CS594, Python Programming Lab

(https://www.iitg.ac.in/asahu/cs594/)

Assignment IV : Finding Similarity between Books

Deadline : 11.55 PM IST, 25th October 2020, 2 marks deduction per day after deadline

Write a Python Program to identify 10 similar pairs of text books from a given set of 64 text books in txt format. A simple (but crude) way to measures similarity is given as following:

- Identify top 100 frequent (normalize) occurring words in a text file. Consider only numeric character and alphabet character (by converting both small and capital alphabets to capital form) of the text file. Normalize the word count of a file by dividing total number of words in the file.
  - Remove commonly occurring six specific words “a”, “and”, “an” , “of” “in” and “the” from the file at the time of counting frequent word, total number of word and the normalize frequency.
- Similarity index of two files can be calculated as sum of normalize value of all the frequent words occurred in both the files.
- Create a 64x64 size similarity matrix.
- Report top ten similar pairs of text books using the similarity matrix. You need to exclude the self similarity.

(64 text books in txt format **Available at https://www.iitg.ac.in/asahu/cs594/)**

Submission procedure:

- **Create a folder with your name/roll number, put all the source codes and readme files in that folder**
- **Send your assignments code in compressed folder (tgx/zip/gz) to asahu < at > iitg < dot > ac < dot > in with "CS594: Assignment<III> , < RollNo > " as subject before the deadline**
- Please embed comments, how to run and required inputs properly in the code, or a separate readme file.
- **Please do not send the provided input files for the assignment**
- **Submitted code will be checked for Plagiarism**