**DS201: Statistical Programming**

Farhan Alam

12340740

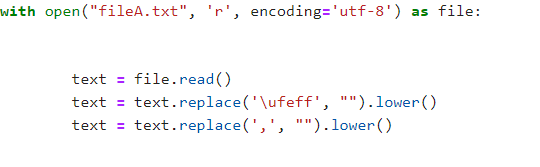
**Assignment no-1**

**Question no.1(a)**

**Introduction :** Given a file named fileA.txt we need to find the no. Of times a certain alphabet occurs in the file and then find its probability of occurring and then we need to rank the top 10 characters with respect to highest number of occurrence .

**Data** : The file is a text file which contains a sequence of alphabet with given to us is as follows .

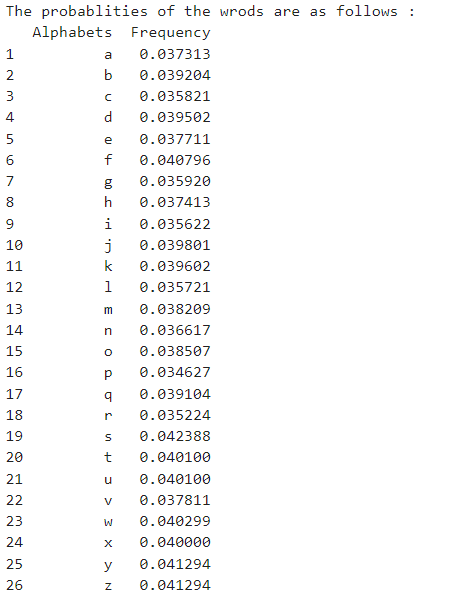
**Methodology:** Firstly the file is read through the following command:

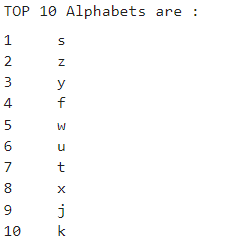


After this the characters are appended into a list then a list of length 26 is declared which stores the frequency of each character in the file then these two lists are used for formation of a DataFrame which consists of the 26 alphabets and their frequency , now the frequency column of the data frame is divided by the total no.of characters in fileA(10050 words) through which we get the probablity of the occurence of each alphabet. Now the Data Frame is sorted with respect to the probablity column and then the top 10 alphabets are displayed.The complet code for the process is as follows :-



**Result:** The output obtained from the code was as follows:





From the above we can infer that the word s was the highest occuring word and the word p was the least occuring word and the top 10 words were the one as printed above.

**Question no.1(b)**

**Introduction :** Given a file named fileB.txt we need to find the no. Of times a certain alphabet occurs in the file and then find the entropy for the given word document which is given by:

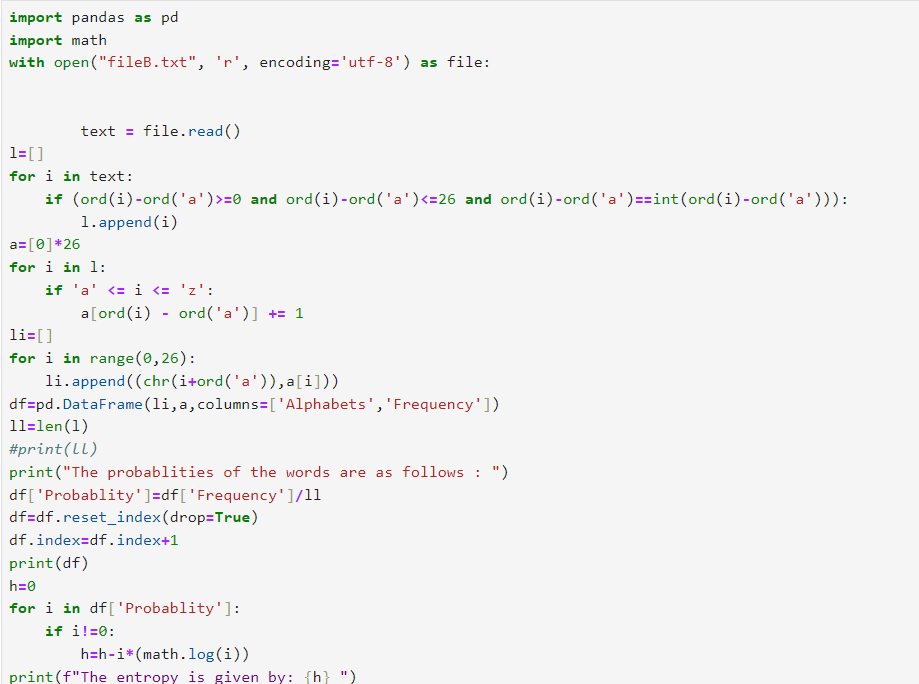


Where h represents the entropy and pi represent the probablity of the ith event.

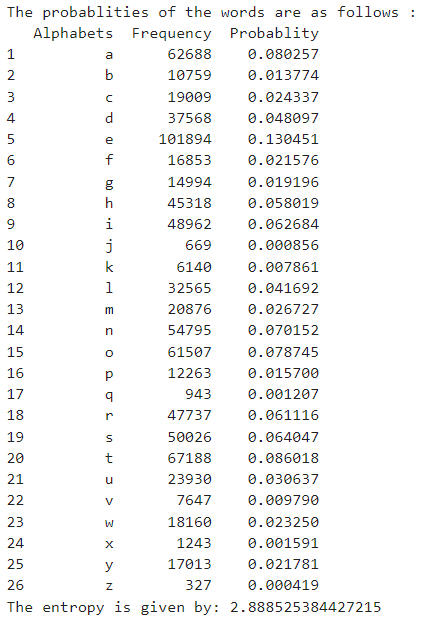
**Data** : The file is a text file which contains a sequence of alphabet with given to us is as follows .

**Methodology**: The same methodology as Question 1(a) was followed for extracting the words and then for calculating the frequency then the probability . Then additionally we traverse through the probability column and then added the value  to a variable which was declared zero initially thus calculating the entropy for the document.

The code used was as follows:-



**Result:** The result obtained was as follows:-



The entropy calculated was as follows : 2.888525384427215

**Question no.1(c)**

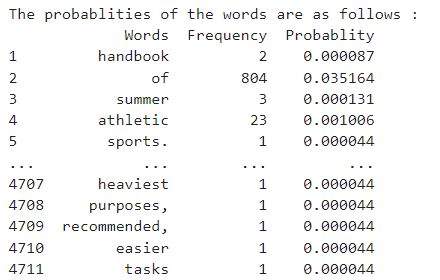
**Introduction :** Given a file named fileC.txt and fileD.txt we need to find the no. Of times a certain word occurs in the file and then find its probability of occurring and then we need to rank the top 10 words with respect to highest number of occurrence .

**Data** : The file is a text file which contains a sequence of alphabet with given to us is as follows :-fileC.txt and fileD.txt

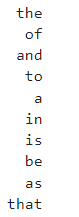
**Methodology:** Same as followed in question 1(a) and the same code is also used as the question except that here the operations were to be performed for the no. of words . The code used (for file C)was as follows:

Similar code was used for fileD.txt.

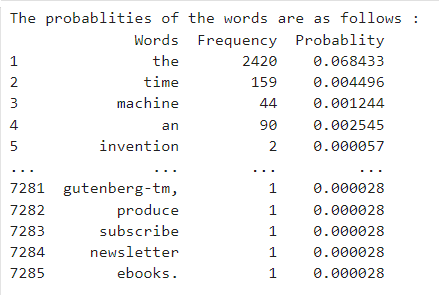
**Result:**The result obtained were as follows (FileC.txt):



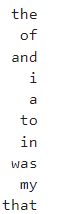
The top 10 words were:



Similarly for fileD.txt:



The top 10 words were:



We can observe that in both the document ‘the’ was the most used word and words like ‘of’ ‘and’ occupied the top spots in both the documents from this we can conclude that the most reccuring words are not a proper criteria to differentiate documents.

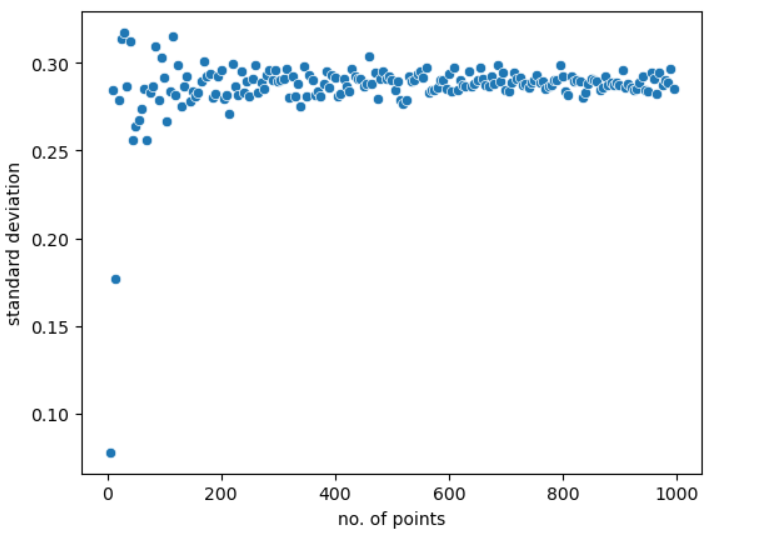
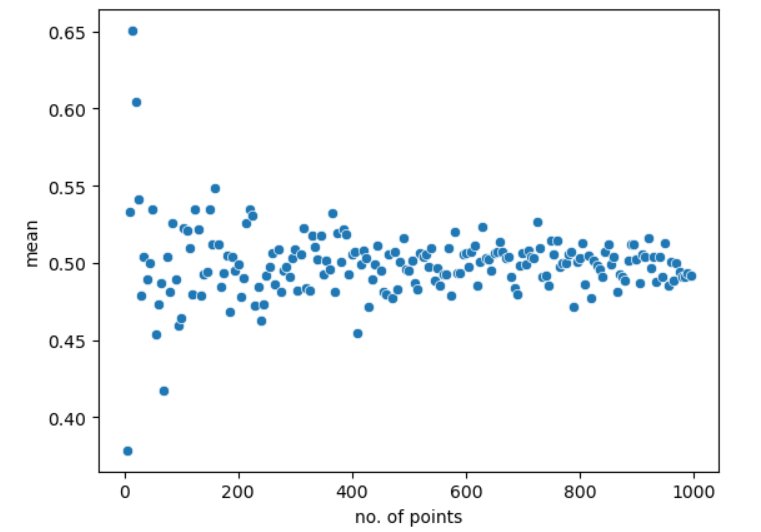
**Question no.2**

**Introduction :** This experiment investigates the behavior of random variables generated from a uniform distribution between 0 and 1. The goal is to explore how the mean and variance of a sample of *n* randomly generated numbers change as *n* increases. By generating different sample sizes and calculating the mean and variance for each, we can observe how these statistics evolve and whether they approach stable values as the number of samples grows.

**Methodology:** For the question we are iterating n from 5 to 1000 in addition of 5 each time and then declaring an array of the length of n where we append numbers ranging from 0 to 1 using a random no. Generator that follows the uniform distribution we know that mean for uniform distribution between 0 and 1 is 0.5 and the standard deviation is nearly 0.28 Everytime we calculate mean and standard deviation of the array then we plot the graph for the same with respect to no.of points the code used is as follows :-



**Result:** The resulting graph obtained are as follows :



We can observe that as n becomes larger the mean and the standard deviation tend to the theoretical value of mean and the standard deviation thus satisfying the central limit theorem.

**Question no.3**

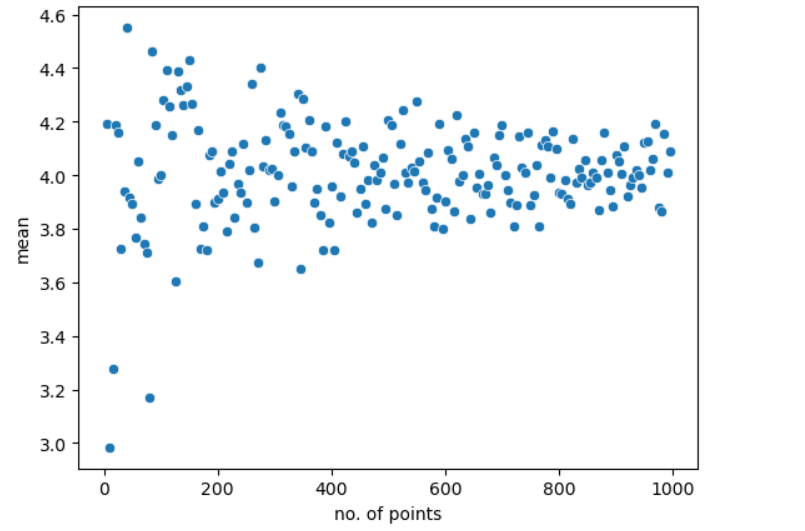
**Introduction :** This experiment investigates the behavior of random variables generated from a gaussian distribution with mean 4 and standard deviation 3. The goal is to explore how the mean and variance of a sample of *n* randomly generated numbers change as *n* increases. By generating different sample sizes and calculating the mean and variance for each, we can observe how these statistics evolve and whether they approach stable values as the number of samples grows.

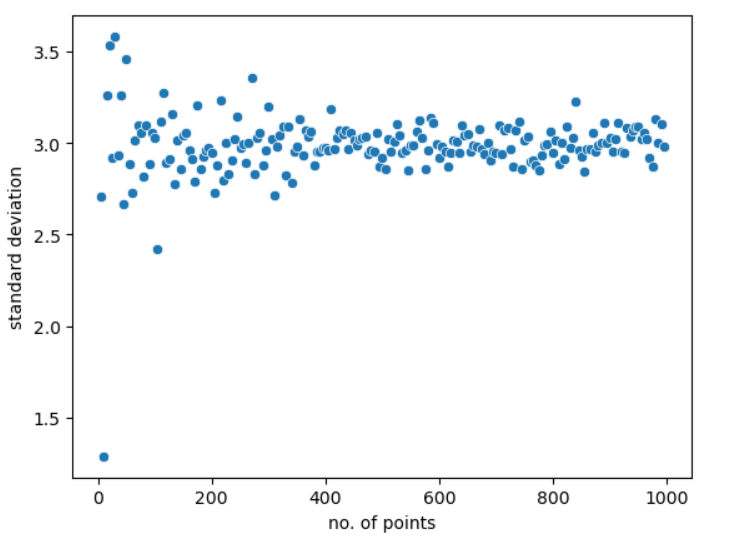
**Methodology:** The same methodology as question no.2 was used but the uniform random number generator function was exchanged with

A gaussian random no. Generator having mean 4 and standard deviation 3 the code used was as follows:-



**Result :** The following output was obtained :





Here again we can see that the actual value of mean and variance starts tending to its true value as value of n increases therefore proving the central limit theorem to be true.