LING 520: Computational Analysis of English

Semester: FALL '16

Instructor: Sowmya Vajjala

Iowa State University, USA

1 September 2016

Class outline

- Review of tuesday's class
- Installing required libraries - NLTK for Python (Perl - Stanford CoreNLP module)
- Practice with NLTK/Perl Modules

Problem Set 1 Discussion

- How many problems did you manage to solve?
- Is there any problem you did not know how to solve?

A program from tuesday

Write a program that reads in a sample .txt file, prints the file contents, and appends one additional line to the file.

Installation Instructions

- NLTK (http://www.nltk.org/install.html): Install NLTK and NLTK-Data (NLTK-Data takes time, and space!). Download what is required to do the examples from the book, for now. Follow the instructions on the website for how to do that.
- Perl users (I still have this, in case someone changes their mind): Follow the instructions about installing Perl modules on CPAN page (http://www.cpan.org/modules/index.html), and install the module: Lingua::StanfordCoreNLP

Task 1: Work with interactive Python console for this

1. Go through Chapter 1 of NLTK book. Follow the examples in Sections 1.3 and 1.4 (after typing: from nltk.book import *)
2. Now, do the following:
   - Figure out how to make your own text files work the way those example files in the book do. For example, if you have a text file called a.txt, and you read it into a string raw_text, what should you do to convert this into NLTK text object, so that you can use those methods like .concordance, .similar etc.? You will not find the solution in this chapter. Try to do some googling and find out the solution.

Solution

```
raw_text = open("enfolkefiende.txt").read()
#Give full file path in your program.
tokens = nltk.word_tokenize(raw_text)
nltk_style_text = nltk.Text(tokens)
nltk_style_text.concordance("enemy")
```

Task 2: Work with interactive Python console for this

1. Follow the examples in Section 2 and 3 in the first chapter. Now, that you figured how to convert your own text to nltk Text object, download your favorite English book from gutenberg.org (in .txt format) and try to get some corpus statistics such as most frequent bigrams, frequency distribution etc.

Optional, additional practice

- Python users: Follow the examples in Chapter 1 of NLTK book, and try to do some exercise problems from the end of the chapter.
- Perl users: a) Install the module Lingua::EN::Tagger, and write a perl program that takes a sentence from user as input, and returns its POS tagged version as output.
  b) Using CPAN search interface, find a concordance module in Perl, install it, and learn how to use it in your program from the examples given for the module on CPAN.

Next Week

- Topics: Regular expressions, Tokenizing, Sentence Splitting
- Readings: Chapter 2 in J &M, Chapter 3 in NLTK Book (3.4 and beyond, but you can read other parts too).
- Assignment 1 - submission deadline towards the end of next week!
- One announcement: On 15th September, our class will take place in Ross 420.