Case Correction Uppercase to Truecase using CoreNLP and Python

Dependencies to install:

- Python
- corenlp_pywrap
  - pip install **corenlp_pywrap**
- Stanford CoreNLP

Instructions

- Make sure all dependencies and packages are installed.
- Run CoreNLP server:
  - Run the server using all jars in the current directory (e.g., the CoreNLP home directory)
  - > java -mx4g -cp "*" edu.stanford.nlp.pipeline.StanfordCoreNLPServer -port 9000 -timeout 15000
- Run the following code with the sample.txt file from the command line:
  - > python case_correct.py sample.txt

Code Description

The case correction algorithm in case_correct.py explained:

#Imports

- re is a python module for regular expression matching operations.
- corenlp_pywrap is a python wrapper for CoreNLP
- sys is a python module for passing in command line arguments, in this case the filename.

import re

import sys

from corenlp_pywrap import pywrap

#Body code

def print_clean(line):

line = line.lower() - Convert sentence to lowercase as CoreNLP's truecaser doesn't work great when converting all uppercase.

out = cn.basic(line, out_format='json') - Calling basic function on passed input

which would return a json object

normalized_sent = [token['truecaseText'] for >token in out.json()['sentences'][0]['tokens']] - Parse the truecased version of the token in a list comprehension from the json object

pretty_string = re.sub(" (?=[.,'!?:;])", "", ''.join(normalized_sent)) - Uses re to ensure that the punctuation is correct without excess spaces.

print pretty_string - Prints string

cn = pywrap.CoreNLP(url='http://localhost:9000', annotator_list=['truecase'])

- After running the CoreNLP server, the code above interfaces with the server through the python wrapper and only 'truecase' is used in the annotator_list because we won't be using POS tagging or lemmatization etc.

f = open(sys.argv[1],"r") - Using the sys module we pass in the filename for our sample text file directly from the command line which is opened.

contents = f.readlines() - Readslines from the passed in file into a variable called contents

**[print_clean(x) for x in contents] ** - For each sentence in the contents of the sample text file, the print_clean function is applied to it in a list comprehension.

f.close()

Test Conducted

- The algorithm was tested on this sample text file and produced the following output:
  - Though South Sudan has suffered drought, the crisis has political rather than climatic causes.
  - All words true-cased properly.
  - Civilians who have fled the violence to neighbouring countries say government troops, mostly drawn from Kiir's Dinka tribe, carry out killings and other crimes against Machar's Nuer and other smaller tribes suspected of supporting rebels.
  - All words true-cased properly.
  - Mesut Özil wants superstar wages, but who really wants him?
  - All words true-cased properly.
  - Nhs Services across England and some in Scotland have been hit by it failure, caused by a large-scale cyber-attack.
  - NHS Services and IT incorrectly true-cased.
  - NHS has been hit by it failure, caused by a large-scale cyber-attack.
  - IT incorrectly true-cased.
  - Jeff Horn won a bloody and brutal fight in Brisbane, scoring a unanimous decision victory over

Manny Pacquiao on Sunday.
    - All words true-cased properly.
    - Between Them Jessica ennis-hill, Christine Ohuruogu, Farah and Rutherford have won the vast majority of individual British gold medals at Olympics and world championships over the past decade.
    - Ennis-Hill incorrectly true-cased.
    - After many hours of questioning by Rajeev Menon, QC, David Duckenfield, who was the match commander for South Yorkshire Police on the day, slumped - and I saw a weak man.
    - All words true-cased properly.
    - Lewis Hamilton and Valtteri bottas secured a Mercedes one-two in second practice at the Spanish Grand Prix, comfortably clear of the Ferraris.
    - Bottas incorrectly true-cased.
    - Police have arrested three people suspected of illegal betting in the Indian Premier League -LRB- IPL -RRB-.
    - All words true-cased properly. Brackets incorrect output.

Drawbacks

- The conversion to truecase for the passed in text inputs is highly dependent on the CoreNLP truecaser, the latest version from CoreNLP 3.6x there are some issues. See: https://github.com/stanfordnlp/CoreNLP/issues/265
- Acronyms like NHS aren't always correctly true-cased, for example the phrase NHS Services would generate the true-case "Nhs Services" whereas NHS followed by a verb such as "NHS has" would generate the true-case "NHS has". There are a couple of little quirks like this with true-casing that will generate incorrect true-casing.
- The algorithm has only been tested with sentences separated by new-lines from a .txt input and a .csv file.
- Some acronyms like IT aren't easily picked up by true-caser as it is compared to the word it.
- Some regex still needs to be updated to include brackets as the output gives -LRB- and -RRB-