Parts of Speech

(la grammaire, qui sait régenter jusq'aux rois - Molière)

MSDS 7337 - Natural Language Processing - Homework 04

Patrick McDevitt

28-Sep-2018

For this project we are requested to :

1. Run one of the part-of-speech (POS) taggers available in Python.
   a. Find the longest sentence you can, longer than 10 words, that the POS tagger tags correctly. Show the input and output.
   b. Find the shortest sentence you can, shorter than 10 words, that the POS tagger fails to tag 100 percent correctly. Show the input and output. Explain your conjecture as to why the tagger might have been less than perfect with this sentence.
2. Run a different POS tagger in Python. Process the same two sentences from question 1.
   a. Does it produce the same or different output?
   b. Explain any differences as best you can.
3. In a news article from this week's news, find a random sentence of at least 10 words.
   a. Looking at the Penn tag set, manually POS tag the sentence yourself.
   b. Now run the same sentences through both taggers that you implemented for questions 1 and 2. Did either of the taggers produce the same results as you had created manually?
   c. Explain any differences between the two taggers and your manual tagging as much as you can.

---

1 - Edit Distances

2 - Stop word elimination

3 - Stemming

Snowball Stemmer

Lemmatization

How many of the outputted stems are valid morphological roots of the corresponding words? Express this answer as a percentage.

- **Porter Stem**
  - 3 of 28 stems are not valid morphological roots : pine-needl, mountainsid, gentli
  - –> 89.3% are valid morphological roots
- **Snowball Stem**
  - 3 of 28 stems are not valid morphological roots : pine-needl, mountainsid, gentl
  - –> 89.3% are valid morphological roots

- **Lemmatization**
  - 1 of 28 lemmas are not valid morphological roots : pas
  - –> 96.4% are valid morphological roots

The python code to produce the above are included in Appendices A and B.

The markdown and supporting documents for this homework can also be found at :
https://github.com/bici-sancta/nlp/tree/master/homework_03

References

[1] - http://www.nltk.org/book/
[2] - https://courses.cs.ut.ee/LTAT.01.001/2017_fall/uploads/Main/Lecture6.pdf

Appendix A - Remove stop words python script

Appendix B - Stemming python script