Lab-2
Meeting Project 1 Requirements

Archita Pathak
University at Buffalo

Frequent Concerns

- Why AWS?
  - Mainly for grading
- Why Solr?
  - Provides you resources for efficient indexing of documents
- What is the point of this project?
  - Understanding fundamental IR concepts
  - Understanding tweets and replies retrieval strategies
  - Working with huge volume of data which will be useful in Project 4
- Are we going to grade the code for tweets crawling?
  - No
  - However, checks will be performed on the data. No sharing of data is allowed.

Task 1 – Tweets Crawling

- At least 33,000 tweets in total with not more than 15% being retweets.
- At least 1000 tweets per person of interest.
- At least 20 replies to each of the tweets posted by the POIs for 5 consecutive dates.
- At least 3000 replies in total across all POIs
- At least 5,000 tweets per language i.e, English, Hindi and Portuguese
- At least 5,000 tweets per country
- At least 1000 tweets containing hashtags/keywords related to person of interest

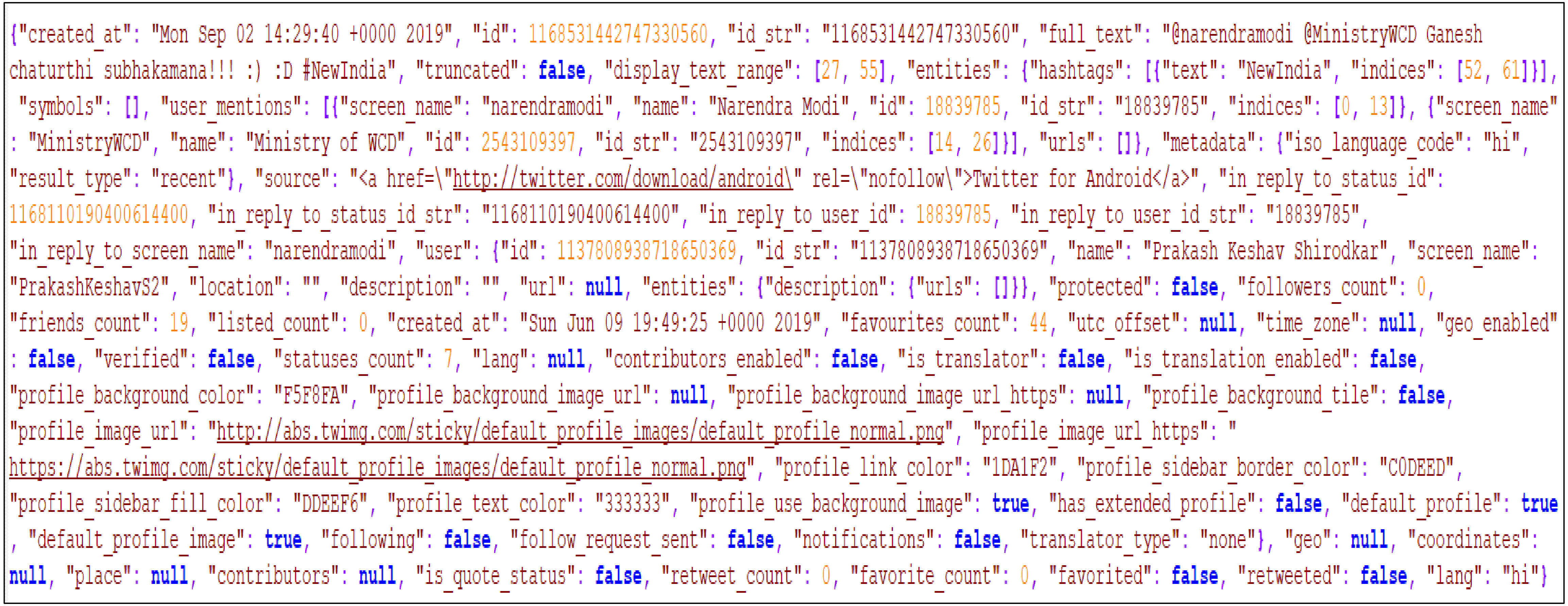
Task 1 - Replies

- Twitter developer policy on Search API:
    - Serves data posted within last 7 days only
    - Strict API rate limiting on the number of tweets you can download or number of requests you can make in a 15 minutes window.
- At least 20 Replies to each tweet posted by POI in 5 consecutive dates
    - Code – Twarc or your own code.
        - Get the list of all tweets containing mentions of the POI
        - Among them, retrieve tweets based on the condition
            - In_reply_to_status_id == tweet_id (where tweet_id is POI’s tweet id)
        - May take long for older tweets (few hours)
        - Max_id and Since_id
    - Daily retrieval
        - Choose 5 days and retrieve daily tweets along with replies for all 15 POIs
        - Scroll through POIs timeline to make sure they have posted something in 5 consecutive dates.

Task 2 – Indexing

- **Solr Field Names**
  1. poi_name : Screen name of one of the 15 persons of interest
  2. poi_id : User Id of one of the 15 persons of interest
  3. verified: Boolean value
  4. country : One of the 3 countries
  5. replied_to_tweet_id : Null for tweet by person of interest else tweet id to which the reply is made.
  6. replied_to_user_id : Null for tweet by person of interest else user id to which the reply is made.
  7. reply_text : Text of the reply to a particular tweet, if replied_to_tweet_id is not null
  8. tweet_text : Default field
  9. tweet_lang : Language of the tweet from Twitter as a two letter code.
  10. text_xx : For language specific fields where xx is at least one amongst en (English), hi (Hindi) and pt (Portuguese)
  11. hashtags : if there are any hashtags within the tweet text
  12. mentions : if there are any mentions within the tweet text
  13. tweet_urls : if there are any urls within the tweet text
  14. tweet_emoticons : if there are any emoticons within the tweet text
  15. tweet_date : Tweet creation date rounded to nearest hour and in GMT
  16. tweet_loc (optional): Geolocation of the tweet. You need to have coordinates in this field.
- Any operation you are doing during indexing, same operations should be done during querying as well.

Task 2 – Indexing Example



```
{"created_at": "Mon Sep 02 14:29:40 +0000 2019", "id": 1168531442747330560, "id_str": "1168531442747330560", "full_text": "@narendramodi @MinistryWCD Ganesh
chaturthi subhakamana!!! :) :D #NewIndia", "truncated": false, "display_text_range": [27, 55], "entities": {"hashtags": [{"text": "NewIndia", "indices": [52, 61]}],
 "symbols": [], "user_mentions": [{"screen_name": "narendramodi", "name": "Narendra Modi", "id": 18839785, "id_str": "18839785", "indices": [0, 13]}, {"screen_name"
: "MinistryWCD", "name": "Ministry of WCD", "id": 2543109397, "id_str": "2543109397", "indices": [14, 26]}], "urls": []}, "metadata": {"iso_language_code": "hi",
"result_type": "recent"}, "source": "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android</a>", "in_reply_to_status_id":
1168110190400614400, "in_reply_to_status_id_str": "1168110190400614400", "in_reply_to_user_id": 18839785, "in_reply_to_user_id_str": "18839785",
"in_reply_to_screen_name": "narendramodi", "user": {"id": 1137808938718650369, "id_str": "1137808938718650369", "name": "Prakash Keshav Shirodkar", "screen_name":
"PrakashKeshavS2", "location": "", "description": "", "url": null, "entities": {"description": {"urls": []}}, "protected": false, "followers_count": 0,
"friends_count": 19, "listed_count": 0, "created_at": "Sun Jun 09 19:49:25 +0000 2019", "favourites_count": 44, "utc_offset": null, "time_zone": null, "geo_enabled"
: false, "verified": false, "statuses_count": 7, "lang": null, "contributors_enabled": false, "is_translator": false, "is_translation_enabled": false,
"profile_background_color": "F5F8FA", "profile_background_image_url": null, "profile_background_image_url_https": null, "profile_background_tile": false,
"profile_image_url": "http://abs.twimg.com/sticky/default_profile_images/default_profile_normal.png", "profile_image_url_https": "
https://abs.twimg.com/sticky/default_profile_images/default_profile_normal.png", "profile_link_color": "1DA1F2", "profile_sidebar_border_color": "C0DEED",
"profile_sidebar_fill_color": "DDEEF6", "profile_text_color": "333333", "profile_use_background_image": true, "has_extended_profile": false, "default_profile": true
, "default_profile_image": true, "following": false, "follow_request_sent": false, "notifications": false, "translator_type": "none"}, "geo": null, "coordinates":
null, "place": null, "contributors": null, "is_quote_status": false, "retweet_count": 0, "favorite_count": 0, "favorited": false, "retweeted": false, "lang": "hi"}
```

Task 2 – Indexing Example Cont.

Field	Values
poi_name	**narendramodi**
poi_id	**18839785**
verified	**false**
country	**India**
replied_to_tweet_id	**1168110190400614400**
replied_to_user_id	**18839785**
reply_text	**@narendramodi @MinistryWCD Ganesh chaturthi subhakamana!!! :) :D #NewIndia**
tweet_text (default)	**@narendramodi @MinistryWCD Ganesh chaturthi subhakamana!!! :) :D #NewIndia**
tweet_lang	**hi**
text_hi	**Ganesh chaturthi subhakamana**
hashtags	**NewIndia**
mentions	**narendramodi, MinistryWCD**
tweet_urls	**none**
tweet_emoticons	**:) :D**
tweet_date	**2019-09-02T15:00:00Z**
tweet_loc (optional)	**none**

Additional Information

- Dry run on 12th Sept. (time will be announced later through Piazza post)
- Grades of dry run will not be counted for final grading
- Grades breakdown has been released in the new version of project requirement document
- After dry run is completed, you will get an email summarizing the grading.
- Note that submission **must be** on time as it’s auto-graded script.