Deerwalk Institute of Technology



Lab 6: Statistical Analysis of Content of Text File

(Artificial Intelligence)

Submitted by	: Submitted 1	to

Name: Shiva Tripathi

Roll No: 532 Batch: 2019

Section: A 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:

-	· \ IIaana\ Ch	. i \ 7	Data Lanall	\Programs\Python\Python36-32\python.exe C:/Users/Shiva/PycharmProjects/AI Practical/Part A 1.py
- 1	Rank	Word	Frequen	ncy
1.	: :			:
- 1	1	thou	1 54	443
-1	2	thy	38	812
-1	3	shall	36	608
-1	4	thee	31	104
-1	5	good	1 28	888
- 1	6	lord	1 27	747
-1	7	come	25	567
-1	8	sir	1 25	543
-1	9	let	1 23	367
-1	10	would	1 23	321
-1	11	well	22	280
-1	12	love	1 20	010
-1	13	man	19	987
-1	14	hath	1 19	917
-1	15	like	18	864
-1	16	know	1 17	763
- 1	17	one	1 17	761
-1	18	upon	1 17	751
- 1	19	go	1 17	749
-1	20	us	17	743

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:

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

Output:

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

- 2. Calculate the following word conditional probabilities:
- (a) P(court | The) (b) P(word | his) (c) P(qualities | rare) (d) P(men | young)

[Read $P(B \mid A)$ as "the probability with which word B follows word A". Note: $P(B \mid A) = count(A;B)$

| count(A)]

Output:

- 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:

	\Programs\Python\Python36-32\py	vthon.exe C:/Users/Shiva/PycharmProjects/AI_Pra
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:

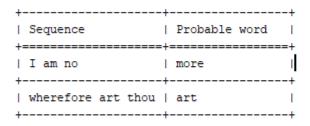
C:\Users\Shiva\AppData\Local\Programs\Python\Python36-32\python.exe C:/Users/Shiva/PycharmProjects/AI_Practical/Part_B_4.py Probability for independent words

+	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



Process finished with exit code 0