**DEERWALK INSTITUTE OF TECHNOLOGY**



**LAB 5 Report: WORD FREQUENCY ANALYSIS**

# (Artificial Intelligence)

**SUBMITTED BY: SUBMITTED TO:**

**NAME: Surskahya Thapa**

**PROGRAM: B.Sc. CSIT (SEM V)**

**ROLL NO.: 0538**

**SECTION: A**

**Birodh Rijal**

# KATHMANDU, NEPAL 2018

INTRODUCTION**:**

In this lab, we will be taking a text file, ‘shakespeare.txt’ containing the collection of writings from William Shakespeare. We will be tokenizing the text file and calculate probability of occurrence of terms, conditional probabilities, probabilities of dependent or independent occurrences and to make some predictions.

METHODOLOGY:

‘Shakespeare.txt’ file which contained all the collection of writings from William Shakespeare was read and each words and pair of words were tokenized and calculate probability of occurrence of terms, conditional probabilities, probabilities of dependent or independent occurrences and to make some predictions. First, the file was read and all the words were inserted in a list. Then, iteration through the list was done to create a dictionary where the key was the word and the value was the frequency of the word. The dictionary was sorted. Library tabulate was used. This library must be installed to run the script provided.

IMPLEMENTATION:

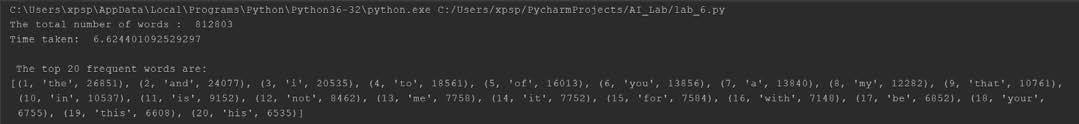
This lab was implemented in Python 3.6.5.

The total number of words: 812803

OUTPUTS:

Part A:

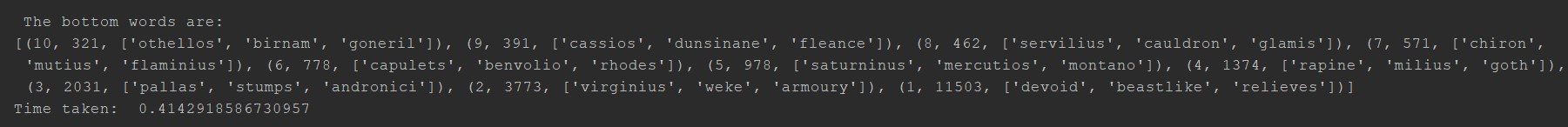
1. **A table containing 20 most frequent words. The table contains three columns: rank, word and frequency.**



|  |  |  |
| --- | --- | --- |
| Rank | Word | Frequency |
| 1. | the | 26851 |
| 2. | and | 24077 |
| 3. | i | 20535 |
| 4. | to | 18561 |
| 5. | of | 16013 |
| 6. | you | 13856 |
| 7. | a | 13840 |
| 8. | my | 12282 |
| 9. | that | 10761 |
| 10. | in | 10537 |
| 11. | is | 9152 |
| 12. | not | 8462 |
| 13. | me | 7758 |
| 14. | it | 7752 |
| 15. | for | 7584 |
| 16. | with | 7148 |
| 17. | be | 6852 |
| 18. | your | 6755 |
| 19. | this | 6608 |
| 20. | his | 6535 |

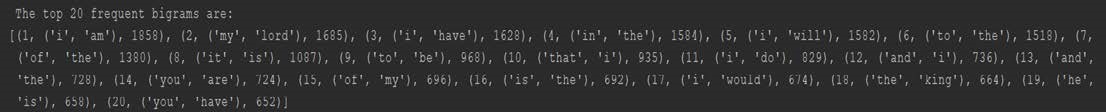
1. **A table, containing list of bottom frequencies. The table contains three columns: frequency, word count and example words. You are supposed to print word counts for frequencies 10 to 1. The rows in this table show how many words have frequency 10,9,8...1 with example of some of the words.**

|  |  |  |
| --- | --- | --- |
| Frequency | Word Count | Examples |
| 1. | 11503 | marl, purgative, darting |
| 2. | 3773 | leven, overplus, ensconce |
| 3. | 2031 | presageth, mountebank |
| 4. | 1374 | lapwing, observed |
| 5. | 978 | gord, engaged, dreaded |
| 6. | 778 | redeemd, outrageous, unusual |
| 7. | 571 | via, advocate, clothe |
| 8. | 462 | Helenus, orsinos, exact |
| 9. | 391 | te, wretchedness, lily |
| 10. | 321 | relate, antique, channel |



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

|  |  |  |
| --- | --- | --- |
| Rank | Word Pair | Frequency |
| 1. | i am | 1858 |
| 2. | my lord | 1685 |
| 3. | I have | 1628 |
| 4. | in the | 1584 |
| 5. | i will | 1582 |
| 6. | to the | 1518 |
| 7. | of the | 1380 |
| 8. | it is | 1087 |
| 9. | to be | 968 |
| 10. | that i | 935 |
| 11. | I do | 829 |
| 12. | and I | 736 |
| 13. | and he | 728 |
| 14. | you are | 724 |
| 15. | of my | 696 |
| 16. | is the | 692 |
| 17. | I would | 674 |
| 18. | the king | 664 |
| 19. | he is | 658 |
| 20. | you have | 652 |



Part B:

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

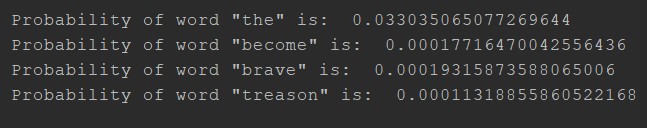
**1. Calculate the relative frequency (probability estimate) of the words:**

**(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 word is calculated as P(word) = count(word)/total\_no\_of\_word

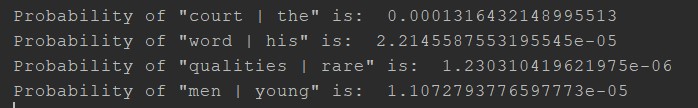
The count of a word is determined from the dictionary (where the key is the word and the value id the frequency). The total\_no\_of\_words is the length of the wordlist itself.



**2. Calculate the following word conditional probabilities:**

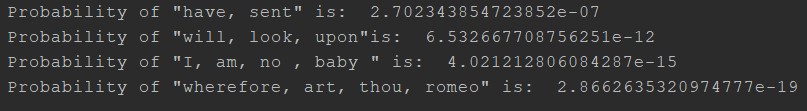
**(a) P(court | The) (b) P(word | his) (c) P(qualities | rare) (d) P(men | young) [Read P(B | A) as “the probability with which word B follows word A". Note: P(B | A) = count(A;B)| count(A) ]** The probability is calculated as:

P(A/B) = count(a,b)/count(a) count(a,b) was extracted from the bigram dictionary and count(a) was extracted from word dictionary.



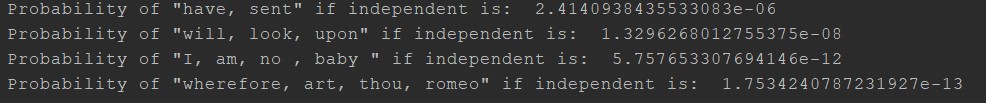
1. **Calculate the probability: (a) P(have, sent) (b) P(will, look, upon) (c) P(I, am, no, baby) (d) P(wherefore, art, thou, Romeo) Hint à use the chain rule (multiplication rule):**

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)



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

If the words are considered to be independent the probabilities are given as P(A,B,C,D) = P(A)\*P(B)\*P(C)\*P(D), the result is shone below.



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

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

The most probable word to follow **"I am no**" is: **more**

The most probable word to follow "**wherefore art thou**" is: **Romeo**

