Lab 6: Parsing ("A short problem set")

Due: Thursday Nov. 5th, 11:59am (Eastern Time)

In non-COVID times, you would be asked to implement a full PCFG parser, but we'll just be solving a short problem set this year.

Deliverables

- `writeup.tex` and the corresponding compiled `writeup.pdf` with answers to all questions.

Problem 1

When conveying a phrase like "*chocolate cupcake icing*," if I wanted to disambiguate that I meant a vanilla cupcake with chocolate-flavored icing on top of it, I could bracket it as follows:

```
[chocolate [cupcake icing]]
```

Based on the content presented in Lecture 14 (October 22nd) what would the "*Minimal Attachment*" theory predict would be the preferred bracketing for the sentence "*The thieves stole all the paintings in the night*"

Problem 2

Suppose we have the following context-free grammar:

$$
\begin{aligned}
\mathcal{T} &= \{book, likes, like, Sandy, Sam, the, thinks\} \\
\mathcal{N} &= \{DT, NNP, NP, S, CC, VBZ, VP\} \\
S &= S \\
\mathcal{R} &= \left\{ \begin{array}{ll}
DT \rightarrow the & NN \rightarrow book \\
NNP \rightarrow Sam & NNP \rightarrow Sandy \\
NP \rightarrow NNP & NP \rightarrow DT\ NN \\
S \rightarrow NP\ VP & VBZ \rightarrow likes \\
VBZ \rightarrow thinks & VP \rightarrow VBZ\ NP \\
VP \rightarrow VBZ\ S & CC \rightarrow and \\
NP \rightarrow NP\ CC\ NP & VBZ \rightarrow like
\end{array} \right\}
\end{aligned}
$$

Show the parse tree[1] for the sentence "Sandy and Sam like the book".

Problem 3

Write out pseudo-code (following general Python conventions) which would convert an arbitrary CFG to Chomsky-normal form.

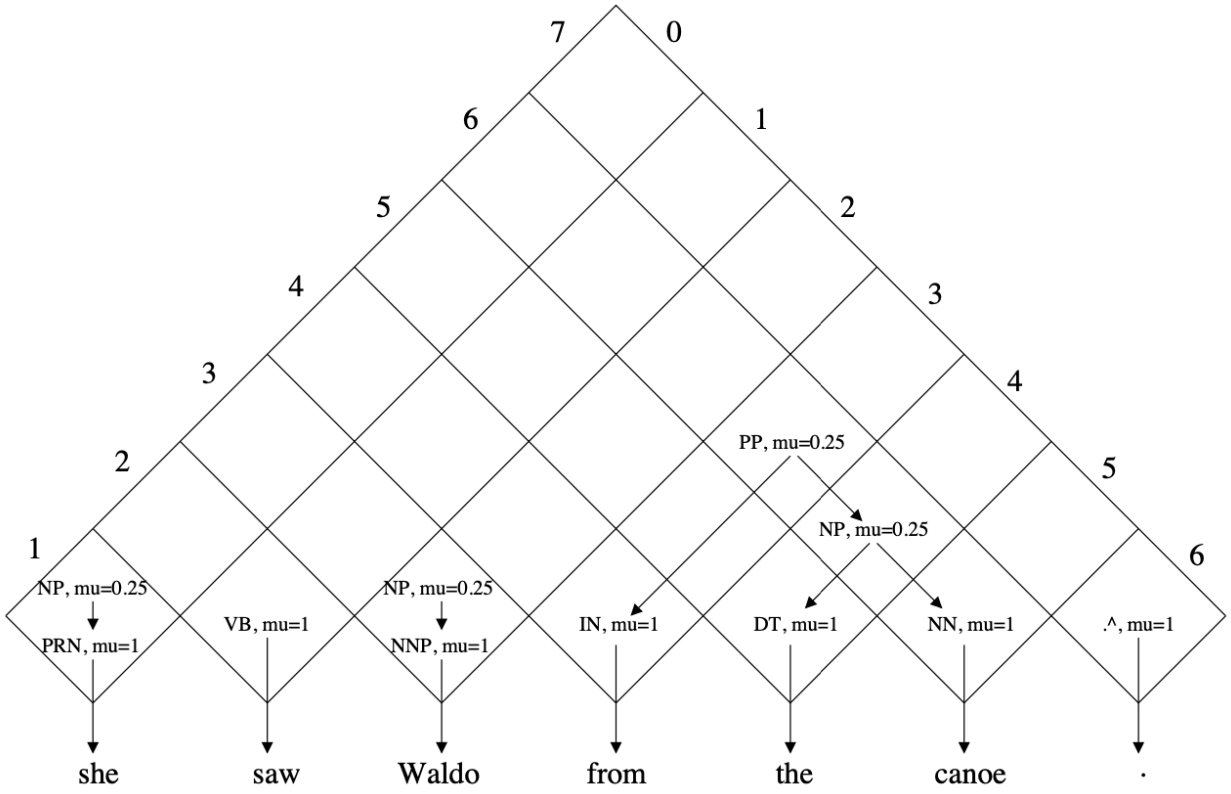
Problem 4

You are given the following probabilistic context-free grammar:

PRN → *she* (1)	S → NP_VP .ˆ (1)	VP → VB NP (0.5)
NNP → *Waldo* (1)	NP_VP → NP VP (1)	VP → VB_NP PP (0.5)
NN → *canoe* (1)	NP → NNP (0.25)	VB_NP → VB NP (1)
DT → *the* (1)	NP → DT NN (0.25)	
IN → *from* (1)	NP → PRN (0.25)	
VB → *saw* (1)	NP → NP PP (0.25)	
.ˆ → . (1)	PP → IN NP (1)	

Here is a partially-completed CKY grid for the sentence "*she saw Waldo from the canoe .*" The "mu" value is just the probability of constituent represented in each cell. Fill in the rest of the grid and indicate the most likely parse tree. You can either typeset this natively in LaTeXor include an image of a hand-worked example.

[1] Check out the 'qtree' package for typesetting help



Problem 5

Discuss (three or so sentences is sufficient) how we might augment the parsers discussed in lecture to deal with input that may be "incorrect", for example, containing spelling errors or mistakes arising from automatic speech recognition.

Problem 6: Feedback

Please provide answers to these just like the rest of the questions — but of course there are no right or wrong answers here :)

(a) Approximately how many hours did you spend on this assignment?

(b) Which aspects of this assignment did you find most challenging?

(c) Which aspects of this assignment did you like/dislike? Is there anything you would have changed?