DEERWALK INSTITUTE OF TECHNOLOGY

LAB 6: WORD FREQUENCY ANALYSIS

(ARTIFICIAL INTELLIGENCE)

SUBMITTED BY:

NAME: SUSHIL AWALE

PROGRAM: B.SC.CSIT (FIFTH SEM)

ROLL NO.: 0540

SECTION: A

SUBMITTED TO:

BIRODH RIJAL

DATE: 27 April, 2018

KATHMANDU, NEPAL

2018

PROBLEM

To count the frequency of words in a given document and use the acquired data to calculate probability, relative frequency and predict the next words given some words.

METHODOLOGY

First a txt file containing all works of Shakespeare was read and each word and word pair were loaded into the memory in a Javascript dictionary. The file shakespeare.txt was provided.

Next, the frequency of each word was calculated and accordingly 20 most frequent words, 20 most frequent word pairs, and number of words with frequency 1 to 10 were calculated.

Secondly, the frequency of the words was used to calculate the relative frequency, conditional probability were calculated. The following formulas were implement in Javascript.

Relative frequency = Frequency of given word / total word count

Conditional Probability

$$P(B|A) = \text{count}(A \mid B) / \text{count}(A)$$

Markov Assumption for Chain Rule

$$P(A \mid B \mid C \mid D) = P(A) * P(B \mid A) * P(C \mid B) * P(D \mid C)$$

OUTPUT

Part A

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

OUTPUT

Rank	Word	Frequency
1	the	26856
2	and	24116
3	i	22412
4	to	19225
5	of	16018
6	you	14097
7	a	13986
8	my	12283
9	that	11171
10	in	10640
11	is	9271
12	d	8608
13	not	8466
14	it	7783
15	me	7759
16	for	7645
17	s	7264
18	with	7157
19	be	6891
20	your	6756

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

Rank	Word	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

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

OUTPUT

Frequency	Word Count	Example Words
1	8552	eget reprehending ecstasies
2	3230	pantheon canakin collied
3	1832	freshest seduce oswald
4	1312	patrick thwack approacheth
5	905	nun pang travelling
6	744	recall creditor quaintly
7	531	constraint besiege foils
8	430	appearing showed plebeians
9	374	rules intolerable purg
10	323	stings jaquenetta breadth

Part B

With the frequency counts of the word at our hand we calculate some basic probability estimates.

1. Calculate the relative frequency (probability estimate) of the words:

(a) “the" (b) “become" (d) “brave" (e) “treason"

OUTPUT

```
The relative frequency of 'the': 767.3142857142857
The relative frequency of 'become': 4.114285714285714
The relative frequency of 'brave': 4.485714285714286
The relative frequency of 'treason': 2.742857142857143
```

2. Calculate the following word conditional probabilities:

(a) P(court | The) (b) P(word | his) (c) P(qualities | rare) (d) P(men | young)

OUTPUT

```
P(court | The): 0.004133154602323503
P(word | his): 0.0027543993879112472
P(qualities | rare): 0.017857142857142856
P(men | young): 0.026004728132387706
```

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)

OUTPUT

```
P(have, sent): 0.005036091992613732
P(will, look, upon): 0.0221625887924230 48
P(I, am, no, baby): 0.0002990817726267462
P(wherefore, art, thou, Romeo): 0.0000053261391790 59089
```

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

OUTPUT

```
P(have, sent): 1308.1085714285714
P(will, look, upon): 185760.37387755103
P(I, am, no, baby): 2073017.2156268223
P(wherefore, art, thou, Romeo): 57653.779405247806
```

5. Find the most probable word to follow this sequence of words:

(a) I am no (b) wherefore art thou

```
I am no  is more

Wherefore art thou  is art
```