DCBD Assignment1

Arpan Biswas (MDS202011),
Chandrashish Prasad (MDS202015),
Megha Chakraborty (MDS202022)

April 2021

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

We were able to achieve a Total Space Gained score of **98.02%**. Earlier, we got a score of 98.44%. However, that led to some loss of address data so we decided to go with a different (current) code which gave us the maximum space saving to the best of our ability without loss of data.

2 What data processing steps did you perform?

1. Look for address HTML tag.
   Used `soup.find_all()` to find all `address` tags on the html page. If found, we `append(element.get_text())` it to `address_lines`.
2. Find and recognise HTML tags with class and id attributes that are meant for addresses.
   If found, we `append(element.get_text())` it to `address_lines`.
3. Remove Scripts, Styles, Head, Meta and Images.
4. Remove HTML tags that might be a Heading or Menu.
5. Extract all the text.

6. Remove unnecessary spaces and newlines from the text.
7. Remove chunks of text in long lines that are not close to any comma.

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

Possible scopes of improvement :

1. eliminate more html tags
2. eliminate footnotes if not containing address
3. if uniform template of webpages, then more specific improvements can be made
4. We could have collected more addresses using which we could have trained a recurrent neural network for parsing addresses

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

It was difficult at first. In the beginning the approach was towards recognizing or parsing addresses, but not much progress was made. After carefully studying the input HTML files, the idea was to think of the things that are definitely not part of addresses like headings, etc. Then some significant progress materialized. Once this hurdle was crossed, it got fun from there. Working as a team and sharing ideas and perspectives made the project more interesting and the work was done efficiently.