1a.Identify and make changes to Nutch configuration to deal with politeness:

b. Identify and make changes to Nutch configuration to deal with URL filtering.

c. Create team identification information and label your Nutch bot via Nutch configuration.

Paste screenshots of our nutch-site.xml

2c.

Identify at least 10 different image MIME types that you encounter while crawling:

image/jpeg,image/jpg , image/gif, image/png, image/ico,image/tiff,image/bmp,image/ief

2d.

Errors:

java socket not found : timeout exception

502 error

403 error

404 error

For 2d get Gauravs segments and crawldb for statistics.

URLS:

1.www.midwestguntrader.com--problem cookie

2.www.advanced-armanent.com- error name not resolved .. dns lookup failed

3. www.davidsonsinc.com- requires a specific FFL number for dealer resgistration

4. www.firearmlist.com- website is not clickable as it is for sale

5.www.firearmslist.com- dns problem

6.[http://www.freegunclaXssifieds.com/-](http://www.freegunclaxssifieds.com/-) dns lookup problem

7.<http://www.fhclassifieds.com/-> website not found or moved resource

8. <http://www.comprayventadearms.com/-> website not found

9.<http://www.comprayventadearms.com/-> not found

10. <http://www.midwestgtrader.com/-> not found

11.<http://newmexicoguntrader.com/-> error

2e. Deliver the crawl statistics from your crawls of each repository. i. Crawl statistics should be information e.g., about what URLs crawled, the HTTP response, etc. You can deliver this as a text file included with your final deliverable zip outlined in Section 6.

Install Selenium in shivens laptop and run final crawl to get craw statistics for point 5.

3b. Extend and develop a plugin for protocol-interactive-selenium to better handle the problem URLs you have identified

5. Re-run your Weapons crawls with the enhanced Tika and Nutch Selenium you’ve built in step a. Identify at least 100 URLs that you have difficulty fetching and identify why

b. Are the URLs present from 2d still?

c. Did the enhanced Tika parsing assist with that?

d. Deliver your updated crawl statistics from each repository your crawl

7

c. Compare the resultant URLs from using this Filter to your results from Step 6 – were your deduplication algorithms more effective? Which ones (6a or 6b) and why?

d. Deliver your updated crawl statistics using the Similarity Scoring Filter.