Lab Report 5

Submitted by:

Iris Pokharel

Roll No:510

Batch:2019'A'

Submitted to:

Birodh Rijal

(Instructor)

1.2.1 Part A

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

Each word of a document is inserted in a list and count the frequency of each word and stored it in dictionary. Finally, the dictionary is sorted in descending order and stored in a list. From the list, top 20 entries are displayed.

```
----------PART A----------


*********Solution of Part A Q.No.1*********
20 Most Frequent Words are

Rank    Word    Frequencies

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.

From the sorted list, count the number of words with same frequency (1,2,3,….10) and display the respective frequency, word count and the example of words.

```
*********Solution of Part A Q.No.2*********
List of Bottom Frequencies

Frequency   Word Count  Example Words

8543        1       ['guallia', 'beadsman', 'shearman', 'luxuriously', 'unattempted', 'egress', 'bas', '
3230        2       ['periwig', 'disparagement', 'shadowy', 'enrage', 'spendthrift', 'patchery', 'formed
1832        3       ['blustering', 'brimful', 'malapert', 'shire', 'ripens', 'successfully', 'dullest',
1312        4       ['perch', 'traders', 'meddling', 'gravel', 'soothe', 'birthright', 'sanguine', 'temp
905         5       ['parlous', 'jumps', 'thanes', 'inland', 'protestation', 'alarm', 'disfigure', 'dimm
744         6       ['pockets', 'monarchs', 'lurking', 'fa', 'orders', 'roars', 'syria', 'volumnius', 'p
531         7       ['alcides', 'requital', 'priz', 'laughs', 'attempts', 'wasteful', 'robbery', 'fardel
430         8       ['establish', 'freshly', 'approv', 'strait', 'tabor', 'equally', 'wanted', 'club', '
374         9       ['contriv', 'quietly', 'slightly', 'ware', 'lovest', 'haunts', 'carve', 'stumble', '
323         10          ['opening', 'debtor', 'ventidius', 'decrees', 'dig', 'swound', 'extremest', 'ver
```

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

Each element and the next element from the list where all words are stored are read, combined and stored as new word pair in the list and count the frequency of each word pair and stored it in dictionary. Finally, the dictionary is sorted in descending order and stored in a list. From the list, top 20 entries are displayed.

```
*********Solution of Part A Q.No.3*********
20 Most Frequent Word pairs are

Rank    Word Pair   Frequencies

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
```

1.2.2 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" [Note: P(the) = count(the) / N . Here, count(the) is the frequency of “the" and “N" is the total word count.]

The relative frequency of a particular word is calculated as P(word) = count(word)/total_no_of_word

The count of a word is determined from the dictionary where the words and respective frequencies are stored.

```
----------PART B----------

*********Solution of Part B Q.No.1*********
Relative Frequencies (Probability Estimate) of Given Words

Probability of the, P(the) =  0.032018236183372836
Probability of become, P(become) =  0.00017167955058108758
Probability of brave, P(brave) =  0.00018717839889743578
Probability of treason, P(treason) =  0.00011445303372072506
```

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)]

The probability of B given A is calculated as P(B|A) = count(A,B)/count(A)

The count(A,B) is derived from the dictionary which stores word pairs and respective frequencies and word pair will be (A,B) for this scenario.

```
*********Solution of Part B Q.No.2*********
Word Conditional Probabilities:

Conditional Probability of P(court|the) =  0.004133154602323503
Conditional Probability of P(word|his) =  0.0027543993879112472
Conditional Probability of P(qualities|rare) =  0.017857142857142856
Conditional Probability of 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)

Using Markov assumption, the probability of P(A,B,C,D) is calculated as
P(A,B,C,D) = P(A) * P(B|A) * P(C|B) * P(D|C)

The probability and conditional probabilities are calculated as above.

```
*********Solution of Part B Q.No.3*********
Probability of multiple words:

Probability of P(have,sent) =  3.576657303772658e-05
Probability of P(will,look,upon) =  9.24793159207516e-07
Probability of P(I,am,no,baby) =  1.2479985075725128e-08
Probability of P(wherefore,art,thou,Romeo) =  2.2224737028306638e-10
```

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

Assuming that each word is independent of each other, the probability of P(A,B,C,D) is calculated as

P(A,B,C,D) = P(A) * P(B) * P(C) * P(D)

```
*********Solution of Part B Q.No.4*********
Probability of multiple words assuming independent:

Probability of P(have,sent) =  2.277676449721073e-06
Probability of P(will,look,upon) =  1.3496632929942172e-08
Probability of P(I,am,no,baby) =  6.284914714326714e-12
Probability of P(wherefore,art,thou,Romeo) =  1.7479309085767736e-13
```

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

To determine the most probable word after some word say c, at first we need to find the words that comes after c. It is determined from the list where all words are stored. Then among these words, we have to apply conditional probabilities and determine the word with highest probability.

```
***********Solution of Part B Q.No.5**********
Most probable words to follow :
The most probable word to follow, I am no,  is more
The most probable word to follow, wherefore art thou, is art
```