Distributed Computing and Big Data
ASSIGNMENT 1

1. INTRODUCTION

In the context of this project, we are tasked with the creation of a utility for Ms./Mrs. Ria, designed to extract information from a substantial collection of webpages. Our approach involves processing each HTML file to minimize its content, ensuring that no address data is omitted in the process.

A Python script was provided to us, which is capable of reading HTML files from a designated input directory and subsequently storing them in a specified output directory. We have incorporated the requisite code to facilitate data processing, resulting in the output directory containing a corresponding file for each input file, albeit with a significantly reduced size. This method ensures efficient data extraction without compromising on the completeness of the information.

2. REPORT

2.1. Our Achievement

The average reduction in data size that our team was able to accomplish with the provided input is 98.11%

2.2. What we did?

The Python libraries *re*, *pathlib*, and *BeautifulSoup* from the *bs4* package were utilized in this project. We have implemented a method to identify postal codes, which we have assumed to be any six-digit number that does not commence with a zero. For each identified postal code, we have extracted an address, which we have defined as the 200 characters preceding the postal code and the 15 characters following it. In addition, if a city name is mentioned and is followed by a comma within 15 characters, we have extracted an address, which we have defined as the 190 characters preceding the city name and the 25 characters following it.

2.3. What more could we do?

Given additional time, our team could further explore and implement various Natural Language Processing methodologies to enhance the efficiency of our data reduction strategy.
In particular we could perform data processing on a large corpus which can give us a probabilistic idea of what the surroundings of an address looks like.

2.4. Our Challenges

Given that our team had not previously engaged in text extraction or web scraping methods, the task presented a considerable challenge. The primary difficulty stemmed from the lack of a standardized structure for addresses. The inconsistency in the inclusion of postal codes and state names complicated address identification. Furthermore, despite the existence of address tags in HTML, many did not utilize them.

REPORT BY:
Nooh Ali MDS202337
Harsh Arora MCS202208
Syed Aslah Ahmad Faizi MCS202222