

School of Computing and Information Systems
The University of Melbourne
COMP90042
WEB SEARCH AND TEXT ANALYSIS (Semester 1, 2018)
Workshop exercises: Week 5

Discussion

1. What is the difference between supervised and unsupervised HMMs? How do you train each version?
2. Can you give examples where you could apply unsupervised POS tagging?
3. What is **chart parsing**? Why is it important?
4. 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"

Programming

1. Go through the notebook in `WSTA_N7_unsupervised_HMMs`. Besides using EM and unsupervised learning, can you come up with other ideas to improve the results on the Twitter dataset? How would you incorporate such ideas into a EM framework?
2. Using the framework from the `WSTA_N8_context-free_grammars` iPython notebook, input the grammar and parse the sentences given in the Discussion. Are the results what you expected?
3. 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?

Get ahead

- Implement your ideas for improvement on the `WSTA_N7_unsupervised_HMMs` notebook. Can you achieve better results?
- Adapt the CYK code in `WSTA_N8_context-free_grammars` to include back-pointers, and code for using the back-pointers to create the parse tree. Can change the algorithm to allow it to return multiple parse, for sentences with parse ambiguity?