Library6Btweets
Final Presentation

By: Pranav Dhakal, Yash Bhargava, Daniel Burdisso, Kenneth Powell, Anna Herms

Edward A. Fox
CS 4624: Multimedia, Hypertext, and Information Access
Virginia Tech, Blacksburg VA 24061
Date: 12.07.2021
Project Overview

Our goal is to consolidate over 6 billion tweets from the Digital Library Research Laboratory (DLRL) into a unified schema. Currently there are three types of collection systems that are storing the tweet data:

- yourTwapperKeeper (YTK)
- Social Feed Manager (SFM)
- Digital Methods Initiative Twitter Capture and Analysis Toolset (DMI-TCAT)

Technology utilized: Python, MySQL database, JSON objects
Timeline

**Sept.**
- **Milestones 1 & 2**
  - Establish team roles
  - Meet with client
  - Establish requirements and milestones
  - Complete Python scripts for reading in YTK and SFM data

**Oct.**
- **Milestones 3 & 4**
  - Complete script for reading in DMI-TCAT data
  - Begin prototyping for JSON collection level schema
  - Finalize JSON unified tweet master schema

**Nov.**
- **Milestones 5-8**
  - Finalize JSON collection-level schema
  - Complete script(s) for converting old data to new schema
  - Test script for converting tweet data to the JSON master schema
Work Completed: Summary

- Tweet converter scripts
  - SFM individual tweet converter
  - SFM collection-level converter
  - YTK individual tweet converter
  - YTK collection-level converter
  - DMI-TCAT individual tweet converter
  - DMI-TCAT collection-level converter
- Testing scripts
  - Unit tests
  - Tweet data validators
- The next slides will discuss some of this work
## Finalized Individual Tweet Schema

<table>
<thead>
<tr>
<th>SFM</th>
<th>DMI-TCAT</th>
<th>YTK</th>
<th>New Schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>contributors</td>
<td>contributors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>created_at</td>
<td>created_at</td>
<td>created_at</td>
<td></td>
</tr>
<tr>
<td>display_text_range[0]</td>
<td>display_text_range.start</td>
<td></td>
<td></td>
</tr>
<tr>
<td>display_text_range[1]</td>
<td>display_text_range.end</td>
<td></td>
<td></td>
</tr>
<tr>
<td>entities.hashtags</td>
<td>entities.hashtags</td>
<td></td>
<td></td>
</tr>
<tr>
<td>entities.media</td>
<td>entities.media</td>
<td></td>
<td></td>
</tr>
<tr>
<td>entities.media.url</td>
<td>entities.media.url</td>
<td></td>
<td></td>
</tr>
<tr>
<td>entities.media.id_str</td>
<td>entities.media.id</td>
<td></td>
<td></td>
</tr>
<tr>
<td>entities.urls (NOTE: actually a multi-element dict/struct)</td>
<td>entities.urls</td>
<td></td>
<td></td>
</tr>
<tr>
<td>entities.user_mentions</td>
<td>entities.user_mentions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>place.country</td>
<td>geo.country</td>
<td></td>
<td></td>
</tr>
<tr>
<td>geo.coordinates[1]</td>
<td>geo.lat</td>
<td>geo.coordinates_1</td>
<td>geo.latitude</td>
</tr>
<tr>
<td>geo.coordinates[2]</td>
<td>geo.lng</td>
<td>geo.coordinates_0</td>
<td>geo.longitude</td>
</tr>
<tr>
<td>geo.type</td>
<td>geo.type</td>
<td>geo.type</td>
<td></td>
</tr>
<tr>
<td>id_str</td>
<td>id (BigInteger)</td>
<td>id</td>
<td>id</td>
</tr>
<tr>
<td>in_reply_to_screen_name</td>
<td>in_reply_to_screen_name</td>
<td></td>
<td></td>
</tr>
<tr>
<td>in_reply_to_user_id</td>
<td>in_reply_to_user_id</td>
<td></td>
<td></td>
</tr>
<tr>
<td>in_reply_to_status_id</td>
<td>in_reply_to_status_id</td>
<td></td>
<td></td>
</tr>
<tr>
<td>is_quote_status</td>
<td>is_quote_status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lang</td>
<td>lang</td>
<td>lang</td>
<td></td>
</tr>
<tr>
<td>metadata.iso_language_code</td>
<td>iso_language_code</td>
<td>lang.iso_code</td>
<td></td>
</tr>
<tr>
<td>metadata.result_type</td>
<td>result_type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>favorite_count (is this for the tweet or user?)</td>
<td>metrics.favorite_count</td>
<td></td>
<td></td>
</tr>
<tr>
<td>retweet_count</td>
<td>metrics.retweet_count</td>
<td></td>
<td></td>
</tr>
<tr>
<td>source (HTML tag - will have to extract the ex source)</td>
<td>source</td>
<td></td>
<td></td>
</tr>
<tr>
<td>full_text</td>
<td>Text</td>
<td>text</td>
<td></td>
</tr>
<tr>
<td>user</td>
<td>user</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Finalized Collection-Level Schema**

<table>
<thead>
<tr>
<th>Final schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
</tr>
<tr>
<td>description</td>
</tr>
<tr>
<td>count</td>
</tr>
<tr>
<td>tweet_ids</td>
</tr>
<tr>
<td>collection_terms</td>
</tr>
<tr>
<td>wikipedia</td>
</tr>
<tr>
<td>create_time</td>
</tr>
<tr>
<td>metrics.retweet_count</td>
</tr>
<tr>
<td>metrics.like_count</td>
</tr>
<tr>
<td>metric.reply_count</td>
</tr>
<tr>
<td>metric.quote_count</td>
</tr>
</tbody>
</table>
Work Completed: SFM Individual Tweet Converter

Start of SFM Tweet Object

Start of New Schema Tweet Object
Work Completed: YTK Individual Tweet Converter

Sample tables in YTK

<table>
<thead>
<tr>
<th>Tables_in_twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>archives</td>
</tr>
<tr>
<td>processes</td>
</tr>
<tr>
<td>rawstream</td>
</tr>
<tr>
<td>z_1</td>
</tr>
<tr>
<td>z_10</td>
</tr>
<tr>
<td>z_11</td>
</tr>
<tr>
<td>z_12</td>
</tr>
<tr>
<td>z_13</td>
</tr>
<tr>
<td>z_14</td>
</tr>
<tr>
<td>z_15</td>
</tr>
<tr>
<td>z_2</td>
</tr>
<tr>
<td>z_6</td>
</tr>
<tr>
<td>z_7</td>
</tr>
<tr>
<td>z_9</td>
</tr>
</tbody>
</table>

14 rows in set (0.01 sec)

JSON output

- table_z_1.json
- table_z_2.json
- table_z_6.json
- table_z_7.json
- table_z_9.json
- table_z_10.json
- table_z_11.json
- table_z_12.json
- table_z_13.json
- table_z_14.json
- table_z_15.json

Connection was successful
Fetching all table names (except archives, processes, and rawstream)
table: z_1 # of tweets: 208978
table: z_10 # of tweets: 17129
table: z_11 # of tweets: 195
table: z_12 # of tweets: 11619
table: z_13 # of tweets: 739
table: z_14 # of tweets: 94167
table: z_15 # of tweets: 2
table: z_2 # of tweets: 172240
table: z_6 # of tweets: 5096
table: z_7 # of tweets: 32060
table: z_9 # of tweets: 207827
Successfully converted data to JSON!
Total time taken to convert 750052 lines is 0:01:12.162199

Running the script
Work Completed: YTK Collection Tweet Convertor

Sample Tables in YTK

Archives Table Sample

Partial Example of YTK Collection Output
Work Completed: DMI-TCAT

Sample tables in MySQL

```
mysql> use twitter_data_test
Database changed
mysql> show tables;

+------------------+
| Tables_in_twitter_data_test |
| budget2015_hashtags |
| budget2015_media    |
| budget2015_mentions |
| budget2015_places  |
| budget2015_tweets  |
| budget2015_urls    |
| budget2015_withheld |
+------------------+
```

Query code example

```python
$ python dmi_converter.py tweet.json
Username: root
Password:
Host: localhost
Database: twitter_data_test
Connection was successful
Converting data into schema...
Total time taken to convert 100 tweets is 0:01:05.423917
```
Work Completed: SFM Converter Unit Tests

```python
def test_sample_3(self):
    in_datum = {
        "full_text": "RT @slogesh: Sir my name is logeshwaran and I'm passed ",
        "lang": "en",
        "retweet_count": 10000,
        "metadata": {
            "iso_language_code": "en",
            "result_type": "recent"
        },
        "source": "<a href="http://twitter.com/download/android" rel="noreferrer">
    }
    out_datum = {
        "created_at": None,
        "id": None,
        "text": "RT @slogesh: Sir my name is Logeshwaran and I'm passed on",
        "entities": None,
        "lang": "en",
        "lang_iso_code": "en",
        "source": "<a href="http://twitter.com/download/android" rel="noreferrer">
    }
    self.assertEqual(out_datum, convert(in_datum))
```

$ python test/test_convert_sfm.py

..............
Ran 15 tests in 0.002s

OK

Test case example

How to execute the tests
Validator Scripts

- Used for testing
- Two scripts:
  - Validator for individual tweet data
  - Validator for collection-level data
- Sanity checks all data in a file
Future Work

- Performance Improvement of Scripts - Time taken to convert tweets
- Converting 6 billion tweets - Can take significant amount of time
- Link SFM and DMI-TCAT to Events Archives collection table
References

- Twitter API Reference
- Twitter V2 Data Dictionary
- Tweet Collection Management Report
- Events Archiving
  - http://eventsarchive.org/
Acknowledgements

We would like to thank our client, Xinyue Wang, and Dr. Fox for their help on this project.
Any Questions?