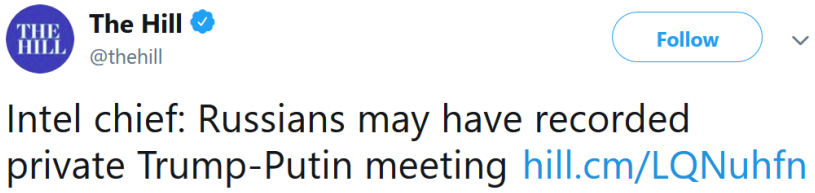
CS6200 Information Retrieval
Fall 2019
Instructor: Omar Alonso

Homework #5. Crawling and link statistics

Objective: Implement a simple focused crawler by extracting links from Twitter. Using the Twitter API (https://developer.twitter.com/), your task is to use a hashtag as query and extract links from Twitter. For example, say that your query is "trump" and you retrieve via the API the following tweet from @thehill.



As we can see, the above text contains a link that points to a real web page. You need to find if such page contains links and follow them using the crawling techniques described in class. Your task is to extract tweets that contain links and use those links as seed for a crawler.

This assignment involves:

1. Given a hashtag (you can define which one), extract a sample of tweets that contain links. Please use English only hashtags.
2. A single-threaded crawler that uses the output of (1) to crawl the web with two parameters: number of links and depth (levels).
3. Normalize links
4. Extract 20,000 links and compute the following statistics from the generated data set:
    a. Number of unique links extracted
    b. Frequency distribution by domain
    c. Breakdown of links by type (e.g., text, image, video)
    d. Average link depth
    e. For each crawled page, compute the number of incoming and outgoing links. Report the top-25 pages with the highest number of incoming and outgoing links.
    f. Plot the top-50 domains ranked by highest number of incoming links. Note that this is a computation for domains (e.g., cnn.com, bbc.co.uk) and not individual pages.

What you need to submit

1. Your tweet extractor in Java source code
2. Your single-threaded focused crawler in Java source code
3. The output statistics and charts for item (4).
4. Documentation on how to compile and run the code.