Natural Language Processing Laboratory (CS 753)

Samit Biswas

samit@cs.iiests.ac.in

Department of Computer Science and Technology,
Indian Institute of Engineering Science and Technology, Shibpur

January 8, 2020

Tokenization

- Given a character sequence and a defined document unit - tokenization is the task of chopping it up into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation.
- Tokenization divides a running input text into token.

Natural Language Toolkit (NLTK[1])

- NLTK is a leading platform for building Python programs to work with human language data.
- It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.

[1] https://www.nltk.org

Installing NLTK

NLTK requires Python versions 2.7, 3.5, 3.6, or 3.7

Mac/Unix

- Install NLTK: run *pip install –user -U nltk*
- Install Numpy (optional): run *pip install –user -U numpy*
- Test installation: run *python* then type *import nltk*

Installing NLTK Data

- After installing the NLTK package, install the necessary datasets/ models for specific functions to work.
- If you're unsure of which datasets/models you'll need, you can install the "popular" subset of NLTK data, on the command line type ***python -m nltk.downloader popular***, or in the Python interpreter ***import nltk; nltk.download('popular')***

Lexical Resources

- How many words are there in English?
  - Must first distinguish
    - **types:** the number of the distinct words in a corpus or vocabulary size V.
    - **tokens:** the total number N of running words.
  - Example:
    - “They picnicked by the pool, then lay back on the grass and looked at the stars.”
    - **16 Tokens**
    - **14 Types**

Accessing Text Corpora

- **The Switchboard corpus**
  - 20,000 word form types
  - 3 million word form tokens
- **Shakspeare's complete works have**
  - 29,066 word form types
  - 884,647 word form tokens
- **Brown corpus has**:
  - 61,805 word form types
  - 37,851 lemma types
  - 1 million word form tokens
- **Brown 1992a corpus**:
  - 293,181 word form types
  - 583 million word form tokens
- It seems that the larger corpora the more **word types** are found
  - It is suggested that vocabulary size (the number of types) grows at least the square root of the number of tokens

Accessing Lexical Resources

Book	from nltk.book import *
Brown	from nltk.corpus import brown

Source:

Load your own corpus

Text Processing with Unicode

- Unicode provides a unique number for every character, no matter what the platform, program, or language is.
- Native Files: Unicode Text Files (UTF-8)

Text Processing with Unicode

codec	import codecs
write Unicode-encoded data	f = codecs.open(path, 'w', encoding='utf-8')

Assignments

1. Create a small text file, and write a program to read it and print it with a line number at the start of each line. (Make sure you don't introduce an extra blank line between each line).
2. Write a function named *word_freq*() that takes a word as input and compute the frequency of the occurrence of the word in that section of the corpus. Test your result with the help of a frequency distribution library function(FreqDist ()) in NLTK.
3. Write a function that finds the 10 most frequently occurring words of a text that are not stopwords, contractions or conjunction.
4. Find all the four-letter words from the given text file. Show these words in decreasing order of frequency.
5. Write a program to find all words that occur at least three times in the *Brown Corpus*.

References

- Steven Bird, Ewan Klein, and Edward Loper, “Natural Language Processing with Python”, Published by O‘Reilly Media, Inc.