Assignment 4 : CS 751

HARISH

May 1, 2015

0.1 Question 1

Using the pages from A3 that boilerpipe successfully processed, download those representations again reprocess them with boilerpipe. I have used jusText library to remove HTML templates. It removes headers, footers and navigation links from the page. I have written a code in python named "htmlnonhtml.py" which is located in "q1-A4" folder. This code removes boilerplate content for the links present in "responseAfterCurlCalls" located in "q1-A4" folder .

Only 6100 links were unique , out of this 1496 are successful . Most of them are non-HTML skipped because of non content , Images on links , 404, 302 errors etc.

jusText was successful for pages that followed HTML standards. Pages which are properly organized .

The size of the file after removing boiler plate in A3 is "553KB". The size of the file after removing boiler content in A4 is "573 KB".

The time difference between A4 and A3 is

Time(A4)-Time(A3) : 28 days