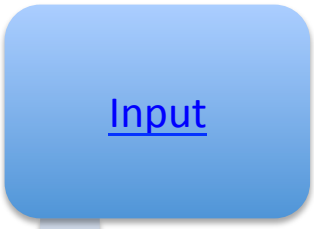
Instant Visualization of twitter data

Team Dahlia:
Meihao Chen, Yitong Wang

Our objectives

- Visualize the massive amount of twitter data (tweets) in a timely fashion
- Ideal for fast information extraction
- Reproducible in Python
- To be utilized by a variety of industries and technical/non-technical professionals

Pipeline: How do we get there?



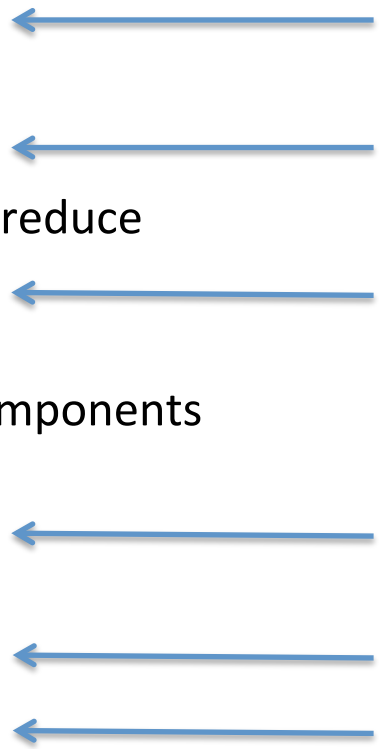
- Twitter data is stored as json.gz files, and usually of large sizes
- Fixed format

- Obtain preliminary results using and **jq** (json field extractor and Map-Reduce/**python streaming**
- Further transformation of data

- D3, NVD3 in JavaScript
- Twitter Bootstrap template

Product: How does it work?

- Folder structure:
  - Dahlia
    - Data
    - Proc
      - ./bashFilter
      - ./countMapreduce
      - ./Proc_d3
    - Vis
      - ./bower_components
      - ./css
      - ./data
      - ./js
      - Index.html
    - run.sh



Input data
intermediate data

Filter out useful fields

Process for visualization,
output goes to Vis/data

Data directly used for visualization

Visualization result in html

Trigger data processing

Extracting information from data

- User information:
  - User name
  - Mentions
  - Statuses count
- Tweet information:
  - Text/Retweeted text
  - Location of the tweet (random sample with coordinates)
  - Source
  - Retweeted user

Visualization

- Draw a Summary of the Dataset:
  - what are people talking about:
    - Word cloud
    - Popular tweets
    - Tweet sequence
  - who are sending tweets:
    - Pie charts on hashtags, geolocation, source, usermentions
    - Mention network
  - Interesting events:
    - Time series

Now take a look at our Demo!

Following Work

- Wrapping up data pipeline
- Interactive plot
  - Field selection
  - Query function: type-in query or drop down menu
- Dashboard visualization structure
  - Make it compact: combine related plots together
  - Make it friendly: design layout