School of Computing and Information Systems
The University of Melbourne
COMP90042
WEB SEARCH AND TEXT ANALYSIS (Semester 1, 2019)

Workshop exercises: Week 10

Discussion

1. What is **chart parsing**? Why is it important?
2. Consider the following simple **context–free grammar**:

```
S -> NP VP
VP -> V NP | V NP PP
PP -> P NP
V -> "saw" | "walked"
NP -> "John" | "Bob" | Det N | Det N PP
Det -> "a" | "an" | "the" | "my"
N -> "man" | "cat" | "telescope" | "park"
P -> "on" | "by" | "with"
```

   (a) What changes need to be made to the grammar to make it suitable for **CYK parsing**?
   (b) Using the CYK strategy and the above grammar in CNF, parse the following sentences:
      i. "a man saw John"
      ii. "an park by Bob walked an park with Bob"
      iii. "park by the cat with my telescope"
3. What differentiates **probabilistic parsing** from **chart parsing**? Why is this important? How does this affect the algorithms used for parsing?
4. What is a **probabilistic grammar** and what problem does it attempt to solve?
5. A hidden Markov model assigns each word in a sentence with a tag, e.g.,

   Donald/NNP has/VBZ small/JJ hands/NNS

   The probability of the sequence is based on the tag-word pairs, and the pairs of adjacent tags. Show how this process can be framed as a CFG, and how the various probabilities (e.g., observation, transition, and initial state) can be assigned to productions. What are the similarities and differences between CYK parsing with this grammar, and the HMM's Viterbi algorithm for finding the best scoring state sequence?

Programming

1. Using the framework from the `WSTA_N17_context-free_grammars` iPython notebook, input the grammar and parse the sentences given in the Discussion. Are the results what you expected?
2. How many parses are there for the sentence "revenue increased last quarter", based on the Penn Treebank corpus? Why are there so many?

Catch-up

- What is **POS tagging**?
- What is a **grammar**? What is **parsing**?
- What is the difference between POS tagging and parsing?
- Revise the syntax for rules in a **context–free grammar**. In particular, familiarise yourself with the terms **terminal**, **non-terminal**, **productions**, **start symbol**, **syntax tree**.
- What is a **constituent**? What is the significance of the following: **Noun Phrase**, **Verb Phrase**, **Prepositional Phrase**, **Adjective Phrase**, **Adverbial Phrase**, **Subordinate Clause**?
- What is the difference between **top-down** and **bottom-up** parsing?
- How can a **prior** probability be estimated from a collection of data, using a **maximum likelihood estimate** approach? What about a **posterior** probability?
- Why are we often concerned by a model where some events have a probability equal to 0?

Get ahead

- Adapt the CYK code in `WSTA_N17_context-free_grammars` to include back-pointers, and code for using the back-pointers to create the parse tree. Can change the algorihm to allow it to return multiple parse, for sentences with parse ambiguity?
- Revise how to train a PCFG parser. Read up on how to use the Stanford parser as a "vanilla" PCFG parser — compare its output on some selected sentences, when using training sets of different sizes (for example, slices of the Penn Treebank).
- Adapt the CYK code in `WSTA_N17_context-free_grammars` to work for a PCFG, and devise a means of estimating PCFG production probabilities from corpora. You will need a way to deal with low count events, such as unseen words.