|  |  |
| --- | --- |
| **Ex. 11** | **EXPLORING FILE OPERATIONS** |
| **Date:15/04/24** |  |

**Aim:**

To explore file operations in Python by writing programs for the following:

1. Create a new text file and write 10 lines into it.
2. Read the text file created above and count the number of words, characters, and lines in it. Find the frequency of occurrence of each word and character.
3. Read only the first 2 lines of the text file created above and print them. Then skip the next 4 characters and print the remaining characters in the line. Then, skip 3 lines and print the remaining lines.
4. Create a text file with several lines. Compute the unigram, bigram, and trigram probabilities (that is, probability of occurrence of each word, pair of words, and combination of 3 words). Remove all punctuation marks from the text before computing the probabilities. Store the computed probabilities in dictionaries and display them.

**Algorithm:**

**(a)**

**STEP 1:** Define a long multi-line string variable `text` with information about the Chambal River.

**STEP 2:** Open a new file named "text" in write mode.

**STEP 3**: Write the contents of the `text` variable to the file.

**STEP 4:** Close the file to ensure all data is saved and resources are properly released.

**(b)**

**STEP 1**: Initialize variables to track the number of words, characters, and create dictionaries to hold word and character frequencies.

**STEP 2**: Open the file "text" in read mode**.**

**STEP 3:** Iterate over each line in the file, and for each line, increment the line count.

**STEP 4:** Split each line into words, update the total number of words, and calculate the number of characters in each word.

**STEP 5:** Update the word frequency dictionary for each word and character frequency dictionary for each character in the words.

**STEP 6:** Print the total number of lines, words, and characters, and display the frequency of each wordand character from the file.

**(c)**

**STEP 1:** Open the file named "text" in read mode.

**STEP 2:** Read and print the first two lines from the file.

**STEP 3**: Read the next line, skip the first four characters, and then print the remaining part of this line.

**STEP 4**: Skip the next three lines in the file.

**STEP 5**: Print all remaining lines in the file after the skipped lines**.**

**(d)**

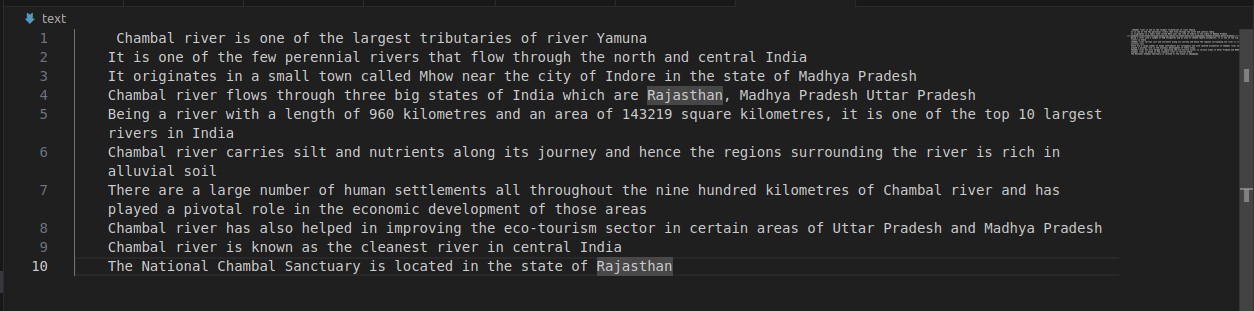
**STEP 1:** Open the file named "text" and read its contents into a variable. Split the content into words to create a list of words.

**STEP 2**: Create a list of unique words (unigrams) and calculate the probability of each word appearing in the text. Print the unigram probabilities.

**STEP 3:** Construct a list of consecutive word pairs (bigrams) from the text. Calculate and print the probability of each bigram occurring based on the frequency of the first word in the bigram.

**STEP 4:** Generate a list of consecutive word triplets (trigrams) from the text. Calculate and print the probability of each trigram based on the frequency of the corresponding bigram**.**

**Text file:**



**Program:**

**(a)**

text=''' Chambal river is one of the largest tributaries of river Yamuna

It is one of the few perennial rivers that flow through the north and central India

It originates in a small town called Mhow near the city of Indore in the state of Madhya Pradesh

Chambal river flows through three big states of India which are Rajasthan, Madhya Pradesh Uttar Pradesh Being a river with a length of 960 kilometres and an area of 143219 square kilometres, it is one of the top 10 largest rivers in India

Chambal river carries silt and nutrients along its journey and hence the regions surrounding the river is rich in alluvial soil

There are a large number of human settlements all throughout the nine hundred kilometres of Chambal river and has played a pivotal role in the economic development of those areas

Chambal river has also helped in improving the eco-tourism sector in certain areas of Uttar Pradesh and Madhya Pradesh

Chambal river is known as the cleanest river in central India

The National Chambal Sanctuary is located in the state of Rajasthan'''

f=open("text","w")

f.write(text)

f.close()

**(b)**

num\_lines=0

num\_words=0

num\_char=0

word\_freq={}

char\_freq={}

with open("text","r") as file:

   for line in file:

       num\_lines+=1

       words=line.split()

       num\_words=len(words)

       for word in words:

           num\_char+=len(word)

           if word in word\_freq.keys():

               word\_freq[word]+=1

           else:

               word\_freq[word]=1

               for char in word:

                   if char in char\_freq.keys():

                       char\_freq[char]+=1

                   else:

                       char\_freq[char]=1

print("Number of lines in the file:",num\_lines) print("Number of words:",num\_words) print("Number of characters:",num\_char) print("Frequency of occurence of words:",word\_freq) print()

print("Frequency of occurence of characters",char\_freq)

**(c)**

with open("text","r") as file:

 for i in range(2):

    print(file.readline())

    remaining\_chars=file.readline()[4:]

    print(remaining\_chars)

for i in range(3):

    file.readline()

for line in file:

    print(line)

unigram=[]

bigram=[]

trigram=[]

uni\_prob={}

bi\_prob={}

tri\_prob={}

f=open("text","r")

file=f.read()

words=file.split()

num\_words=len(words)

unigram=[word for word in set(words)]

uni\_prob={word:format(words.count(word)/num\_words,'.3f') for word in unigram}

print(uni\_prob)

bigram=[(words[i],words[i+1]) for i in range(len(words)-1)]

bi\_prob={(w1,w2):format(bigram.count((w1,w2))/words.count(w1),'.3f') for w1,w2 in set(bigram)}

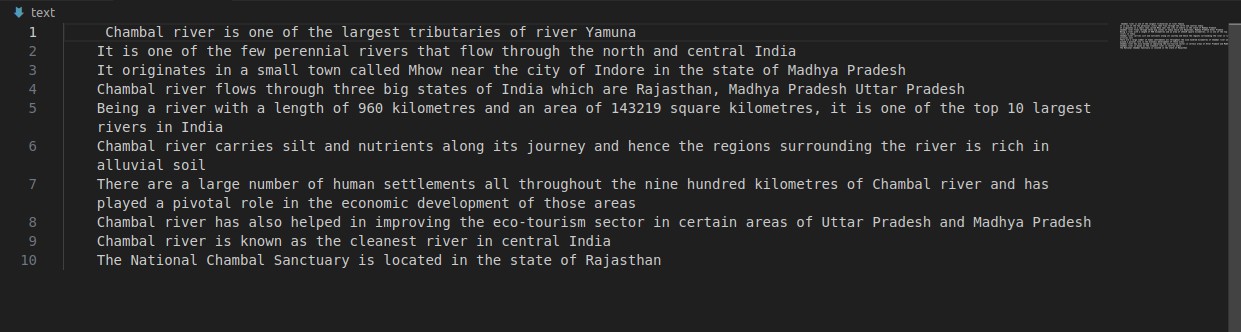
print(bi\_prob)

trigram=[(words[i],words[i+1],words[i+2]) for i in range(len(words)-2)]

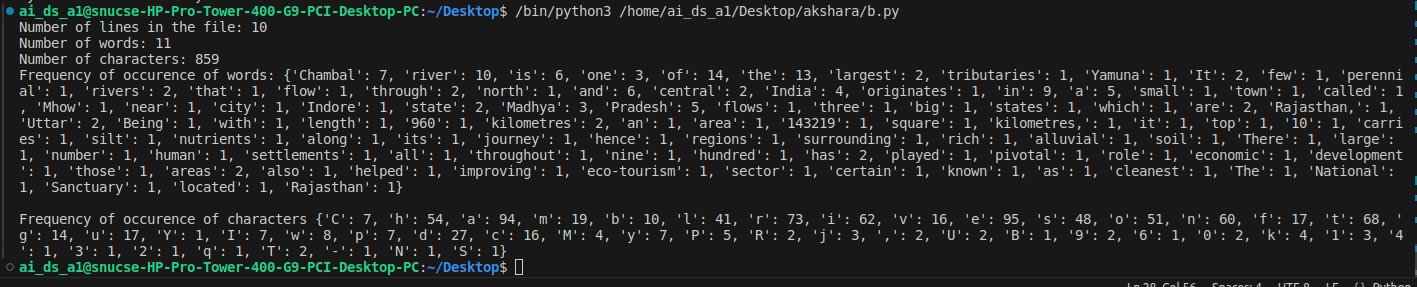
tri\_prob={(w1,w2,w3):format(trigram.count((w1,w2,w3))/bigram.count((w1,w2)),'.3f') for w1,w2,w3 in set(trigram)}

print(tri\_prob)

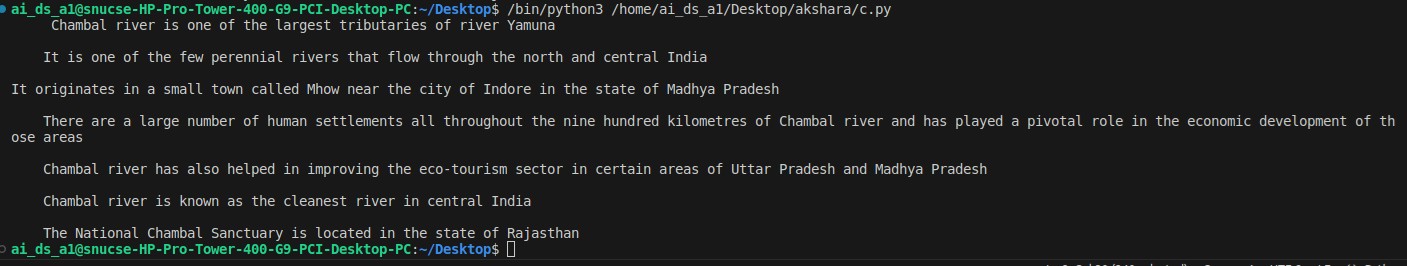
**Screenshot of Output: (a)**



**(b)**



**(c)**



**(d)**

**Result:**

Thus, programs have been written and executed to explore file operations in Python.