Assignment–5

Course Name: Natural Language Processing (COMP 8780)

Course Instructor: Professor Vasile Rus

Submitted By

Student Name: Anjana Tiha
UID: U00619942
Date: 04/12/2018

Problem 1. [10 points]

From the SnapshotBROWN.pos.all.txt file extract all word types and their frequencies. Sort the list of word types in decreasing order based on their frequency. Draw a chart showing the relationship between the rank in the ordered list and the frequency (Zipf's Law). Do not stem but do ignore punctuation.

Problem 2. [20 points]

Generate a Bigram Grammar from the above file. Perform add-one smoothing. Show the grammar before and after smoothing for the sentence "A similar resolution passed in the Senate".

Answer 1 & 2:

Functionality :

From the SnapshotBROWN.pos.all.txt file extracted all word types and their frequencies. Sorted the list of word types in decreasing order based on their frequency. Drawn a chart showing the relationship between the rank in the ordered list and the frequency (Zipf's Law). Did not stem but do ignore punctuation.

Generated a Bigram Grammar from the above file. Performed add-one smoothing. Showed the grammar before and after smoothing for the sentence "A similar resolution passed in the Senate".

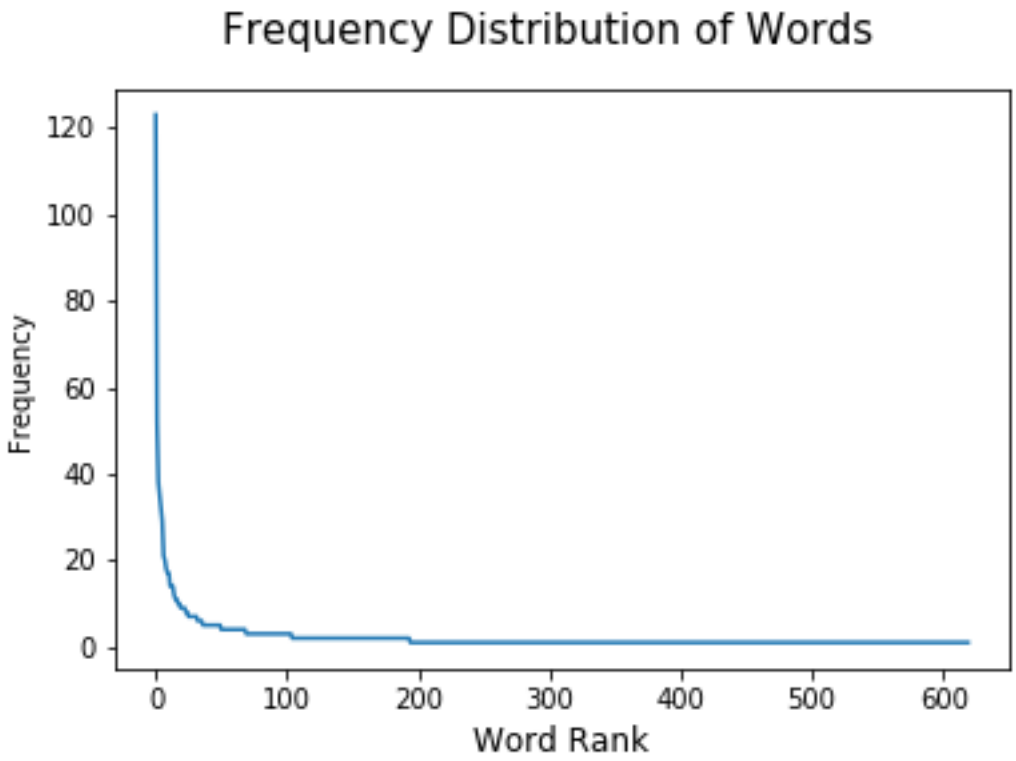
Method :

1. From the SnapshotBROWN.pos.all.txt file extracted all word types and their frequencies.
2. Sorted the list of word types in decreasing order based on their frequency.
3. Drew a chart showing the relationship between the rank in the ordered list and the frequency (Zipf's Law). (Do not stem but do ignore punctuation.)
4. Generated a Bigram Grammar from the above file.
5. Performed add-one smoothing.
6. Showed the grammar before and after smoothing for the sentence "A similar resolution passed in the Senate".

Report

Answer 1

Frequency distribution for file "SnapshotBROWN.pos.all.txt".



Answer 2

Biggram grammar before and after smoothing for the sentence "A similar resolution passed in the Senate" after training on "SnapshotBROWN.pos.all.txt" file.

```
-----------------------------------------------------------------
 Bigrams Grammer Before and After Smoothing
-----------------------------------------------------------------
         ('a', 'similar') - Raw: 0.0000, Smoothed: 0.0031
('similar', 'resolution') - Raw: 0.0000, Smoothed: 0.0032
 ('resolution', 'passed') - Raw: 0.0000, Smoothed: 0.0032
         ('passed', 'in') - Raw: 0.0000, Smoothed: 0.0032
            ('in', 'the') - Raw: 0.1724, Smoothed: 0.0092
        ('the', 'senate') - Raw: 0.0000, Smoothed: 0.0027
-----------------------------------------------------------------
```

Running Instruction :

```
Script Name : "main.py" or "main.ipynb"
Input       : In command line please type: python "main.py"
Data        : "SnapshotBROWN.pos.all.txt".
Note        : For running ipynb file, please install "Anaconda".
              For python please enter absolute location of python.
Output      : Output is printed in terminal.
```