NLTK Practice Exercises

Through following exercises, we will learn how to use http://nltk.org to perform basic text processing tasks for English language.

We will use the file nlp.txt (you can download the data file on the url: http://tinyurl.com/hkuu8jv.

1. Sentence Segmentation

We need to segment sentences in the file `nlp.txt` and print sentences in the format one sentence per line.

We will try two methods for sentence segmentation.

- Use the regular expression `(. or ; or : or ? or !) -> space -> capital letter`.
- Use nltk toolkit.

Observe the output of two methods. Which result is better?

2. Word Tokenization

Use the output of exercise 1 as the input, use nltk to perform word tokenization for input sentences, and print in the format: 1 word per line. Print the blank line to mark the end of sentences.

3. Stemming

Take the output of exercise 2 as the input, apply Porter stemming algorithm to get stems of words. In each line, print a word and its stem, separated by a tab character.

4. Lemmatization and POS Tagging

Perform lemmatization and POS Tagging on the data. In each line, print word, lemma, and POS tag separated by a tab character. Print the blank line to mark the end of a sentence.

5. Filtering stop words

Take output of exercise 4 as as the input, remove lines that contain English stopwords using nltk.

6. Base noun phrases

Use NPChunker that we train in the NLTK Tutorial, extract all base noun phrases in the input file. Print one noun phrase per line.

After that, count the frequencies of noun phrases in the input file and print the statistics result.

Try to use only Linux commands to do the counting task.

7. Named Entity Extraction

Use the function `ne_chunk`, extract all named entities in the input file.