DCBD Assignment 1

Project Group members

Alok Dhar Dubey – MDS202304
Gauranga Kumar Baishya – MDS202325
Rohit Roy – MDS202340

1 Best Savings

We were able to achieve a saving percentage of 98.14% on the given five HTML pages. The code is written in such a way that it collects minimum text from the web pages while ensuring that most addresses are included.

2 Data Processing Steps

There are a couple of steps we incorporated in our code. But the main idea that we followed was, every address will have at least one of the following three things - city or district, state or union territory, and pincode. Without any of these, no address can be determined uniquely. Hence we made a list of all the cities, districts, states, union territories, and possible unique pincodes. Then we just used regex on the web page to find the occurrences of such words. On finding a match, we apply a left and right buffer to it, i,e. how much more characters do we want to take in our address. Then, having collected all these intervals, we combined all of them to get a collection of disjoint intervals. Then finally, we just used these intervals to extract parts of the webpage to include in the output.

3 Additional Steps

There are a few more things we could have included in our project. Right now, our code is prone to spelling mistakes. There can be an address where there is a mistake in the spelling of the state name. Then that address might not be considered for extraction. For this, we could have included an edit distance search, i.e., searching for all the words on the web page which are a few edit distance apart from a place's name. This will make our program more robust and better at collecting addresses.

Also, we could have spent more time to find an optimal magnitude for the left and right buffer. A magnitude that is not too large so as to include large amount of unnecessary text, and not too small as to leave out some parts of the addresses. We could have tested our code for a few random web pages with addresses and approximated appropriate values for these buffers.

4 Task Difficulty

Overall, the assignment was challenging and fun, where all of us applied our logic to make rules for extracting address, gave counter examples to each other's opinions, and tried to generalize the rules as much as possible. We came by many challenges in the processes, like parts of codes not working properly, rules getting failed on which we imposed so much trust, and even discarding big chunks of codes. It also had fun events like suddenly realizing a better way to perform the same function in a code in a smaller time complexity. All these things together led us to our final code and a savings percentage of around **98%**.