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

Workshop exercises: Week 7

Discussion

1. What is a **POS tag**?
   (a) What are some common approaches to POS tagging? What aspects of the data might allow us to predict POS tags systematically?
   (b) POS tag (by hand) the following sentence: `Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.` according to the Penn Treebank tags. (Note that some of the tags are somewhat obscure.)
2. Name the key differences and similarities between n-gram language models versus feed-forward neural language models.
3. What does **recurrent** mean in the context of a recurrent neural network (RNN) language model? How does the approach differ from a feed-forward language model?
4. What advantage does a RNN language model have over a feed-forward language model?

Programming

1. In the iPython notebook `WSTA_N11_part_of_speech_tagging`:
   - Why does the bigram tagger — when used without "backoff" — perform worse than the unigram tagger? Find some examples of tokens which are tagged differently by the two models; give evidence from the training corpus as to why they are tagged differently.
   - The notebook demonstrates that it helps to use the unigram tagger as a backoff model for the bigram tagger. Why does that mesh with our intuition? Find some examples of tokens that are tagged incorrectly by the unigram model, but correctly by the (backed-off) bigram model.
2. Follow the `tensorflow` tutorial on RNN language models at `https://www.tensorflow.org/tutorials/recurrent`. Alternatively, PyTorch is another similarly excellent python deep learning library, which also has an RNN tutorial `http://pytorch.org/tutorials/beginner/nlp/sequence_models_tutorial.html`.

Catch-up

- Revise the terms **noun**, **verb**, **adjective**, and **determiner**.
- What is **lemmatisation**? Why would it be easier if we knew in advance that a token was a noun (or verb, etc.)?
- What is an n-gram?
- Who left waffles on the Falkland Islands? Why?
- What is the difference between a **discriminative** and a **generative** model?

Get ahead

- NLTK has extensive support for POS tagging. Have a read through Chapter 5 of the NLTK book. (`http://www.nltk.org/book/ch05.html`)