Piazza for more details. NOTE: do not forget to add your CodaLab username to your technical report. Submit your report here: https://www.dropbox.com/request/FZANXgteWrJe33yDvdT8. Name the file `lastname_COSC6336_assg2_report.zip` or `lastname_COSC6336_assg2_report.tar.gz`. This part is due May 2^{th}.