
DATA WAREHOUSING AND DATA MINING LAB (CSD-421) LAB ASSIGNMENT 3

AKSHAT RAJ VANSH (185520)

January 27, 2022

Computer Science Department
National Institute of Technology, Hamirpur

Contents

1 Program 1

1.1 Question

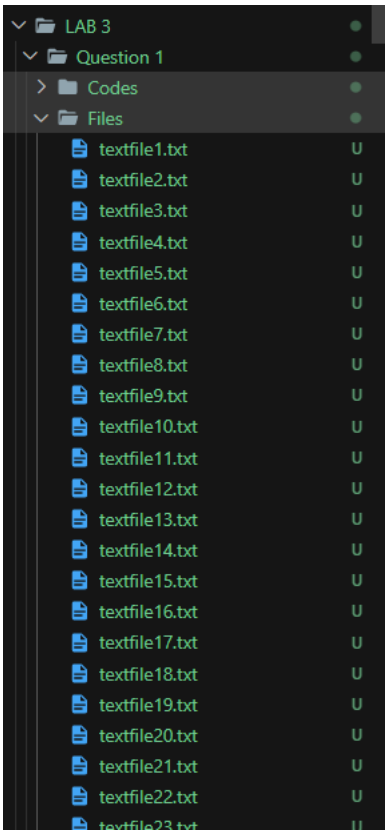
Create 500 txt files in a directory. Every file contains 20,000 lines and every line contain random string of length 20 characters.

1.2 Code

```
__author__ = 'Akshat Raj Vansh'
__version__ = '0.1.0'
__license__ = 'MIT'

import random
import string
import os


def generateRandomLines():
    text = []
    for i in range(0, 20000):
        text.append(''.join(random.choices(
            string.ascii_letters+string.digits, k=20))+'\n')
    return text


def writeToFile(filename):
    file = open(filename.encode('unicode_escape'), 'w+')
    randomString = generateRandomLines()
    file.writelines(randomString)
    file.close()


def main():
    directory = os.getcwd()
    directory = directory[:directory.rfind("\\")]
    filename = directory+'\Files\textfile'
    for i in range(1, 501):
        writeToFile(filename+str(i)+'.txt')


if __name__ == '__main__':
    main()
```

1.3 Output

Question 1 Outputs



2 Program 2

2.1 Question

Calculate the execution time to convert all the file to upper case. Save the results in csv file as given below.

2.2 Code

```
__author__ = 'Akshat Raj Vansh'
__version__ = '0.1.0'
__license__ = 'MIT'

import time
import os
import random
import string


def encodeString(word):
    return word.encode('unicode_escape')


def updateFile(filename):
    file = open(encodeString(filename), 'r')
    originalText = file.read()
    file.close()
    file = open(encodeString(filename), 'w+')
    originalText = originalText.splitlines(keepends=True)
    updatedText = [x.upper() for x in originalText]
    file.writelines(updatedText)
    file.close()


def generateRandomLines():
    text = []
    for i in range(0, 20000):
        text.append(''.join(random.choices(
            string.ascii_letters+string.digits, k=20))+'\n')
    return text


def writeToFile(filename):
    file = open(filename.encode('unicode_escape'), 'w+')
    randomString = generateRandomLines()
    file.writelines(randomString)
    file.close()


def rollback():
    directory = os.getcwd()
    directory = directory[:directory.rfind("\\")]
    filename = directory+'\Files\\textfile'
    for i in range(1, 501):
        writeToFile(filename+str(i)+'.txt')


def writeResults(result):
```

```
    file = open('result.csv', 'w+')
    row_out = []
    res_out = []
    for row in result:
        row_out.append(','.join(x for x in row))
    for res in row_out:
        res_out.append('\n'.join(x for x in row_out))
    file.writelines(res_out)
    file.close()

def main():
    directory = os.getcwd()
    directory = directory[:directory.rfind("\\")]
    filename = directory+'\Files\\textfile'

    table = [['No. of Files', 'Time Taken (sec)']]
    for j in range(1, 6):
        rollback()
        begin = time.time()
        for i in range(1, j * 100 + 1):
            updateFile(filename+str(i)+'.txt')
        end = time.time()
        table.append([str(j*100), str(end-begin)])
        print('Finished updating {n} files in {t} seconds'.format(n=j*100, t= end-begin))
    writeResults(table)


if __name__ == '__main__':
    main()
```

2.3 Output

Question 2 Outputs