Twitter Stream Crawler Project

UCR CS172 - Winter 2015

Nicolas Lawler and Henry Garcia

Collaboration Details

- Henry
    - connected to the twitter stream, saved the tweets to json files, implemented the script for crawling the tweets, set up the documentation
- Nicolas
    - implemented the link identifying logic, implemented the title grabbing functionality, performed disk io as necessary to support the above, refined the documentation

Overview of System

- System Libraries
    - json-simple 1.1.1
    - jsoup 1.8.1
    - twitter4j 4.0.2
    - JavaSE 1.7
- Architecture
    - The System is split into two main parts, the Crawler and the TitleFetcher.
        - The Crawler connects to the twitter streaming API and downloads a number of tweets (specified by the user), writing them to disk in batches of files. The size of the files are determined by the user
        - The TitleFetcher takes as input the directory that the tweets were saved in, and loads them a file at a time. It searches the text of the tweet for links, grabs the title of the html response when it finds one, and adds that title to the json of the tweet. All the tweets are then written back to the same file that they came from.
- Crawling/data collection strategy
    - Crawler
        - The crawler first checks the directory for a file containing saved tweet ids. If the file exists the hashmap will be loaded with the tweet ids, otherwise the hashmap will start empty.

        - The tweet crawler first creates a filter in order to grab more tweets that are in English and have a location. After creating the filter it establishes a connection with twitter's stream, using account and consumer tokens provided through https://dev.twitter.com/
        - After establishing a connection with the twitter stream the system will start to grab tweets. First it will check if the tweet exists in the hashmap. If it does, then it will move onto the next tweet.
        - If the tweet has not been crawled yet, then it will add the id to the hashmap, add the id to the hashmapque, and check if it has geolocation information.
        - If the tweet has geolocation, then it will be added to the saveTweetQueue. otherwise it will move onto the next tweet.
        - The hashmapque and saveTweetqueue are used to store the information into the files after the crawler has reached a certain amount of crawled tweets (100 by default).
        - When saving the tweets the crawler will check if the file is larger than that specified by the user. The crawler will continue to increment the tweet file until it finds a file that is no longer larger than the file size provided by the user (useful if the user has already crawled tweets and doesn't want them to be placed in already crawled files).
    - TitleFetcher
        - The TitleFetcher loads a File[] full of all the files in the provided directory, and then processes each file one at a time, skipping over files that don't have tweets.
        - It parses the json into a JSONObject line by line and inserts that object into the input list. Then, for each item in the list, take the tweet and pass it off to a thread pool, where it will be serviced by the next available thread. Each worker begins by checking if the tweet has been processed before. If it hasn't, we get the text from the JSONObject, and attempt to find a url. If a url is found, we follow it and wait for an html response. If the response comes through, we parse it and get the title. Then, we add that title as a new field in the JSONObject. If the link turned out to be bad or the title was empty, we mark the tweet with a field "hasBadLink" so that it won't be searched again. Whether or not any fields have been added, we enqueue the tweet onto the output queue and the thread moves on to the next available tweet.
        - When all the tweets have been processed, the contents of the output queue are written back to the same file that they came from. Then the process starts again for the next file.

- Data Structures
    - HashMap
        - For hashing crawled tweets so that they won't be crawled again
    - ConcurrentLinkedQueue
        - For thread-safe writes when collecting processed tweets.
    - ArrayList
        - For loading and storing sets of tweets.
    - JSONObject
        - A simple json wrapper over a java HashMap, allows automatic encoding and decoding of JSON strings, used to represent tweets

Limitations

- If a file called "hashedtweets.txt" or a tweet file in the following format "tweets.*.json" exist and are not formated the way the crawler expects them to be, then the program will likely throw parsing errors and not function correctly.
- If the user cancels (^C) the system while the system is writing to the hashed tweet file or the tweetfile, then there is a possibility that the file will be corrupt and throw a parsing error on startup (if run in the same directory).
- The regular expression that matches to urls in the text string isn't fully robust. for example, it matches to urls ending in '.', and the resulting http request fails in a case where it may have succeeded.

Deployment

./runCrawlerExecutable.sh <Tweets> <File Sizes (mb)> <Threads> <outputdir>

OR...

./runCrawlerExecutable.sh <Tweets> <File Sizes (mb)> <Threads> <outputdir> <save Every # Tweets>

<Tweets> refers to the number of tweets the user wants to grab.

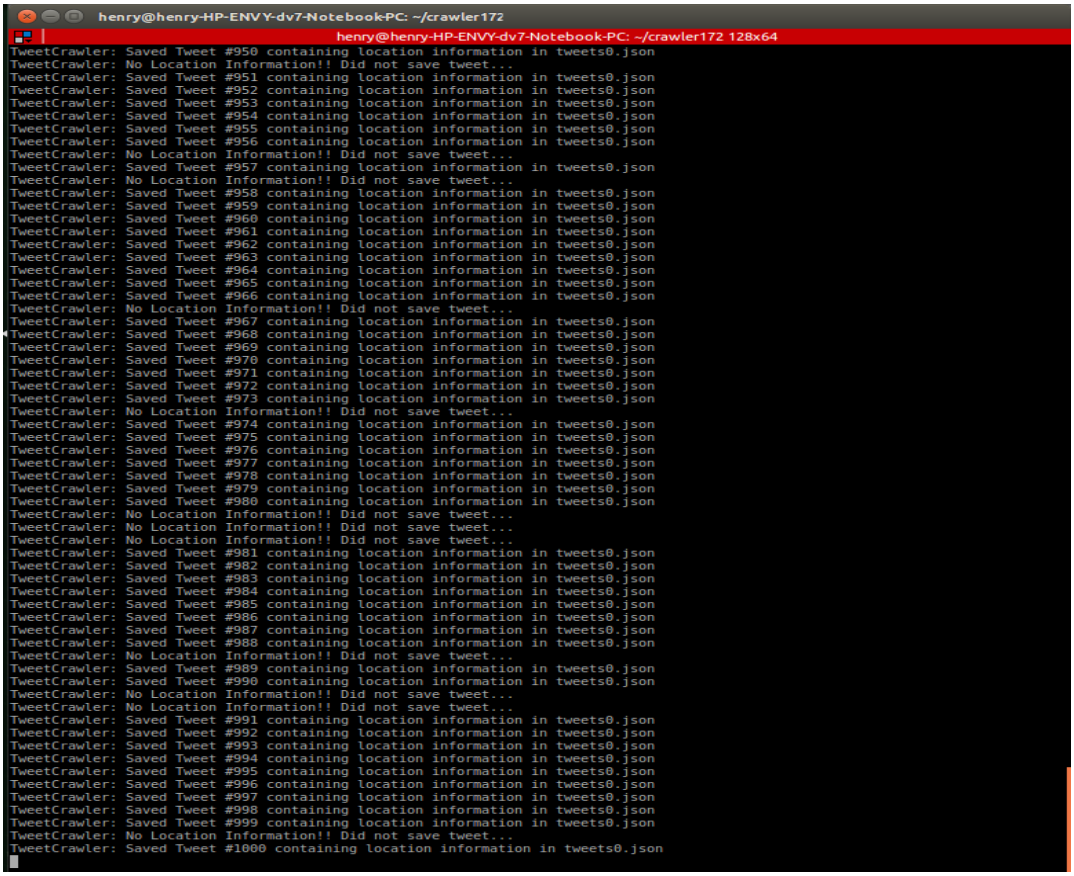
<File Sizes (mb)> refers to the size of files the user wants the tweets to be stored in

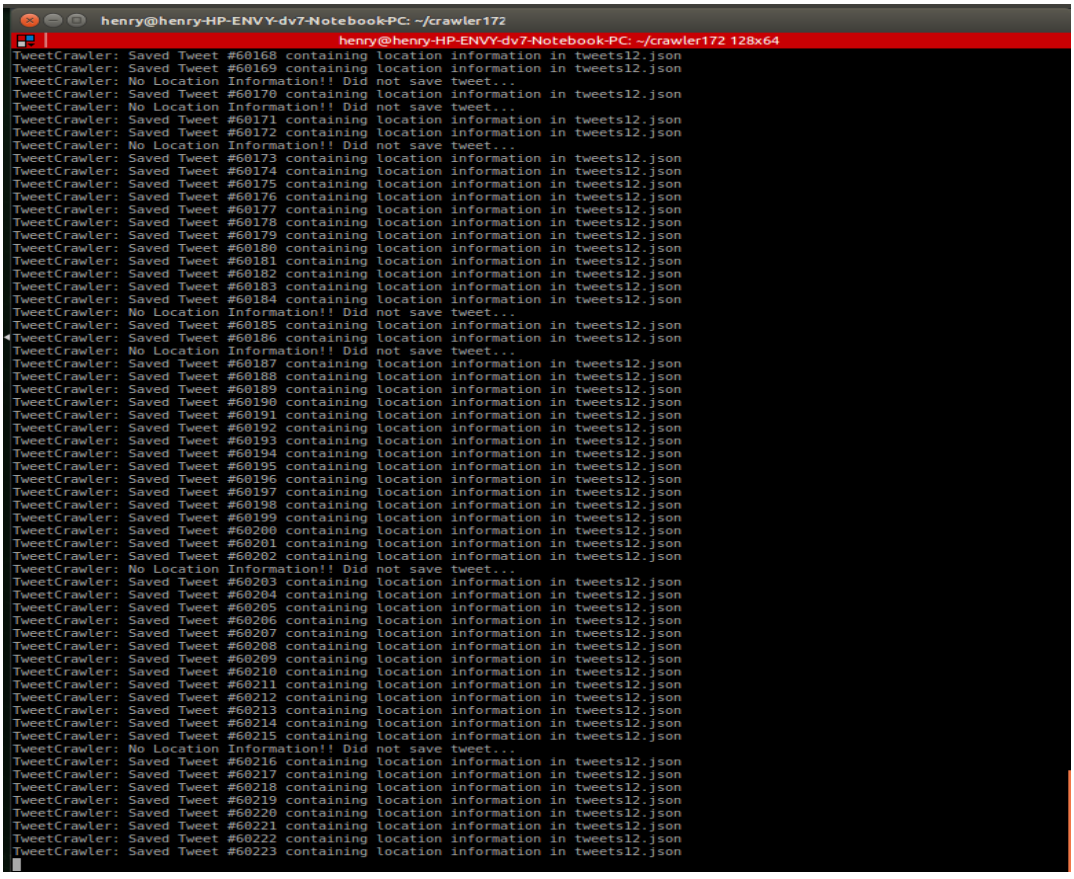
<Threads> specifies the number of threads the user wants to create

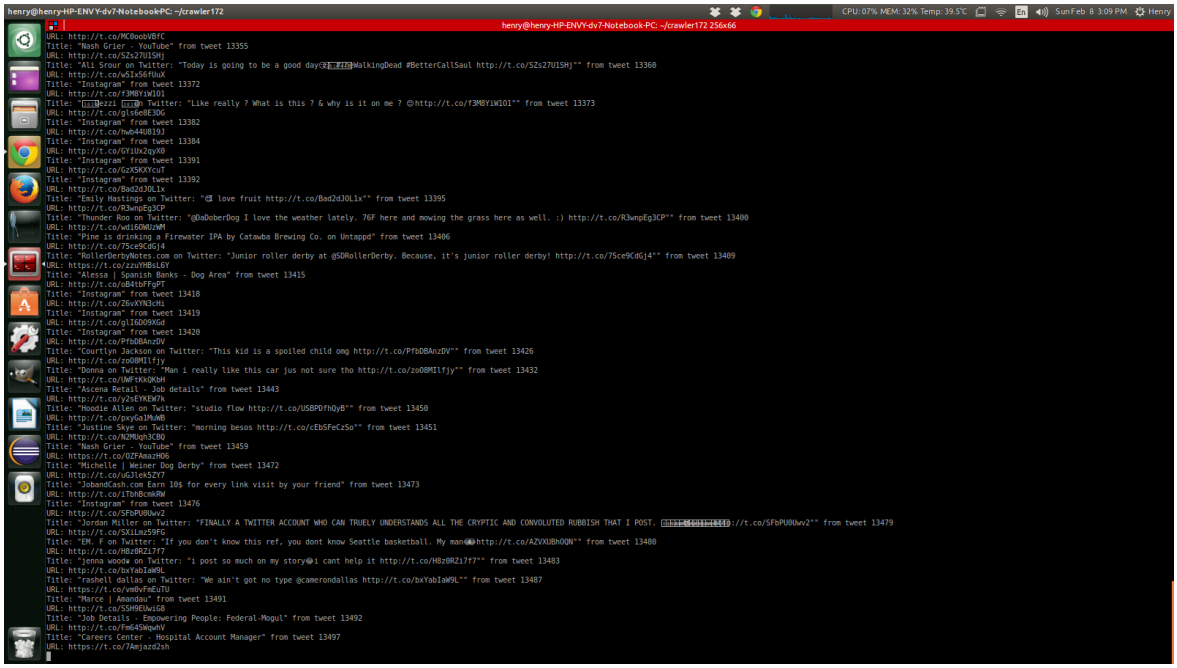
<outputdir> specifies the directory in which the user would like to store all of the data.

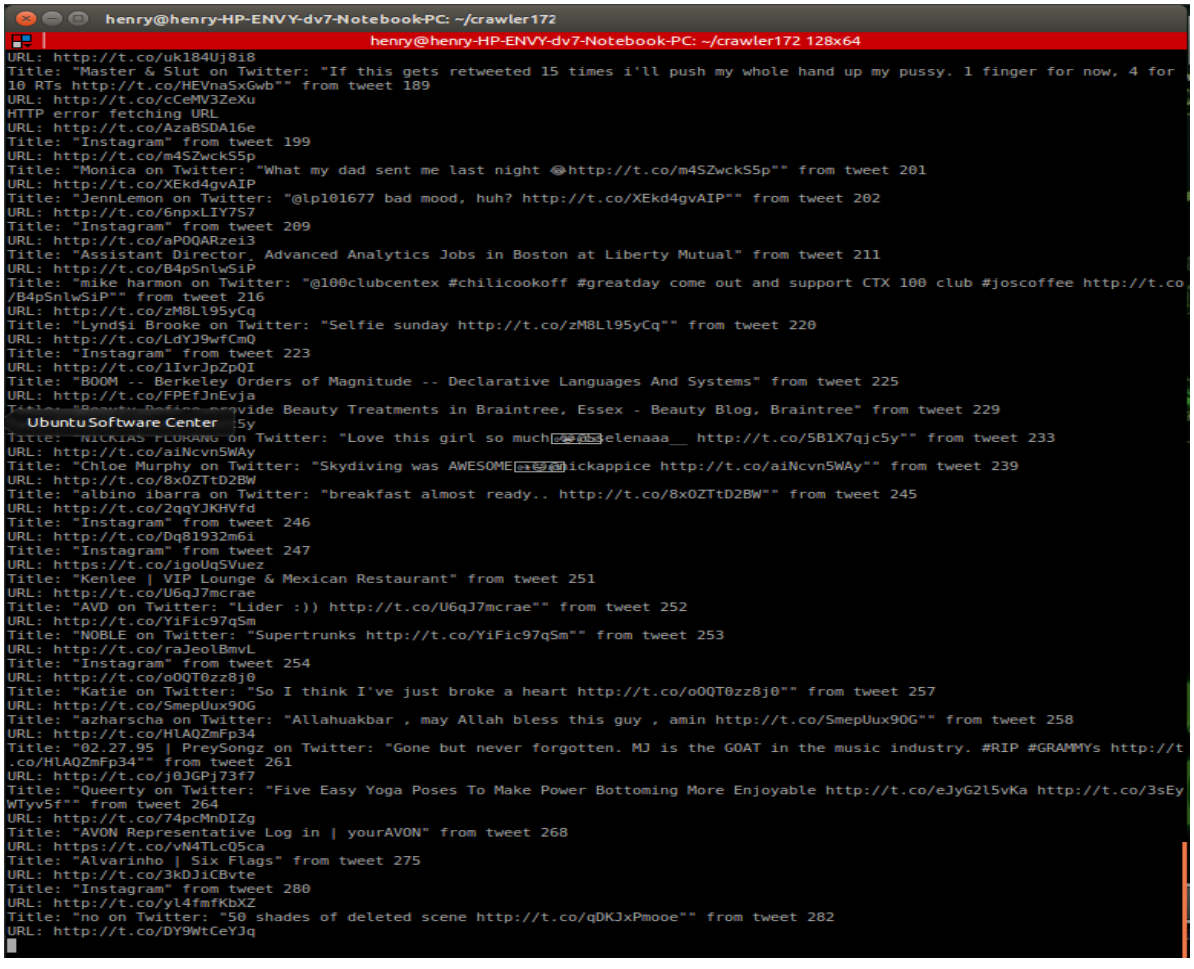
<save Every # Tweets> specifies how many tweets the user would like to obtain before the crawler saves the tweetids and tweets into the corresponding files. 100 is the default setting.

Screenshots






henry@henry-HP-ENVY-dv7-NotebookPC: ~/crawler172
CPU: 07% MEM: 32% Temp: 39.5°C
Sun Feb 8 3:09 PM
Henry
henry@henry-HP-ENVY-dv7-Notebook-PC: ~/crawler172 256x66


henry@henry-HP-ENVY-dv7-Notebook-PC: ~/crawler172
henry@henry-HP-ENVY-dv7-Notebook-PC: ~/crawler172 128x64
URL: http://t.co/uk184Uj8i8
Title: "Master & Slut on Twitter: "If this gets retweeted 15 times i'll push my whole hand up my pussy. 1 finger for now, 4 for 10 RTs http://t.co/HEVnaSxGwb"" from tweet 189
URL: http://t.co/cCeMV3ZeXu
HTTP error fetching URL
URL: http://t.co/AzaBSDA16e
Title: "Instagram" from tweet 199
URL: http://t.co/m4SZwckS5p
Title: "Monica on Twitter: "What my dad sent me last night http://t.co/m4SZwckS5p"" from tweet 201
URL: http://t.co/XEkd4gvAIP
Title: "JennLemon on Twitter: "@lp101677 bad mood, huh? http://t.co/XEkd4gvAIP"" from tweet 202
URL: http://t.co/6npxLIY7S7
Title: "Instagram" from tweet 209
URL: http://t.co/aPOQARzei3
Title: "Assistant Director, Advanced Analytics Jobs in Boston at Liberty Mutual" from tweet 211
URL: http://t.co/B4pSnlwSiP
Title: "mike harmon on Twitter: "@l00clubcentex #chilicookoff #greatday come out and support CTX 100 club #joscoffee http://t.co/B4pSnlwSiP"" from tweet 216
URL: http://t.co/zM8Ll95yCq
Title: "Lynd$i Brooke on Twitter: "Selfie sunday http://t.co/zM8Ll95yCq"" from tweet 220
URL: http://t.co/LdYJ9wfCmQ
Title: "Instagram" from tweet 223
URL: http://t.co/1IvrJpZpQI
Title: "BOOM -- Berkeley Orders of Magnitude -- Declarative Languages And Systems" from tweet 225
URL: http://t.co/FPEfJnEvja
Title: "Beauty Salons provide Beauty Treatments in Braintree, Essex - Beauty Blog, Braintree" from tweet 229
Ubuntu Software Center
Title: "NICKIAS FLORANG on Twitter: "Love this girl so much elenaaa__ http://t.co/5B1X7qjc5y"" from tweet 233
URL: http://t.co/aiNcvn5WAy
Title: "Chloe Murphy on Twitter: "Skydiving was AWESOME ickappice http://t.co/aiNcvn5WAy"" from tweet 239
URL: http://t.co/8x0ZTtD2BW
Title: "albino ibarra on Twitter: "breakfast almost ready.. http://t.co/8x0ZTtD2BW"" from tweet 245
URL: http://t.co/2qqYJKHVfd
Title: "Instagram" from tweet 246
URL: http://t.co/Dq81932m6i
Title: "Instagram" from tweet 247
URL: https://t.co/igoUqSVuez
Title: "Kenlee | VIP Lounge & Mexican Restaurant" from tweet 251
URL: http://t.co/U6qJ7mcrae
Title: "AVD on Twitter: "Lider :)) http://t.co/U6qJ7mcrae"" from tweet 252
URL: http://t.co/YiFic97qSm
Title: "NOBLE on Twitter: "Supertrunks http://t.co/YiFic97qSm"" from tweet 253
URL: http://t.co/raJeolBmvL
Title: "Instagram" from tweet 254
URL: http://t.co/oOQT0zz8j0
Title: "Katie on Twitter: "So I think I've just broke a heart http://t.co/oOQT0zz8j0"" from tweet 257
URL: http://t.co/SmepUux90G
Title: "azharscha on Twitter: "Allahuakbar , may Allah bless this guy , amin http://t.co/SmepUux90G"" from tweet 258
URL: http://t.co/HlAQZmFp34
Title: "02.27.95 | PreySongz on Twitter: "Gone but never forgotten. MJ is the GOAT in the music industry. #RIP #GRAMMYs http://t.co/HlAQZmFp34"" from tweet 261
URL: http://t.co/j0JGPj73f7
Title: "Queerty on Twitter: "Five Easy Yoga Poses To Make Power Bottoming More Enjoyable http://t.co/eJyG2l5vKa http://t.co/3sEyWTyv5f"" from tweet 264
URL: http://t.co/74pcMnDIZg
Title: "AVON Representative Log in | yourAVON" from tweet 268
URL: https://t.co/vN4TLcQ5ca
Title: "Alvarinho | Six Flags" from tweet 275
URL: http://t.co/3kDJiCBvte
Title: "Instagram" from tweet 280
URL: http://t.co/yl4fmfKbXZ
Title: "no on Twitter: "50 shades of deleted scene http://t.co/qDKJxPmooe"" from tweet 282
URL: http://t.co/DY9WtCeYJq