Aniket Santra MDS202106

DCBD Assignment2

(1) What is the best savings you could achieve on the given input?

Ans:- I have achieved the savings of 97.68%

(2) What data processing steps did you perform?

Ans:- Checking the input files we can see there could be an address tag in html file containing address or a tag whose id or class(not href) does have address string or there could be a arbitrary tag having address string(with address). So we use RegEx in python and checking each above condition to extract the addresses. If all of the condition fails we use convert the html text file in normal txt and finding the keywords like ‘address’, ‘ADDRESS’,‘contact’, ‘contact us’, ‘pincode’ and extracting the string before or after the keywords.

(3) If provided more time, what more could you have done to improve your savings score?

Ans:- I could learn and use different Natural Language Processing method to improve the savings score if I could have more time.

(4) How easy/difficult was this task? What challenges did you come across?

Ans:- I faced much difficulty while doing this assignment as I haven’t done such text extraction, web scraping method before. This is completely first time.

The challenges I faced that first of all the html text files are not of fixed structure; there is no well-known fixed structure of address also. So I had to take into consideration different ways for the different html files to extract the addresses.