## Chennai Mathematical Institute

DISTRIBUTED COMPUTING AND BIG DATA        DEADLINE: JAN 30, 2023 11:59
PM. MAX MARKS: 10.

---

Instructions:

(1) Submit your assignment solution as a single zip file on moodle. It is sufficient if one person in the team submits the assignment. Ensure all team member names are clearly mentioned in your submission.

(2) You should work in a team of size four. For teams that do not have four members, a penalty of one mark will be applied.

---

## PROBLEM STATEMENT

Indian addresses show up on several web pages. For example, see `https://www.cmi.ac.in/merchant.php#Contact`. Figure 1 shows the address part in a red rectangle. Note that there may be multiple addresses appearing in different parts of the same web page.
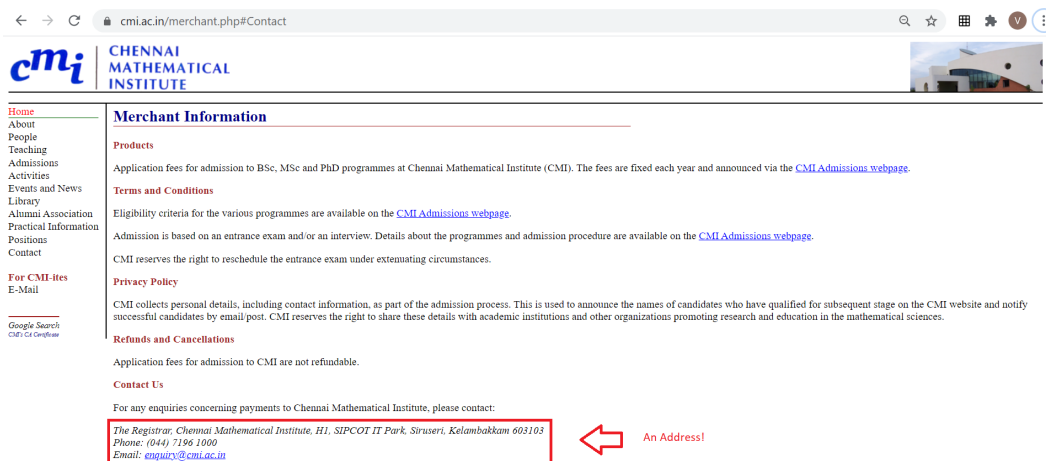


FIGURE 1. Address in a webpage

Ria's startup has a huge crawled set of web pages. The pages are available as html files in a folder. A sample of five such pages is provided to you as part of this assignment. Ria wants to develop a tool that can extract addresses from such web pages. Before developing such a tool, she wants you to take each html file and reduce it contents. You must keep the address and try to remove as much other content from the input files as possible.

Indian addresses are written in multiple ways. So, you cannot make strong assumptions on the address format. For example, addresses may appear in a single line or

across multiple lines. Addresses may not contain pincode. For evaluation, only websites containing at least one address in English will be chosen.

A Python script is given to you which reads html files from a specified input folder and saves them to an output folder. Insert the necessary lines to process the data so that the output folder has one file for each corresponding file in the input folder with a reduced size. Note that this python script already achieves 91.74% space savings.

You cannot use any existing address detection library since Ria's startup is a direct competitor. You are welcome to use any programming language of your choice.

Along with your code, please include a one/two page report in pdf format explaining the following:

(1) What is the best savings you could achieve on the given input?
(2) What data processing steps did you perform?
(3) If provided more time, what more could you have done to improve your savings score?
(4) How easy/difficult was this task? What challenges did you come across?

**Evaluation**. *A test set consisting of 10 such web pages will be used for evaluation. You will lose marks if any part of the address gets changed or removed. The input and output folders are compared to compute the total bytes saved. Your assignment will be scored based on the savings you could achieve on the test set, and the quality of processing you perform on the input data.*