**LAB ASSIGNMENT 1**

**Name : Dipesh Khichada**

**Student Id : x22143271**

1. Take a text file (uploaded to Moodle) as an input of your program. Read its content. Write frequency of the words of the text in a file. Use python’s dictionary data structure for storing frequencies of the words. [hint: consider word as key and frequency of that word as value.

Step 1 Read File

Graphical user interface, text, application, email

Description automatically generated

Text, letter

Description automatically generated

Step 2 Frequency of Word

Graphical user interface, text, application, email

Description automatically generated

Step 3 Dictionary using Key Value Pair

2 Write most- and least-frequent words (100) in a file.

Most frequent words

Graphical user interface, text, application

Description automatically generated

Least Frequent

Graphical user interface

Description automatically generated with medium confidence

3 Count total number of words and total number of unique words in the corpus [corpus is a collection of text; here, set of sentences of the text file is referred as corpus]

Graphical user interface, text, application, email

Description automatically generated

4 Perform tokenisation (i.e separating punctuation from word) on the input text. Write the tokenised text in an output file [for tokenisation use the standard English punctuation list]

Graphical user interface, text, application

Description automatically generated

Text, letter

Description automatically generated

Text, letter

Description automatically generated

5 Perform (2) and (3) on tokenised corpus (from (4)).

A picture containing background pattern

Description automatically generated

Most Frequent Word

Graphical user interface

Description automatically generated with medium confidence

Least Common

A picture containing table

Description automatically generated

6 Perform lowercasing on the corpus text. Write the lowercased text in an output file.

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

7 Perform (2) and (3) on tokenised + lowercased corpus (from (6)).

Most Common Lower Case Word

Graphical user interface, text, application, email

Description automatically generated

Least Common Lower Case Word

Table

Description automatically generated

8 Did you find any difference in statistics when taking three different types of corpora as an input [i.e.

(i) raw corpus, In the raw corpus it is observed that each word along with punctuation was counted and so the number of total words and unique words were more than that of tokenized corpus Also it was observed that some words were joined with punctuation were considered as one word

(ii) tokenised corpus , In Tokenized corpus it was observed that words and punctuation were counted differently and this being the Reason for increase in Word Count and Frequency of Words also increased and unique words were increased

(iii) tokenised and lowered cased corpus In the Tokenized, In the tokenized corpus along with lower case it was observed that the word were considered as one as the camel Casing was Removed The Count dropped as total Number words were lessen and frequency of words also dropped Number of Unique Words were Reduced