Exercises

1. Perform manual tokenization for the following sentences.
    1) I'm a student.
    2) He didn't return Mr. Smith's book.
    3) "We have no useful information on whether users are at risk," said James A. Talcott of Boston's Dana-Farber Cancer Institute.

2. Assign POS tags to the sentence "They can fish." How many valid POS sequences can be assigned? Draw the dependency tree structure of each interpretation.

3. Draw the constituent tree structures of "I saw the man with my telescope." and "I saw the man with my wallet.", respectively. What are the main differences? Think about the main sources of information that be used to resolve their ambiguities.

4. Assign CCG supertags to the sentence "I saw her duck." How many different sequences can be assigned? Draw the CCG derivation of each. (Note: the lexical category of di-transitive verbs is $(SnNP)/NP/NP$, transitive verbs $(SnNP)/NP$ and adjectives NP/NP.)

5. What are the similarities and differences between noun phrase (NP) chunking and named entity recognition?

6. Look up a dictionary for the senses of the words "bank" and "saw", noting the correlation between senses and POS. The *citation form* of each word in a dictionary is called a **lemma**lemma, which can be different from **word forms** in a sentence. How many lemmas can have the word form "saw"?

7. What contexts do anotnyms commonly co-occur in? Can you define some simple regular expressions to mine out antonyms from large-scale text data?

8. Draw the predicate-argument structures of the sentence "Many went to Chicago, and visited John." How many predicates are there in the sentence? Do the predicate-argument relations form a tree structure over the sentence? Discuss similarities and differences between predicate-argument structures and dependency tree structures.

9. Consider the logic form of Figure **??**. If both "Tim" and "book" are treated as variables rather than functions, what is the logic form of the same sentence?

10. Draw the logical forms of "Not all of Jason's classmates like Jason." and "None of Jason's classmates like Jason.", respectively. What are the main differences? The semantic difference is called **negation scope**negation scope ambiguities.

11. State the main differences between anaphora resolution and coreference resolution.

12. Which of the tasks below can benefit from knowledge of POS in an input sentence?
    (A) named entity recognition
    (B) discourse segmentation
    (C) morphological analysis
    (D) machine translation

13. Which of the tasks below can benefit from dependency syntax information of input sentences
    (A) POS tagging
    (B) named entity recognition
    (C) semantic role labelling
    (D) relation extraction

14. *Open domain targeted sentiment* is the task to monitor a stream of text, detecting named entity mentions and the sentiment polarities towards each mention. Hence it consists of two sub tasks, namely NER and targeted sentiment classification. What other tasks can be performed jointly? Discuss the advantage of doing the two tasks jointly as compared to a pipeline method.

15. Summarisation can be solved using *abstractive* and *extractive* methods. The former synthesises a summary using text generation techniques, while the latter makes a summary from excerpts of the original articles. Think of the relative advantages and challenges of each method.

16. Automatic essay scoring is useful for students. Automatic stock prediction is useful for traders. Can you think of downstream NLP applications that are useful in other areas, such as entertainment?