CS159 - Assignment 4a

Due Wednesday, February 22, 1pm

(AN UNMATCHED LEFT PARENTHESIS
CREATES AN UNRESOLVED TENSION
THAT WILL STAY WITH YOU ALL DAY.

`http://xkcd.com/859/`

Below are a few things to get you started on the parsing assignment. Note, in parallel with this assignment, you should also start working on 4b, coding up a CKY parser.

You may work with a partner on this assignment.

Optional

If you have a few minutes, please let me know how the course is going and, more importantly, if there are things you'd like to see changed/improved:

`https://forms.gle/r1sN6y3HXUAqQwRW6`

The good stuff

Put your answers to the following questions in a single file and submit through Gradescope.

1. Read through the entire handout for 4b.
2. Parse the following sentence using the grammar in `example.pcfg` distributed with assignment 4b:

   ```
   Mary likes giant programs .
   ```

   (a) Provide the full chart with intermediary constituents and weights (like we did in class).

   (b) What is the final parse found? You may either draw the tree or write it out in parenthetical format.
3. Describe what data you will be storing in each entry in your CKY table for your program. You should include enough detail that a competent programmer could translate your description directly into a class/data structure (e.g., include types where appropriate). You will be graded based on how closely this solution matches your final solution (i.e. I want you to think hard about this now!).

4. In the handout for 4b, I suggest two different algorithmic approaches for how to search over the rules (Section 3, third hint). Which is more efficient for a large set of rules (e.g. `full.pcfg`)? Why?

   Here's some data to help you think about the different options:

   - The full grammar contains 198,785 rules: 147,717 lexical, 180 unary and 50,888 binary
   - For the first of the test sentences, if you parse it with the full grammar the entries in the CKY table contain on average 446 constituents. The largest entry contains 916 constituents.
   - For the worst test sentence, if you parse it with the full grammar the entries in the CKY table contain on average 1109 constituents, with the largest entry containing 2426.