

Part 2 - Indexer

Overview of system

The index to store the retrieved tweets is created through ElasticSearch. ``set_index.py`` is a file that initializes the setting of the index. ``insert_index.py`` reads through the ``fetched_tweets.json``, analyzes each json in the file, and inserts it into ElasticSeach.

In the process of parsing, selected objects like [tweet id], [screen name], [location], [text], and [hashtags]. Text retrieval is based if the tweet is truncated because of the 140-character limit.

If a tweet contains any hashtags, it is included in the index stored as an array. If a tweet contains URLs, a function is called to parse the link and store the title of the tweet. Because tweets can contain URLs that link to images, titles are stored based on the elements stored in [entities: urls].

To maintain a sizeable index, the index should be deleted and re-created every time ``twitter_streaming.py`` has run.

Limitations of system

- The index only stores [tweet id], [screen name], [location], [text] from tweet, [timestamp] of tweet, [hashtags], and [urlTitles] of tweets rather than storing all the data from the json.

Instructions on how to deploy the indexer.

While ElasticSearch is running, to insert an already existing .json into the index, run:
`json_index.sh`

.json must be in data/ directory.

Part 3 - Extension

As an extension to the parts above, I created a web-based interface using Flask, a python web-development framework. This interface allows the user to see all the tweets retrieved from the index in the first page. Upon querying, the resulting page shows tweets relevant to the user's input and a score on the right side. Scoring, sorting, and ranking are all done with ElasticSeach. Scoring is the sum of all fields relevant to query.

Limitations of system

- This is a very, very basic UI.
- Displayed tweet attributes are [screen name], [location], [text], and [timestamp]. [score] is also displayed to the right, upon query.
- Searching helpers like "query" and boolean AND OR NOT are not implemented.
- Not category specific
- Some issues that arose are:
 - Querying emojis: this UI doesn't recognize emoji's
 - Doesn't show the words queried before listing results

Instructions on how to deploy the system.

// command to start docker and run .sh , in that order

Open <http://0.0.0.0:8005> in browser of choice.

Screenshots showing the system in action.

The first screenshot shows the 'The Tweets' page. It has a search bar with the text 'Search for a tweet...' and a 'Search' button. Below the search bar, it says 'We have 267 tweets pulled from San Francisco!'. The page lists several tweets with their authors, locations, and timestamps. The second screenshot shows the 'The Queried Tweets' page. It has a search bar with the text 'This is a story all about' and a 'Search' button. Below the search bar, it says 'We have 91 tweets pulled from the index!'. The page lists several tweets with their authors, locations, and timestamps. On the right side of the page, there is a column of scores for each tweet.

Twitter Search Engine

Not Secure | 0.0.0.0:8005

We have 267 tweets pulled from San Francisco!

The Tweets

Search for a tweet... Search

@Lizhnf650 South San Francisco, CA

Happy 22nd Birthday 🍰 to my beautiful daughter, Lily! I'm so proud to be your Mom. Words can't express what you mean to me. Continue working hard to achieve your dreams, persevere like... <https://t.co/vRPP02sc4>
Wed Mar 20 08:23:39 +0000 2019

@PushinWeight California, USA

@savvy_tck @DexDubbs A whole ass mood. 🍑 🍑 🍑
Wed Mar 20 08:24:33 +0000 2019

@JUSTBYRDD San Leandro, CA

@ashleytaylur even though u RT it thought I should tag u 🤔
Wed Mar 20 08:25:56 +0000 2019

@MightyMolasses Sunnyvale, CA

Early morning (I am) #WhiskersWednesday Back to sleep now says 🐾 🐾 🐾 🐾
Wed Mar 20 08:26:16 +0000 2019

Twitter Search Engine

Not Secure | 0.0.0.0:8005

We have 91 tweets pulled from the index!

The Queried Tweets

This is a story all about Search

@little_brenna San Jose, CA

i'm about to be awake during crackhead hours and my dumbass is about to online shop for the second night in a row cause i'm bored
Wed Mar 20 08:55:29 +0000 2019

@mxanthropology Sunnyvale, CA

Hey. It's probably a good idea to be a slight bit more studious about that economist tweet before going off the handle about it. A. It was a stupid fuck tweet. B. They agree and deleted it. C. It didnt represent the article well at all.
Wed Mar 20 08:22:53 +0000 2019

@ValerieCorey1 California, USA

OMG! This is too funny NOT to retweet! #PoorBabyDevinNunes
Wed Mar 20 09:00:08 +0000 2019

@arrrrrrianna California, USA

if there's ever been a tweet that IS ME
Wed Mar 20 08:44:35 +0000 2019

@flawlessbaguete California, USA

This one took a few hours lmao
Wed Mar 20 08:43:42 +0000 2019

@Cvssius_Clv San Jose, CA

@_jeandoe 🐼 🐼 🐼 Ong but all I smell is dro cuz I den definitely rolled up while da load spinning
Wed Mar 20 08:55:50 +0000 2019

6.5164022

5.3856945

5.2695036

4.7004633

4.699002

4.5907607

Home

Part 4 - Other remarks

In the future, I would like to improve upon this system by:

- improving and expanding from the limitations stated in previous parts.
 - include a category/field option for user querying.
 - improve on UI (prettying up the front end, etc.).
 - allow user input on how many tweets to retrieve in crawler.