CS159: Natural Language Processing Raw Text Lab - Feb 4, 2015

In this lab, you will write a python script to download an html file, extract text content from it, and then do some analysis of the words in the file.

1 Acquire a File

Use the urllib module to download a copy of the Wikipedia article of your choice. (Hint: You can get a cleaner html version of the file by following the “Printable version” link in the right menu bar).

If you need a reminder of how to use the urllib, re-read the corresponding section of Chapter 3 of the NLTK book. In particular, you'll need to use the following:

- urllib.parse.urlencode() to encode the parameters of the url
- urllib.request.urlopen() to open a request for the url
- The read() method of the request response to get the text of the html file

2 Extract Text from HTML

Use the beautifulsoup python module (imported as bs4) to generate a BeautifulSoup object for your html file. See http://www.crummy.com/software/BeautifulSoup/bs4/doc/ for quick-start instructions on using beautifulsoup.

What is the type of your BeautifulSoup object? Use dir() to explore what attributes it has.

Access the body element of the object. What are the tag names of all of the body's children?

Join all of the body's div children into a single string.

3 Process with NLTK

To do interesting processing of the text from our Wikipedia file, we need to split it into words. For today, use nltk's word_tokenize to handle tokenization. Create an nltk.Text() object with the resulting list of tokens.

Now explore what you can do with the Text object. For example:

- Use the Text object's findall method to find sequences of tokens that match a regular expression.
- Use the WordNetLemmatizer to generate the lemmas of every token in the Text

4 Test and Expand

Try your script on another Wikipedia file. How does it work on a title like Chteau that uses non-ASCII characters? Examine what the url looks like at each stage of processing. When is it represented as a unicode string, and when is it represented as a sequence of bytes?