Week 0 HW1 - Create your own web crawler

Rishabh Agnihotri

April 12, 2016

1 Instructions

1. This assignment is a group effort.
2. Submission to be uploaded into your group repositories in the folder web_crawler
3. Deadline is the last week of text mining TA session.
4. Please follow the google python styleguide for your code. Pay attention to the guidelines naming, comments and main.
5. Code will be checked for plagiarism. Compelling signs of a duplicated effort will lead to a rejection of submission and will attract a 100% grade penalty.

2

Develop a web crawler as demonstrated in class. Refer to the basic principles discussed as guidelines. Write a document detailing how the crawler functions (it can be specfic to the website you have decided to scrape). How is your code robust? How do you deal with unexpected failures? What logs have you maintained? What features have you implemented specifically to adhere to the robots.txt file of your taget website.

Provide instrictions on ow to run the crawler on a sample of 10 documents/urls etc.