Web Mining

Assignment 3: Find Business Email IDs (5% weightage)

Date Posted: Sep 18, 2014
Date of submission: Sep 25, 2014. 9pm.

Goal: To make students understand basic crawling, and use simple heuristics to handle real world unclean web data to get email ids.

Task: We will provide you with 2000 business webpages crawled from Yelp. Each webpage is an HTML containing details about the business. It does not have the email id, but it has the website address for the business which can be used to find the contact us page for the website and thereby extract its email id. Your task is to obtain structured data for the business: business name, business phone number, business home page URL, contact-us URL for the business, email id for the business.

The file webMiningAS3Data.zip contains 2000 html files from business1.html to business2000.html which correspond to html for business ids 1 to 2000.

Dataset:

1)Public Link: https://dl.dropboxusercontent.com/u/85901834/webMiningAS3Data.zip
2)IIIT DC++: /share/Assignments/Assignment3 under the user WebMining@DC

Submission Instructions: Create a directory with the name "<rollno>_as3".
Within that you need to put in the following files: The result file result.txt, a readme file README.txt and your code directory zipped as code.zip.
The result.txt file should have 2000 lines, 1 line per business. Every line must contain the following values separated by tabs: business id (i.e., the file name for the business), business name, business phone number, business home page URL, contact-us URL for the business, email id for the business.
The README.txt should describe what heuristics you used to locate the contact-us page for the website and also logic for your email-id extractor. README.txt should also contain instructions about how to compile and run your code.
The code directory should contain all your code. Zip the code directory to create code.zip

Finally zip the "<rollno>_as3" directory to get <rollno>_as3.zip and submit the zip file.