CSCI 572: Assignment I Extra Credits Report

Crawling and Deduplication of Weapons Images Using Nutch and Tika

Team number: 33
Team members: Gaurav Shenoy, Mahesh Kumar Lunawat, Pramod Nagarajarao, Presha Thakkar, Karthik Kini

Task 8: [Extra Credits] Nutch Python

We performed following step to learn about the Nutch Python and have included the python file with the result zip file.
1) Installed Nutch□Python using pip
2) Downloaded and build the latest version of Nutch□Trunk in separate terminal
3) Started Nutch REST Server by running $ bin/nutch startserver □port 8081
4) Wrote a python script for crawling of seed urls with Nutch REST Server
5) Created custom configuration file and used it in python script for nutch crawling
6) Were able to GET/POST data using Nutch REST server.

Example:
GET /Config/Custom□config/http.agent.name ———> Team 33 : CSCI 572
GET /Config/default/file.content.limit ———> 65536 POST /db/crawldb { “type” : “stats”, “confId”: “custom”, “crawlId” : “Nutch□Crawl01”, “args”: {“ param” : “value” } } ———> gives crawl statistics with statusValue, avgscore etc.

Task 9: [Extra Credits] Tika similarity and D3

a. Take the resultant JSON output and visualize it using D3.
   We did 1 visualization with data of 250 images which we have attached in folder D3. The other 3 visualizations are made with data of about 25,000 images. The cluster visualization wasn’t able to print to pdf, so we have sent the other two visualizations in D3 as well.
   We have attached json of cluster and circlepacking in the D3 folder as well. This json can be used to view cluster, circlepacking and composite visualization.

b. Are their any interesting clusters? Can you explain the clusters that you see.
   Yes, there are a lot of interesting clusters. Most of the clusters are made due to the a specific property in Metadata of the image. Example a cluster of images with similar heights,widths, etc. Also, few clusters were formed of images from the same site, as they had similar Metadata.

Task 10: [Extra Credits] Crawl using Memex explorer

a. **Describe any bugs you encountered in Memex Explorer**
   Some of the bugs that we encountered are as follows:
   1. We started installing Memex-Explorer by cloning from Github on MAC system. Commands - ./app_setup.sh, source activate memex and supervisord were run on the directory where memex-explorer/source. In the browser http://localhost:8000 was started. Few minutes after installation supervisord stopped working and the server was down.
   2. When the above commands were run on an Ubuntu machine, installation of the Memex-explorer, running the server and the crawling were made with ease.

3. On an Ubuntu machine, when the server is stopped, the control automatically doesn't come out of bash which was running the memex-explorer server. A clean exit wasn't done and hence was bit tough to comprehend manually stopping the server process.

b. Were you able to run your crawls?

Yes, we were able to run our crawls

c. What was missing from Memex Explorer's Nutch capabilities?

1. The Visualization crawl log when we run a crawl command doesn't function appropriately.
2. An enhancement module for the configuration files of Nutch can be introduced which has a drop-down facility for each property and its value.