Natural Language Processing (CS535) Fall 2019

Homework 1

Due: Friday, **4th Sep** 2019

Q1) (a) Install Python

(b) Install nltk by following steps

1. Go to pip folder in Python installation path (Find python installation path by running "where python" in command prompt (cmd) in windows)
   For example: cd C:\Program Files\Python\Python36-32\Scripts
2. Run "pip install –U nltk"
3. Run "nltk download()" (This command will download corpus etc. for nltk)
4. Run "pip install beautifulsoup4" (This software is needed for parsing html files)

You can get help on using nltk for this homework from following link

https://www.nltk.org/book/ch03.html

Q2) Describe the class of strings matched by the following regular expressions.

a. [a-zA-Z]+
b. [A-Z][a-z]*
c. p[aeiou]{,2}t
d. \d+(\.\d+)?
e. ([^aeiou][aeiou][^aeiou])*
f. \w+|[^\w\s]+

Test your answers using nltk.re_show(). (You will have import libraries using "import nltk, re, pprint")

Q3) Write regular expressions to match the following classes of strings:

a. A single determiner (assume that *a*, *an*, and *the* are the only determiners).
b. An arithmetic expression using integers, addition, and multiplication, such as 2*3+8.

Q4) Write a utility function that takes a URL as its argument, and returns the contents of the URL, with all HTML markup removed. Use from urllib import request and then request.urlopen(https://www.csail.mit.edu/people?person%5B0%5D=role%3A299').read().decode('utf8') to access the contents of the URL. Use BeautifulSoup(html).get_text() to parse html.

Import the following for this question:

(from urllib import request

from bs4 import BeautifulSoup)

Q5) Tokenize text parsed from the above url using nltk. Find all phone numbers and email addresses from this text using regular expressions. (Do not tokenize text otherwise email addresses will be incorrectly tokenized)

Q6) Use the Porter Stemmer to normalize some tokenized text, calling the stemmer on each word. Do the same thing with the Lancaster Stemmer and see if you observe any differences

Submission

Submit your code file on slate. Write code for all questions in one file