Collection Management Tweet

CS5604, Information Storage & Retrieval, Fall 2017

Farnaz Khaghani Junkai Zeng Momen Bhuiyan
Anika Tabassum Payel Bandyopadhyay

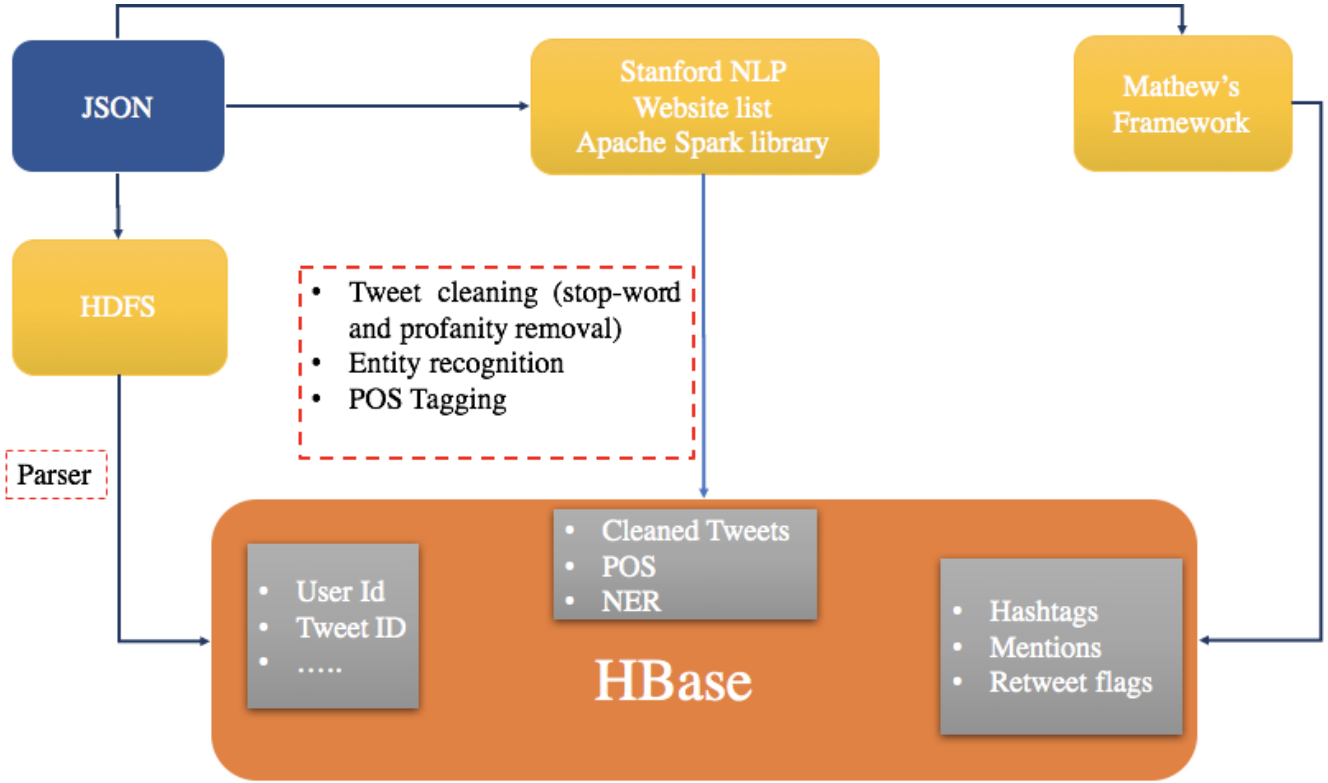
Professor: Dr. Edward Fox

Purpose of CMT

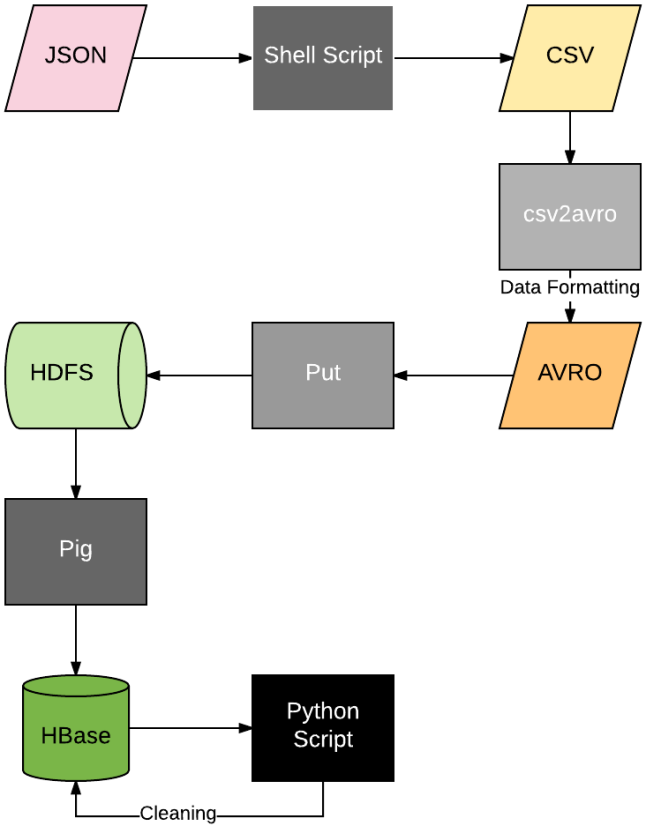
- Processing Tweets of two events:
  - Solar Eclipse (6M Tweets)
  - Las Vegas Shooting (~0.18M tweets)

- Creating a social network database based on the Twitter users and tweets relationships

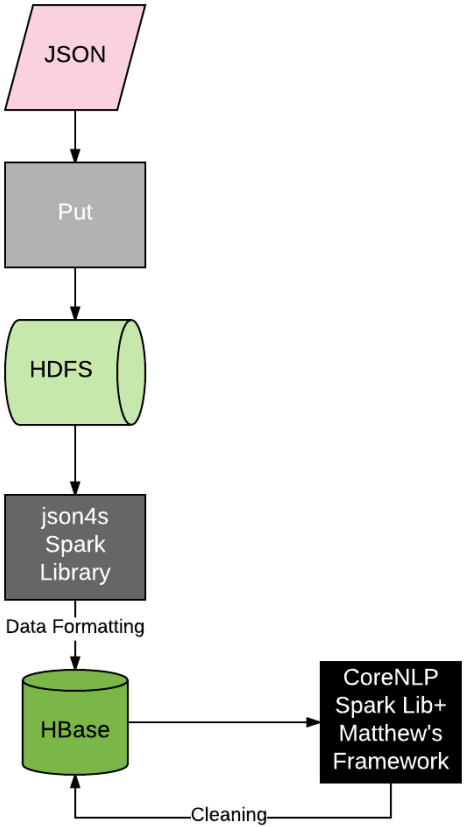
Tweet Processing Overview



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

Schemas Provided in HBase

Column Family	Column-name	Example
clean-tweet	NER	Shooting a Chrome .50 Cal Machine Gun on the Vegas Strip #lasvegas #vegas #shooting #SaturdayMotivation https://t.co/ZroMarY7un
clean-tweet	POS	RT @troyglidden : Scanner ...
clean-tweet	clean-text-cla	security guard shot leg 32nd floor unk hotel vegas shooting
clean-tweet	clean-text-cta	security guard shot leg 32nd floor unk hotel vegas shooting
clean-tweet	clean-text-solr	security guard shot leg 32nd floor unk hotel vegas shooting
clean-tweet	clean-tokens	shooting;chrome;50;cal;machine;gun;vega;strip;lasvega ;vega;shooting;saturdaymotivation;

Schemas Provided in HBase

Column Family	Column Name	Example
clean-tweet	geom-type	
clean-tweet	hashtags	#lasvegas,#vegas,#shooting,#SaturdayMotivation
clean-tweet	long-url	http://freebeacon.com/culture/shooting-a-chrome-50-cal-machine-gun-on-the-vegas-strip/
clean-tweet	mentions	troyglidden
clean-tweet	rt	false
clean-tweet	sner-locations	Vegas;Strip;
clean-tweet	sner-organizations	
clean-tweet	sner-people	
clean-tweet	solr-gemo	

Schemas Provided in HBase

Column Family	Column Name	Example
clean-tweet	spatial-bounding	
clean-tweet	spatial-coord	
clean-tweet	tweet-importance	
clean-tweet	url_visited_cmw	
metadata	collection-id	1024
metadata	collection-name	#shooting #LasVegas
metadata	doc-type	tweet
metadata	dummy-data	false

Schemas Provided in HBase

Column Family	Column Name	Example
tweet	archive-source	twitter-search
tweet	comment-count	-1
tweet	contributor-enabled	false
tweet	created-time	Sat Sep 23 20:08:16 +0000 2017
tweet	created-timestamp	
tweet	geo-0	
tweet	geo-1	
tweet	geo-type	
tweet	language	en

Schemas Provided in HBase

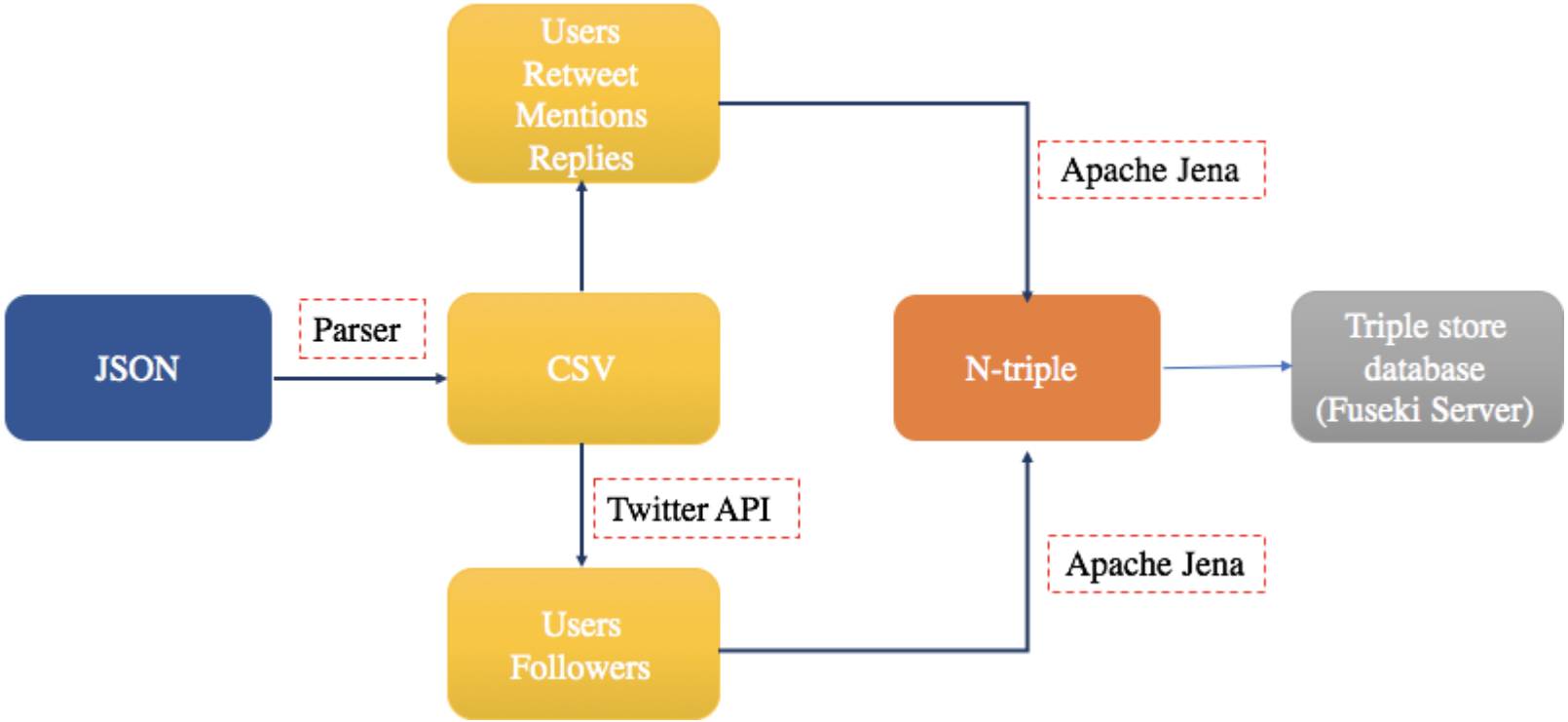
Column Family	Column Name	Example
tweet	like-count	5
tweet	place-country-code	US
tweet	profile-img-url	http://pbs.twimg.com/profile_images/894753143057137666/3U9Y6Di2_normal.jpg
tweet	retweet-count	1
tweet	screen-name	pepesgrandma
tweet	source	Twitter Web Client
tweet	text	Shooting a Chrome .50 Cal Machine Gun on the Vegas Strip \xF0\x9F\x98\x8D\x0A#lasvegas #vegas #shooting #SaturdayMotivation https://t.co/ZroMarY7un
tweet	to-user-id	

Schemas Provided in HBase

Column Family	Column Name	Example
tweet	tweet-deleted	false
tweet	tweet-id	911683653868113920
tweet	url	https://t.co/ZroMarY7un
tweet	user-deleted	false
tweet	user-id	116384038
tweet	user-name	Babushka\xE5\xA5\xB3\xE5\xA3\xAB
tweet	user_favourites_count	42111
tweet	user_followers_count	5569
tweet	user_friends_count	357
tweet	user_lang	en
tweet	user_location	Siberia China
tweet	user_mentions_id_str	Dahboo7
tweet	user_mentions_name	1411455757
tweet	user_statuses_count	31996

Social Network

Overview


Users
Retweet
Mentions
Replies
Apache Jena
JSON
Parser
CSV
N-triple
Triple store
database
(Fuseki Server)
Twitter API
Apache Jena
Users
Followers

Initial Data: JSON


```
"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/mZp50nyXac",
"entities": {
    "symbols": [],
    "user_mentions": [],
    "hashtags": [
        {
            "indices": [...],
            "text": "totaleclipse"
        },
        {...}
    ],
    "urls": [
        {
            "url": "https://t.co/mZp50nyXac",
            "indices": [...],
            "expanded_url": "https://www.instagram.com/p/BXQsEB6D2a2/",
            "display_url": "instagram.com/p/BXQsEB6D2a2/"
        }
    ]
},
"retweeted": false,
"coordinates": null,
"source": "<a href=\"http://instagram.com\" rel=\"nofollow\">Instagram</a>",
"in_reply_to_screen_name": null,
"in_reply_to_user_id": null,
"display_text_range": [...],
"retweet_count": 0,
"id_str": "892447657783853060",
"favorited": false,
"user": {
    "follow_request_sent": false,
    "has_extended_profile": false,
    "profile_use_background_image": true,
    "default_profile_image": false,
    "id": 17401589,
    "profile_background_image_url_https": "https://pbs.twimg.com/profile_background_images/246983128/39-alexander-mcqueen-spring-summer-2005-ss05_-_Copy.JPG",
    "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

Sample pruned CSV file

id	favourite_ count	full_text	user_id	retweeted_status_id	in_reply_to_ user_id	entities_user_mentions
888201064817860613	5	There's going to be a ….	103167711	889882842242707456	15102849	713741422000807937
19199743	2	I gotta buy some solar eclipse ….	264792278	889941327202455553	125485258	2470058834
2762027475	0	Cellphone service could be spotty ….	466665274	889874789611048960	15102849	124197346
224233529	0	Anyone else notice how ….	101144034	889898800411688960	11348282	11348282

Social Network

Objective :

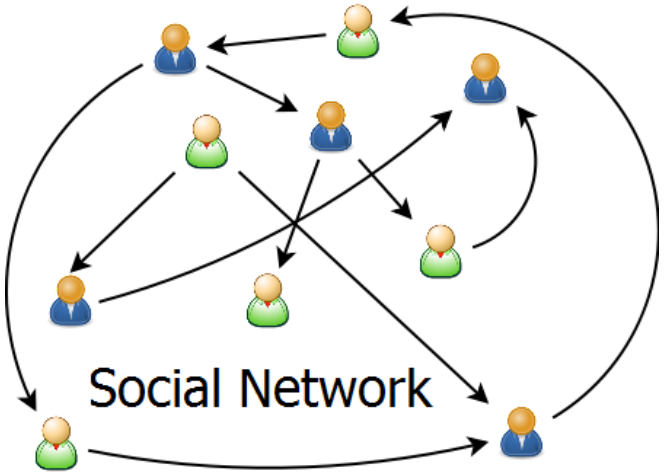
Build a social network to connect the tweets and users relationship

Nodes: 1) Users

2) Tweets

Edges: Existence of the relationship

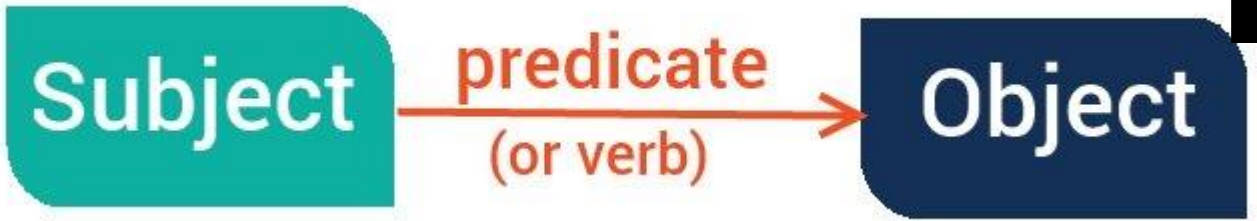
- Retweet
- Mention
- In reply to



RDF triplestore

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/89862009305953484848>

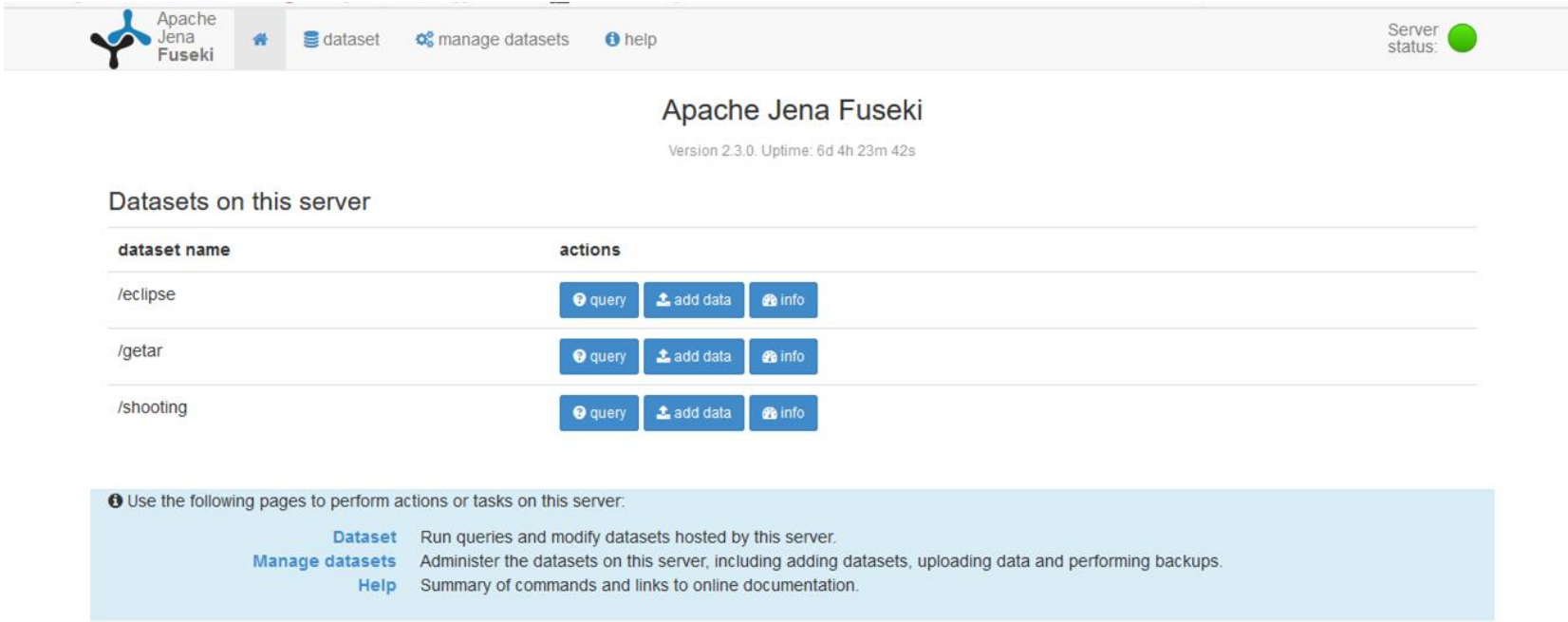
Subject: URI of the userID

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

Predicate: URI of the predicate

Object: userID (string)

Triplestore Database



Triplestore Database

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

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

QUERY RESULTS

Raw Response | Table


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

Front End Team Interface

Sample for fetching query result in JSON:

http://mule.dlib.vt.edu:3030/eclipse/query?query=prefix%20sub:%20%3Chttp://example.org/%3E%20prefix%20pred:%20%3Chttp://xmlns.com/SNR/0.1/%3E%20SELECT%20?y%20WHERE{sub:2351245436%20pred:mentions|pred:in_reply_to|pred:in_retweet_to%20?y.}&wt=json&json.wrf=my_callback

- 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 ~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?