* **Due to Twitter term restrictions, we are not going to share the raw data but the tweet ids for this research purpose shared task.**
* **The downloader script is to download tweet JSON data using the tweet ids in the data folder.**
* **You are also required to register a Twitter dev account to get your credentials and put them as specified in cred.txt. (https://apps.twitter.com/)**
* **The folder also contains train/dev/test data for WNUT 2016 Geotagging Shared Task**

File: train.user.json

Description: Training data for user level tasks

Size: 1M users, 12,827,165 tweets

Format: Each line is a JSON dict which consists of tweet\_id, user\_id, city\_name, city\_latitude, city\_longitude, tweet\_latitude, tweet\_longitude, tweet\_text

Example:

{

"tweet\_id": 12345678,

"user\_id": 12345,

"city\_name": "melbourne-07-au",

"city\_latitude": -35.123,

"city\_longitude": 108.123,

"tweet\_latitude": -35.124,

"tweet\_longitude": 108.124,

"tweet\_text": "This is a demo text"

}

File: validation.user.json

Description: validation data for user level tasks

Size: 10K users, 128,524 tweets

Format: same as train.user.json

Example: same as train.user.json

File: test.user.json

Description: Test data for user level tasks

Size: 10K users, 99,732 tweets

Format: Each line is a JSON dict which consists of hashed\_tweet\_id, hashed\_user\_id, tweet\_text

Example:

{

"hashed\_tweet\_id": 25D55AD283AA400AF464C76D713C07AD,

"hashed\_user\_id": 827CCB0EEA8A706C4C34A16891F84E7B,

"tweet\_text": "This is a demo text"

}

File: train.tweet.json

Description: Training data for tweet level tasks

Size: same as train.user.json

Format: same as train.user.json

Example: same as train.user.json

File: validation.tweet.json

Description: validation data for tweet level tasks

Size: 128,435 tweets which are different from validation.user.json

Format: same as validation.user.json except the user related fields are omitted.

Example: same as validation.user.json except the user related fields are omitted.

File: test.tweet.json

Description: Test data for tweet level tasks

Size: 10K tweets which are different from test.user.json

Format: Each line is a JSON dict which consists of hashed\_tweet\_id, hashed\_user\_id, tweet\_text

Example:

{

"hashed\_tweet\_id": 25D55AD283AA400AF464C76D713C07AD,

"hashed\_user\_id": 827CCB0EEA8A706C4C34A16891F84E7B,

"tweet\_text": "This is a demo text"

}