**Team #32**

Sanjay Jagaeesh

Manohar Thagadur Natraju

Soumya Gowda

Pramod P. Setlur

CSCI–572 Assignment 1

*Crawling using Apache Nutch*

Dark and deep web are two potential places to sell guns illegally. We used Apache Nutch to crawl for images of guns, Apache Tika for content and metadata extraction, and Selenium to automate Ajax interaction and more.

CSCI–572 Assignment 1

## Question 1: Download and configure Nutch to crawl Weapons images as identified in the seed list.

1. The list of files we modified to deal with politeness are as follows:

Nutch-Site.xml:

* + 1. We set http.agent.rotate to ‘true’ and added 7 agent names in agents.txt.
    2. We increased the http.timeout from 10,000 to 80,000.
    3. We changed the http.content.limit to -1 to remove the limitation on the size of the crawling file.
    4. We increased the fetcher.threads.fetch from 10 to 20.

Our Nutch-Site.xml contains the other changes.

* 1. Regex-urlfilter.txt and Automaton-urlfilter.txt
     1. We skipped domains like facebook.com, tumblr.com, twitter.com, etc. to filter majority of images unwanted to us.
     2. We removed the image extensions to avoid skipping images.
  2. Mimetype-filter.txt:
     1. We added image/\* mimetype to allow crawling of documents with images/\* and text/html mimetype.
  3. Suffix-urlfilter.txt:
     1. We commented all the image suffixes to prevent them from exclusion.

1. We enabled rotating bots in agents.txt.
   1. Spider#32Crawler, MasosapaCrawler, etc

## Question 2: Perform crawls of Weapon’s images sites

1. We wrote a python script to generate all the image mimetypes that were found in our crawled data.
   1. We found the following mimetypes:
      1. image/png
      2. image/jpg
      3. image/jpeg
      4. image/gif
      5. image/bmp
      6. image/x-icon
      7. image/tiff
      8. image/x-xbitmap
      9. image/pjpeg
   2. Many of the URLs that weren’t fetched were mainly because of java.net.SocketTimeoutException, HTTP-500, HTTP-403.