Text Technologies – Assignment 1 – Statistics

s0837795

I. Q&A

Q: How long did your crawler run?
A: 851 seconds, if we obey the crawl-delay / request-rate directives, compute all checksums to ensure pages are not stale, and save them to disk. Runs for 44 seconds if we ignore rate directives. Runs for 6 secs if we also don't compute/save checksums.

Q: How many links did you extract in total?
A: 9401, out of which 807 distinct

Q: How many distinct URLs did you encounter during the crawl?
A: 841 – this includes links that led to 404 errors.

Q: How many pages have you fetched?
A: 807 distinct pages were fetched successfully.

Q: Did you encounter any links pointing outside of the domain?
A: Yes, links to facebook.com, turner.com, etc. if I remove the regular expression that limits crawls, however my final implementation by design doesn't capture any of these outside links.

Q: Did you get any errors during fetching? What kind?
A: Yes, a total of 34 (non-unique) HTTP 404 errors. In addition, if the URLs are not well formatted, we can get I/O errors.

II. Statistics

	Delay direct. - yes Checksums - yes	Delays direct. - no Checksums - yes	Delay direct. – no Checksums – no
Running time	851 secs	44 secs	6 secs
Total links extracted	9401	9401	9401
Number of followed pages	807	807	807
404 errors (non-unique)	34	34	34
Delay between requests	1 sec	0 sec	0 sec