Collection Management Tweet

CS5604, Information Storage & Retrieval, Fall 2017

Farnaz Khaghani    Junkai Zeng    Momen Bhuiyan
Anika Tabassum    Payel Bandyopadhyay

Professor: Dr. Edward Fox
Purpose of CMT

● Processing Tweets of two events:
  ○ Solar Eclipse (6M Tweets)
  ○ Las Vegas Shooting (~0.18M tweets)

● Creating a social network database based on the Twitter users and tweets relationships
Tweet Processing Overview

- JSON
- Stanford NLP
  - Website list
  - Apache Spark library
- Mathew’s Framework
- HDFS
- Tweet cleaning (stop-word and profanity removal)
- Entity recognition
- POS Tagging
- Parser

- HBase
  - User ID
  - Tweet ID
  - ...!
  - Cleaned Tweets
    - POS
    - NER
  - HashTags
    - Mentions
    - Retweet flags
Previous Arch.: JSON to HBase
Current Arch.: JSON to HBase
Parsing

- json4s: a json library in scala
- For Las Vegas Shooting dataset (~180k tweet), the parsing took less than 2mins
- Changes:
  - Removal of Multiple Steps: Minimize Data Pre-Processing
  - Overhead: Copying the json file
Cleaning

- Data cleaning
  - NER, POS, Tokenization, Lemmatization: Stanford CoreNLP
  - Hashtag, Mentions, Retweet: Matthew’s Framework
  - Stopword Removal: Spark ML lib
  - Cleaning Punctuation, Removing Profanity, Formatting: Scala Code
- For Las Vegas shooting dataset, data cleaning took less than 2 hour
<table>
<thead>
<tr>
<th>Column Family</th>
<th>Column-name</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean-tweet</td>
<td>NER</td>
<td>Shooting a Chrome .50 Cal Machine Gun on the Vegas Strip #lasvegas #vegas #shooting #SaturdayMotivation <a href="https://t.co/ZroMarY7un">https://t.co/ZroMarY7un</a></td>
</tr>
<tr>
<td>clean-tweet</td>
<td>POS</td>
<td>&lt;em&gt;RT&lt;/em&gt; &lt;em&gt;@troyglidden&lt;/em&gt;: &lt;em&gt;Scanner&lt;/em&gt; ...</td>
</tr>
<tr>
<td>clean-tweet</td>
<td>clean-text-cla</td>
<td>security guard shot leg 32nd floor unk hotel vegas shooting</td>
</tr>
<tr>
<td>clean-tweet</td>
<td>clean-text-cta</td>
<td>security guard shot leg 32nd floor unk hotel vegas shooting</td>
</tr>
<tr>
<td>clean-tweet</td>
<td>clean-text-solr</td>
<td>security guard shot leg 32nd floor unk hotel vegas shooting</td>
</tr>
<tr>
<td>clean-tweet</td>
<td>clean-tokens</td>
<td>shooting;chrome;50;cal;machine;gun;vega;strip;lasvegas ;vega;shooting;saturdaymotivation;</td>
</tr>
<tr>
<td>clean-tweet</td>
<td></td>
<td>Aug 21, 2017</td>
</tr>
</tbody>
</table>
## Schemas Provided in HBase

<table>
<thead>
<tr>
<th>Column Family</th>
<th>Column Name</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean-tweet</td>
<td>geom-type</td>
<td></td>
</tr>
<tr>
<td>clean-tweet</td>
<td>hashtags</td>
<td>#lasvegas,#vegas,#shooting,#SaturdayMotivation</td>
</tr>
<tr>
<td>clean-tweet</td>
<td>mentions</td>
<td>troyglidden</td>
</tr>
<tr>
<td>clean-tweet</td>
<td>rt</td>
<td>false</td>
</tr>
<tr>
<td>clean-tweet</td>
<td>sner-locations</td>
<td>Vegas;Strip;</td>
</tr>
<tr>
<td>clean-tweet</td>
<td>sner-organizations</td>
<td></td>
</tr>
<tr>
<td>clean-tweet</td>
<td>sner-people</td>
<td></td>
</tr>
<tr>
<td>clean-tweet</td>
<td>solr-gemo</td>
<td></td>
</tr>
</tbody>
</table>
### Schemas Provided in HBase

<table>
<thead>
<tr>
<th>Column Family</th>
<th>Column Name</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean-tweet</td>
<td>spatial-bounding</td>
<td></td>
</tr>
<tr>
<td>clean-tweet</td>
<td>spatial-coord</td>
<td></td>
</tr>
<tr>
<td>clean-tweet</td>
<td>tweet-importance</td>
<td></td>
</tr>
<tr>
<td>clean-tweet</td>
<td>url_visited_cmw</td>
<td></td>
</tr>
<tr>
<td>metadata</td>
<td>collection-id</td>
<td>1024</td>
</tr>
<tr>
<td>metadata</td>
<td>collection-name</td>
<td>#shooting #LasVegas</td>
</tr>
<tr>
<td>metadata</td>
<td>doc-type</td>
<td>tweet</td>
</tr>
<tr>
<td>metadata</td>
<td>dummy-data</td>
<td>false</td>
</tr>
</tbody>
</table>
## Schemas Provided in HBase

<table>
<thead>
<tr>
<th>Column Family</th>
<th>Column Name</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>tweet</td>
<td>archive-source</td>
<td>twitter-search</td>
</tr>
<tr>
<td>tweet</td>
<td>comment-count</td>
<td>-1</td>
</tr>
<tr>
<td>tweet</td>
<td>contributor-enabled</td>
<td>false</td>
</tr>
<tr>
<td>tweet</td>
<td>created-time</td>
<td>Sat Sep 23 20:08:16 +0000 2017</td>
</tr>
<tr>
<td>tweet</td>
<td>created-timestamp</td>
<td></td>
</tr>
<tr>
<td>tweet</td>
<td>geo-0</td>
<td></td>
</tr>
<tr>
<td>tweet</td>
<td>geo-1</td>
<td></td>
</tr>
<tr>
<td>tweet</td>
<td>geo-type</td>
<td></td>
</tr>
<tr>
<td>tweet</td>
<td>language</td>
<td>en</td>
</tr>
</tbody>
</table>
# Schemas Provided in HBase

<table>
<thead>
<tr>
<th>Column Family</th>
<th>Column Name</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>tweet</td>
<td>like-count</td>
<td>5</td>
</tr>
<tr>
<td>tweet</td>
<td>place-country-code</td>
<td>US</td>
</tr>
<tr>
<td>tweet</td>
<td>profile-img-url</td>
<td><a href="http://pbs.twimg.com/profile_images/894753143057137666/3U9Y6Di2_normal.jpg">http://pbs.twimg.com/profile_images/894753143057137666/3U9Y6Di2_normal.jpg</a></td>
</tr>
<tr>
<td>tweet</td>
<td>retweet-count</td>
<td>1</td>
</tr>
<tr>
<td>tweet</td>
<td>screen-name</td>
<td>pepesgrandma</td>
</tr>
<tr>
<td>tweet</td>
<td>source</td>
<td>&lt;a href=&quot;<a href="http://twitter.com">http://twitter.com</a>&quot; rel=&quot;nofollow&quot;&gt;Twitter Web Client&lt;/a&gt;</td>
</tr>
<tr>
<td>tweet</td>
<td>text</td>
<td>Shooting a Chrome .50 Cal Machine Gun on the Vegas Strip \texttt{\textcopyright}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>#lasvegas #vegas #shooting #vegas #SaturdayMotivation <a href="https://t.co/ZroMarY7un">https://t.co/ZroMarY7un</a></td>
</tr>
<tr>
<td>tweet</td>
<td>to-user-id</td>
<td>12</td>
</tr>
<tr>
<td>Column Family</td>
<td>Column Name</td>
<td>Example</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>tweet</td>
<td>tweet-deleted</td>
<td>false</td>
</tr>
<tr>
<td>tweet</td>
<td>tweet-id</td>
<td>911683653868113920</td>
</tr>
<tr>
<td>tweet</td>
<td>url</td>
<td><a href="https://t.co/ZroMarY7un">https://t.co/ZroMarY7un</a></td>
</tr>
<tr>
<td>tweet</td>
<td>user-deleted</td>
<td>false</td>
</tr>
<tr>
<td>tweet</td>
<td>user-id</td>
<td>116384038</td>
</tr>
<tr>
<td>tweet</td>
<td>user-name</td>
<td>Babushka\xE5\xA5\xB3\xE5\xA3\xAB</td>
</tr>
<tr>
<td>tweet</td>
<td>user_favourites_count</td>
<td>42111</td>
</tr>
<tr>
<td>tweet</td>
<td>user_followers_count</td>
<td>5569</td>
</tr>
<tr>
<td>tweet</td>
<td>user_friends_count</td>
<td>357</td>
</tr>
<tr>
<td>tweet</td>
<td>user_lang</td>
<td>en</td>
</tr>
<tr>
<td>tweet</td>
<td>user_location</td>
<td>Siberia China</td>
</tr>
<tr>
<td>tweet</td>
<td>user_mentions_id_str</td>
<td>Dahboo7</td>
</tr>
<tr>
<td>tweet</td>
<td>user_mentions_name</td>
<td>1411455757</td>
</tr>
<tr>
<td>tweet</td>
<td>user_statuses_count</td>
<td>31996</td>
</tr>
</tbody>
</table>
Social Network
Overview
"favorite_count": 0,
"full_text": "There's going to be a #totaleclipse on #august21, but you'll only totally see it if you live in: https://t.co/mZpS0NYxac",
"entities": {
  "symbols": [],
  "user_mentions": [],
  "hashtags": [
    {
      "indices": [30],
      "text": "totaleclipse"
    }
  ],
  "urls": [
    {
      "url": "https://t.co/mZpS0NYxac",
      "expanded_url": "https://www.instagram.com/p/BXqE86O2aZ/",
      "display_url": "instagram.com/p/BXqE86O2aZ/"
    }
  ]
},
"retweeted": false,
"coordinates": null,
"source": "<a href="http://instagram.com" rel="nofollow">Instagram</a>",
"in_reply_to_screen_name": null,
"in_reply_to_status_id": null,
"display_text_range": [0, 0],
"retweet_count": 0,
"id_str": "892447657783853000",
"favorited": false,
"user": {
  "follow_request_sent": false,
  "has_extended_profile": false,
  "profile_use_background_image": true,
  "default_profile_image": false,
  "id": 17481588,
  "verified": false,
Pre-processing data for social network

- Using shell scripts for pre-processing the data
- Converting the tweets from JSON to CSV format
- Created a full CSV file with all fields
Challenges of working with JSON file

- Difficult to interpret → JSON formatter
- Large files to process
- Inconsistency in the fields
Commands to convert JSON to CSV

- Used the “jq” library

- Sample usage:

  ```
  cat Eclipse.json | jq -r '. | [.user.id_str, .retweeted_status.id_str, .in_reply_to_user_id,
  .entities.user_mentions[].id] | @csv' > ./Eclipse/Eclipse.csv
  ```

  - The above didn’t worked when there were more than 2 fields having array elements.
  - For those cases, we processed the fields separately, then separated them using semi-colon, “;”, and then merged the files
<table>
<thead>
<tr>
<th>id</th>
<th>favourite_count</th>
<th>full_text</th>
<th>user_id</th>
<th>retweeted_status_id</th>
<th>in_reply_to_user_id</th>
<th>entities_user_mentions</th>
</tr>
</thead>
<tbody>
<tr>
<td>888201064817860613</td>
<td>5</td>
<td>There's going to be a ...</td>
<td>103167711</td>
<td>889882842242707456</td>
<td>15102849</td>
<td>713741422000807937</td>
</tr>
<tr>
<td>19199743</td>
<td>2</td>
<td>I gotta buy some solar eclipse ...</td>
<td>264792278</td>
<td>889941327202455553</td>
<td>125485258</td>
<td>2470058834</td>
</tr>
<tr>
<td>2762027475</td>
<td>0</td>
<td>Cellphone service could be spotty ...</td>
<td>466665274</td>
<td>889874789611048960</td>
<td>15102849</td>
<td>124197346</td>
</tr>
<tr>
<td>224233529</td>
<td>0</td>
<td>Anyone else notice how ...</td>
<td>101144034</td>
<td>889898800411688960</td>
<td>11348282</td>
<td>11348282</td>
</tr>
</tbody>
</table>
Social Network

Objective:
Build a social network to connect the tweets and users relationship

Nodes: 1) Users
2) Tweets

Edges: Existence of the relationship
• Retweet
• Mention
• In reply to
RDF (Resource Description Framework) triplestore is a graph database for storing semantic facts:
- Formally describes the semantics, or meaning, of information
- Represents metadata
- Consists of triples which are based on an Entity-Attribute Value (EAV) model

Selena Gomez follows Coach
What is triplestore?

- Social network is a graph of nodes and edges (Every nodes as a user and edge as a relationship)
- Triplestore stores every node-edge (user-user relationship in simple sentence form)
- Simple sentence: <subject> <predicate> <object>
- Subject: user, predicate: relationship object: user
- We store each user in form of Twitter Ids
Why Triplestore?

- Faster than relational databases
- Support optional schema models, called ontology
- Improve the search and analytics power
- Use of SPARQL Query
Convert CSV to RDF N-Triple File

- Apache Jena Library in Java to convert CSV file to N-Triple (.nt) file

- Apache Jena Fuseki server to store social network (n-triple) data
N Triple file sample

<http://example.org/898620093059534848> \\
\text{Subject: URI of the userID}

<http://xmlns.com/SNR/0.1/mentions> "1021074122" .

\text{Predicate: URI of the predicate}

\text{Object: userID (string)}
Triplestore Database

Datasets on this server

<table>
<thead>
<tr>
<th>dataset name</th>
<th>actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>/eclipse</td>
<td>query, add data, info</td>
</tr>
<tr>
<td>/getar</td>
<td>query, add data, info</td>
</tr>
<tr>
<td>/shooting</td>
<td>query, add data, info</td>
</tr>
</tbody>
</table>

Use the following pages to perform actions or tasks on this server:

- **Dataset**: Run queries and modify datasets hosted by this server.
- **Manage datasets**: Administer the datasets on this server, including adding datasets, uploading data and performing backups.
- **Help**: Summary of commands and links to online documentation.
Triplestore Database

```sql
prefix sub: <http://example.org/>
prefix pred: <http://xmlns.com/SDR/0.1/>

SELECT ?o
WHERE {
  sub:2351245436 pred:mentions|pred:in_reply_to|pred:in_retweet_to ?o
}
```

**QUERY RESULTS**

```
{
  "head": {
    "vars": [ "o" ]
  },
  "results": {
    "bindings": {
      "o": { "type": "literal" , "value": "848515856057479169" }
    }
  }
}
```
Front End Team Interface

Dataset:
Solar Eclipse event : /eclipse
LasVegas Shooting event : /shooting
(Both datasets are Persistent in fuseki server)

URI:
Subject: <http://example.org/>
Predicate: <http://xmlns.com/SNR/0.1/>
Front End Team Interface

Relations:
in_reply_to
mentions
in_retweet_to
followedBy
Sample for fetching query result in JSON:


- Will fetch all mentions, in_reply_to and in_retweet_to ids of user id 2351245436
Time to upload data

- The largest Solar Eclipse file (~373MB) NT file takes ~4 min to upload
- Time to upload whole Solar Eclipse core ~ 12 min
- Time to upload Las Vegas Shooting core ~2 min
Challenges and Future Works

- Fetching Twitter followers, friends takes time, not possible for ~4M users
- Converting directly to n-triple file from JSON
- Parallelizing the conversion to N-Triple
- Storing user names, screen names, followers, friends in social network
- Calculating followers, friends for top $N$ users who have highest number of followers, friends, tweets posted
Acknowledgment

First, we would like to thank Dr. Fox for his constructive comments and guidance during this project. Our thanks are also due to US National Science Foundation for supporting Global Event and Trend Archive Research (GETAR) through IIS-1619028.
Questions?