Midterm – Spark Programming

In this report I will explain the lambda functions used in the program. I will explain the intermediate results yielded while running the program and I will include screenshots of the intermediate steps leading to the final outputs of the program. The screenshots will be from when I run the program on the “Encrypted-1.txt” file because it’s the same process for all 3 files. I will submit the 3 Decrypted files in order to show that the program properly decrypts all three encrypted files.

I will sequentially explain the program from start to finish and explain the functions of my lambda function:

-imported modules from sql.

-set up the environment and started a spark session.

- I read the encrypted file and cached it into memory.

-stored it a string array.

- Displayed the first 10 lines and counted the total number of lines which was 2000.

-An rdd function was used to convert the dataframe into an RDD.

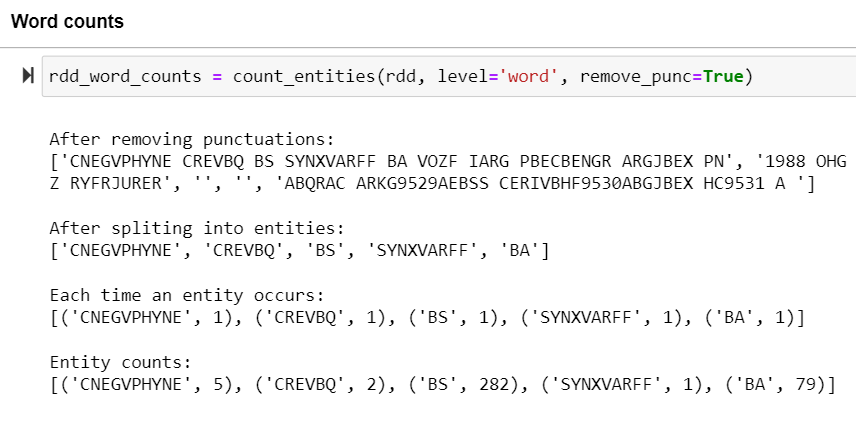
- A remove\_punctuation function to remove all punctuation from the string.

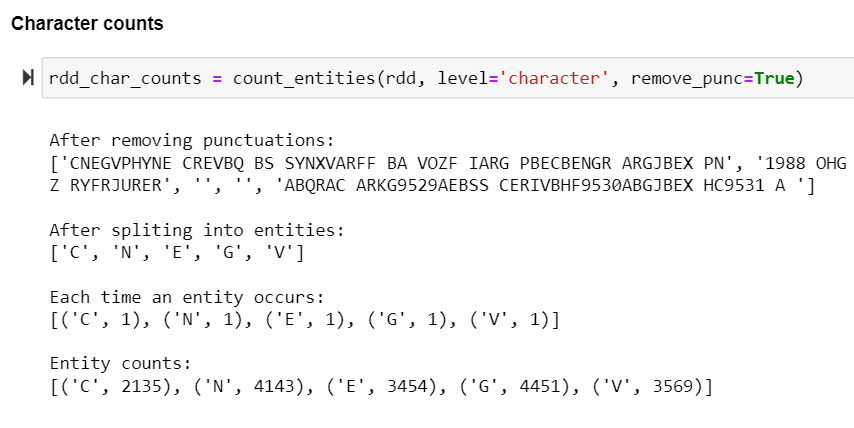
-A split\_character function to remove the blank spaces.

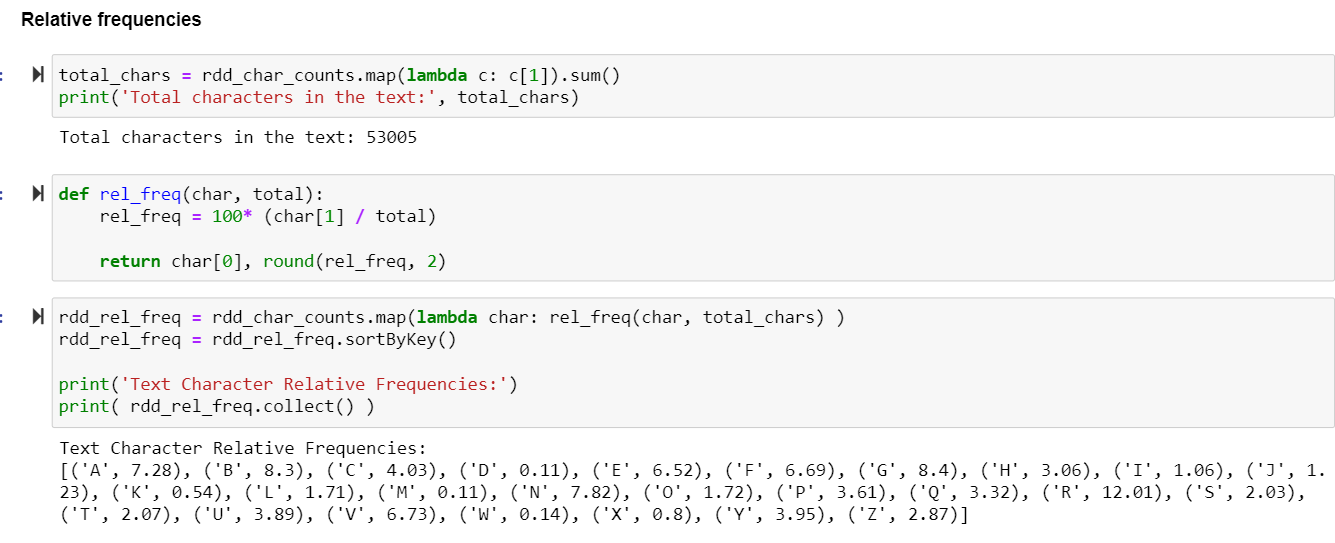
-A count\_entities function to split the string into words and count them.

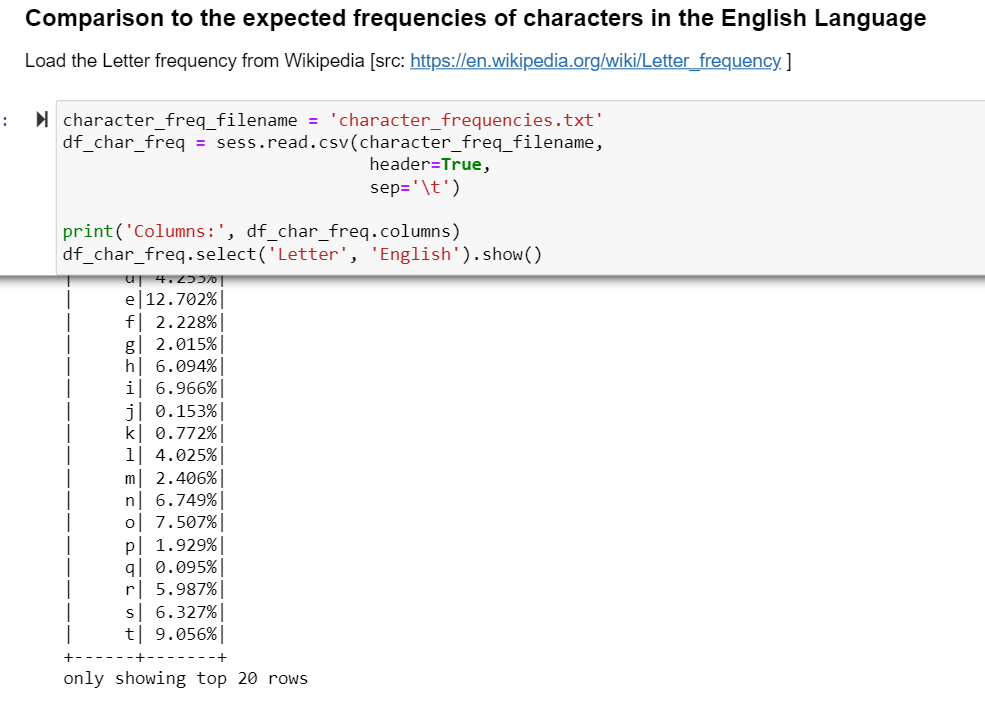
-lamda c to count total characters.

-A reletative frequency function to count the occurrence of each character.



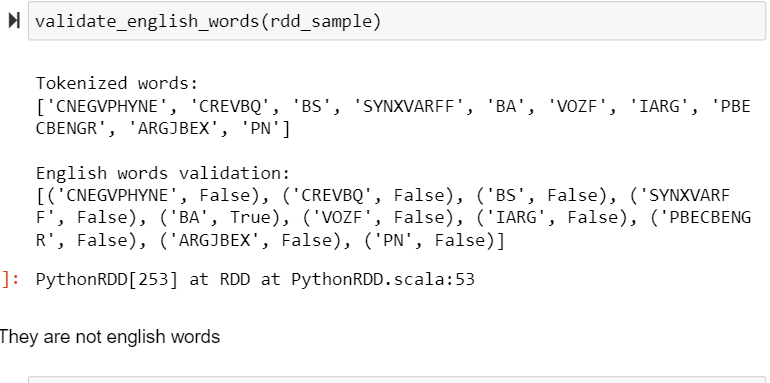






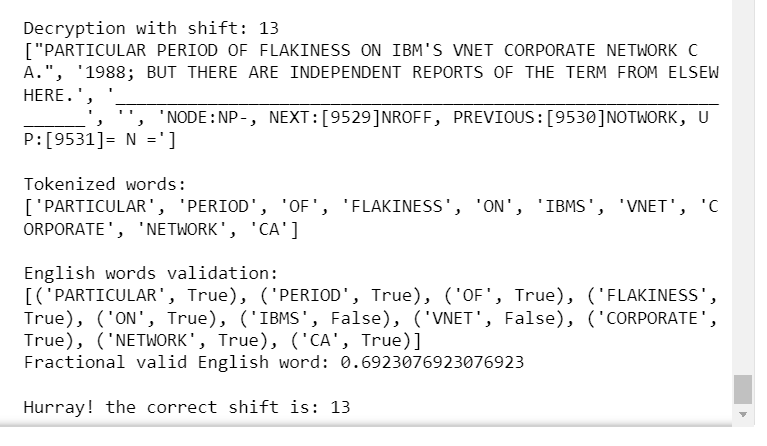


-Comparison of the words to English words using the natural language processing library,



-A decrypt function to find the correct decryption

-There were thirteen shifts of the alphabets until correct English words were produced



-A line lambda function to decrypt the entire file using the newly found shift.

-Output saved to a textfile “DEncrypted-1.txt”