Improvement of data collection

# 1. Users and Tweets

## 1.1. Collect Australian Twitter users

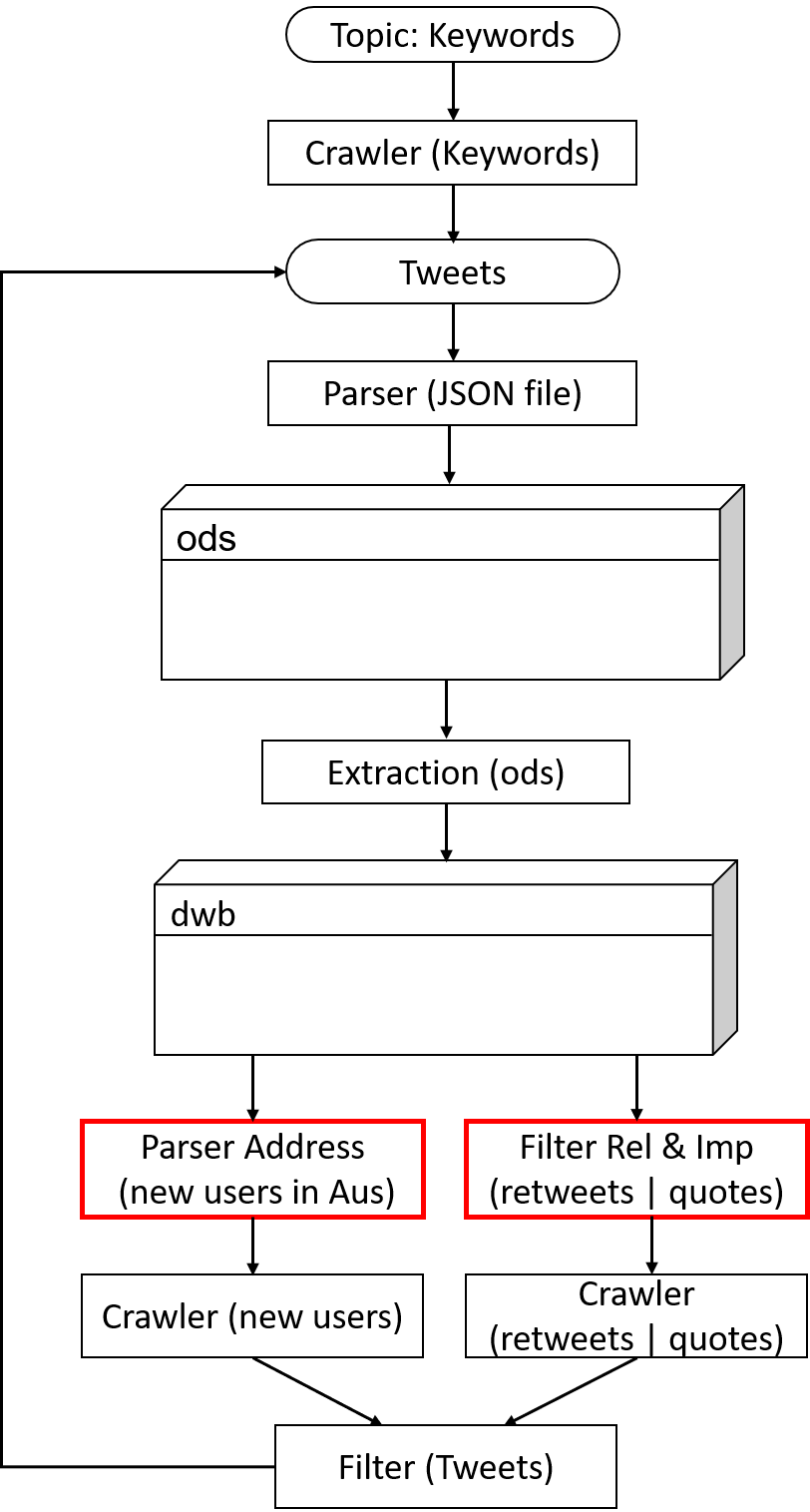
Add module to parse new user, collect the user’s related timeline tweets if the user is from Australia.

## 1.2. Collect related tweets

Check the type of tweet, it is the most important if the type is original, then quote and retweet.

Check whether the tweet exists in Database if the type is retweet and quote. Collect the tweet if it is not in the Database.

## 2.3. Updated diagram



# 2. Truncated tweets

Parse tweets and check whether it is truncated. record the truncated tweets in a table.

Crawler downloads the tweets for downstream analysis.

Truncate table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Status\_id** | **User\_id** | **Screen\_name** | **Type** | **Important** | **URL** | **Related** |
|  |  |  | Origin/RT/RP |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

# 3. Collect replies

Parse and record replies’ ids. David explores the API to collect the replies.

# 4. Resource and time

## 4.1 Main functions

1. Function to parse tweets to add fields of truncated flag and truncated URL
2. Function to optimize parser of user location
3. Function to parse tweet to check original or not
4. Function to parse tweet to check related to topics or not
5. Function to parse key tweets and key users
6. Function to integrate crawler into ETL to improve the performance.

## 4.2 resource

6-person day.