Jan-Apr 2024/Assignment 1 Jan 2024

Chennai Mathematical Institute

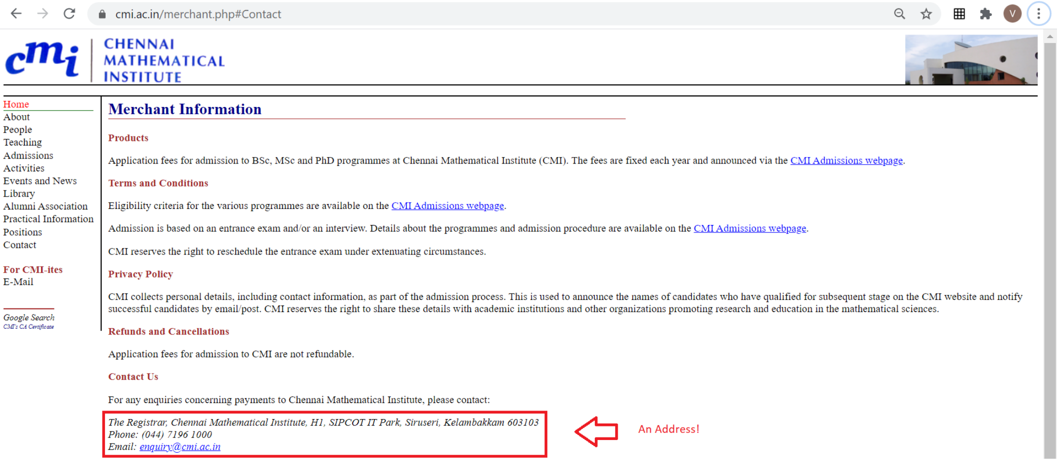
Distributed Computing and Big Data Deadline: Jan 29, 2024 10:00 PM. Max Marks: 10.

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 of three. For any reason, if you cannot be in a team of three, please take email consent from any one TA.
(3) You should not use any AI or ML libraries.
(4) Do not use libraries that do address parsing or detection.
(5) Simple text processing (eg., html handling and string handling) libraries can be used.
(6) Note that the html files given to you will not be used for evaluation. The evaluation set will contain 10 different web pages.

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.



Merchant Information

Products

Application fees for admission to BSc, MSc and PhD programmes at Chennai Mathematical Institute (CMI). The fees are fixed each year and announced via the CMI Admissions webpage.

Terms and Conditions

Eligibility criteria for the various programmes are available on the CMI Admissions webpage.

Admission is based on an entrance exam and/or an interview. Details about the programmes and admission procedure are available on the CMI Admissions webpage.

CMI reserves the right to reschedule the entrance exam under extenuating circumstances.

Privacy Policy

CMI collects personal details, including contact information, as part of the admission process. This is used to announce the names of candidates who have qualified for subsequent stage on the CMI website and notify successful candidates by email/post. CMI reserves the right to share these details with academic institutions and other organizations promoting research and education in the mathematical sciences.

Refunds and Cancellations

Application fees for admission to CMI are not refundable.

Contact Us

For any enquiries concerning payments to Chennai Mathematical Institute, please contact:

The Registrar, Chennai Mathematical Institute, H1, SIPCOT IT Park, Siruseri, Kelambakkam 603103
Phone: (044) 7196 1000
Email: enquiry@cmi.ac.in

An Address!

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 or more addresses in English may 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 answering 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?

In this report, do not forget to name all your team members.

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.*