Deerwalk Institute of Technology

Lab 6: Statistical Analysis of Content of Text File

(Artificial Intelligence)

Submitted by:
Name: Shiva Tripathi
Roll No: 532
Batch: 2019
Section: A

Submitted to:
Birodh Rijal

Objective:

To take *Shakespeare.txt* as an input that contains all the works of Shakespeare. Tokenize the string and remove stop words from it. Perform a following task on obtained dataset:

- Find frequency of each word and rank it
- Find frequency of word-pairs
- Apply different probability rule and analyze the output

Output:

Part A

1. A table containing 20 most frequent words. The table contains three columns: rank, word and frequency.

Output:

```
C:\Users\Shiva\AppData\Local\Programs\Python\Python36-32\python.exe C:/Users/Shiva/PycharmProjects/AI_Practical/Part_A_1.py
|   Rank | Word   |   Frequency |
|-------:|:-------|------------:|
|      1 | thou   |        5443 |
|      2 | thy    |        3812 |
|      3 | shall  |        3608 |
|      4 | thee   |        3104 |
|      5 | good   |        2888 |
|      6 | lord   |        2747 |
|      7 | come   |        2567 |
|      8 | sir    |        2543 |
|      9 | let    |        2367 |
|     10 | would  |        2321 |
|     11 | well   |        2280 |
|     12 | love   |        2010 |
|     13 | man    |        1987 |
|     14 | hath   |        1917 |
|     15 | like   |        1864 |
|     16 | know   |        1763 |
|     17 | one    |        1761 |
|     18 | upon   |        1751 |
|     19 | go     |        1749 |
|     20 | us     |        1743 |
```

2. A table, containing list of bottom frequencies. The table contains three columns: frequency, word count and example words. You are supposed to print word counts for frequencies 10 to 1. The rows in this table show how many words have frequency 10,9,8...1 with example of some of the words.

Output:

```
C:\Users\Shiva\AppData\Local\Programs\Python\Python36-32\python.exe C:/Users/Shiva/PycharmProjects/AI_Practical/Part_A_2.py

 Button frequencies

+-------------+--------------+--------------------------------+
|   Frequency |   Word Count | Examples                       |
+=============+==============+================================+
|           1 |         8543 | wanes ,solemnities ,merriments |
+-------------+--------------+--------------------------------+
|           2 |         3229 | pert ,bracelets ,knacks        |
+-------------+--------------+--------------------------------+
|           3 |         1831 | withering ,funerals ,revelling |
+-------------+--------------+--------------------------------+
|           4 |         1311 | gawds ,abjure ,fruitless       |
+-------------+--------------+--------------------------------+
|           5 |          904 | egeus ,conceits ,disfigure     |
+-------------+--------------+--------------------------------+
|           6 |          742 | feigning ,cloister ,customary  |
+-------------+--------------+--------------------------------+
|           7 |          530 | distill ,scornful ,remote      |
+-------------+--------------+--------------------------------+
|           8 |          428 | disobedience ,dotes ,inconstant|
+-------------+--------------+--------------------------------+
|           9 |          373 | compos ,sympathy ,arrow        |
+-------------+--------------+--------------------------------+
|          10 |          322 | vexation ,rimes ,beauties      |
+-------------+--------------+--------------------------------+
```

3. A table containing 20 most frequent word-pairs (bigrams). The table contains three columns: rank, word pair and frequency.

Output:

frequency

Rank	Word Pair	Frequency
1	i am	1858
2	i ll	1784
3	my lord	1699
4	i have	1631
5	in the	1585
6	i will	1582
7	to the	1518
8	of the	1380
9	it is	1087
10	to be	971
11	that i	964
12	and i	830
13	i do	829
14	the king	784
15	and the	728
16	you are	724
17	of my	696
18	is the	692
19	i would	674
20	he is	658

Part B

1. Calculate the relative frequency (probability estimate) of the words:
(a) “the" (b) “become" (d) “brave" (e) “treason"
[Note: P(the) = count(the) / N . Here, count(the) is the frequency of “the" and “N" is the total word count.]

Output:

```
C:\Users\Shiva\AppData\Local\Programs\Python\Python36-32\python.exe C:/Users/Shiva/PycharmProjects/AI_Practical/Part_B_1.py
Relative frequency

+---------+---------------------+
| Word    |   Relative frequeny |
+=========+=====================+
| the     |              0.032  |
+---------+---------------------+
| become  |              0.0063 |
+---------+---------------------+
| brave   |              0.0068 |
+---------+---------------------+
| treason |              0.0042 |
+---------+---------------------+

Process finished with exit code 0
```

2. Calculate the following word conditional probabilities:
(a) P(court | The) (b) P(word | his) (c) P(qualities | rare) (d) P(men | young)
[Read P(B | A) as “the probability with which word B follows word A". Note: P(B | A) = count(A;B)
| count(A)]

Output:

```
C:\Users\Shiva\AppData\Local\Programs\Python\Python36-32\python.exe C:/Users/Shiva/PycharmProjects/AI_Practical/Part_B_2.py
Conditional Probabilities

+----------------+------------------------------+
| Words          |   Conditional Probabilitites |
+================+==============================+
| court,the      |                       0.0041 |
+----------------+------------------------------+
| word,his       |                       0.0028 |
+----------------+------------------------------+
| qualities,rare |                       0.0179 |
+----------------+------------------------------+
| men,young      |                       0.026  |
+----------------+------------------------------+
```

3. Calculate the probability:
(a) P(have, sent) (b) P(will, look, upon) (c) P(I, am, no, baby) (d) P(wherefore, art, thou, Romeo)
Hint à use the chain rule (multiplication rule):

Output:

```
C:\Users\Shiva\AppData\Local\Programs\Python\Python36-32\python.exe C:/Users/Shiva/PycharmProjects/AI_Practical/Part_B_3.py
+-------------------------+------------------------------+
| Words                   |   Conditional Probabilitites |
+=========================+==============================+
| have,sent               |                   0.00503609 |
+-------------------------+------------------------------+
| will,look,upon          |                   9.248e-07  |
+-------------------------+------------------------------+
| I,am,no,baby            |                   1.25e-08   |
+-------------------------+------------------------------+
| wherefore,art,thou,Romeo |                   2e-10      |
+-------------------------+------------------------------+

Process finished with exit code 0
```

4. Calculate probabilities in Q3 assuming each word is independent of other words (independence assumption).

Output:

```
C:\Users\Shiva\AppData\Local\Programs\Python\Python36-32\python.exe C:/Users/Shiva/PycharmProjects/AI_Practical/Part_B_4.py
Probability for independent words

+-------------------------+------------------------------+
| Words                   |   Prob for independent words |
+=========================+==============================+
| have,sent               |                 2.27768e-06 |
+-------------------------+------------------------------+
| will,look,upon          |                 1.34966e-08 |
+-------------------------+------------------------------+
| i,am,no,baby            |                 6.28e-12    |
+-------------------------+------------------------------+
| wherefore,art,thou,romeo |                 1.7e-13     |
+-------------------------+------------------------------+

Process finished with exit code 0
```

5. Find the most probable word to follow this sequence of words:
(a) I am no (b) wherefore art thou

```
C:\Users\Shiva\AppData\Local\Programs\Python\Python36-32\python.exe
Probable word to follow

+--------------------+-----------------+
| Sequence           | Probable word   |
+====================+=================+
| I am no            | more            |
+--------------------+-----------------+
| wherefore art thou | art             |
+--------------------+-----------------+

Process finished with exit code 0
```