SOSC4300

How Social Media data captures the evolvement of social movement

LEUNG MAN CHOI 20521189

TIN MAN SING 20521256

AU-YEUNG TSZ WAI 20520460

LAM HOI SAN 20506775

Background

How can we observe the trend of the movement?

What are the dynamics of the movement?

How protestors utilize the social media?

Dataset 1: Existing Dataset

- Twitter data from Github
- 18 raw csv files
- 15000 tweets per files
- 03/11/2019 to 20/11/2019
- Tweets, hashtag, username, date created etc....

Dataset 2: Twitter API - follower list

- List of followers of the user appeared in the above data set
- Sample size of 300

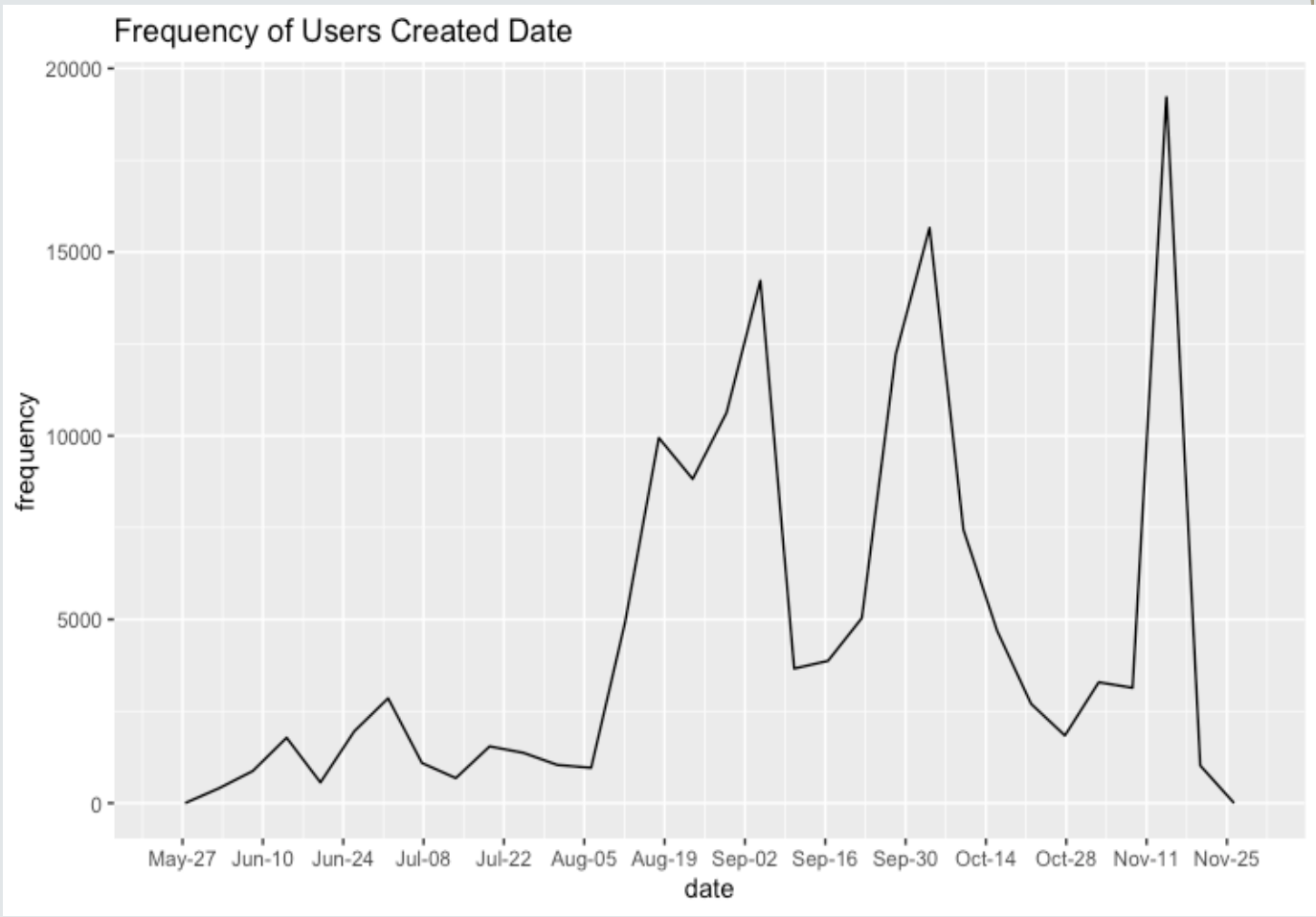
Dataset 3: Twitter API

- Searching data by keywords related to the movement
- Sample size around 500-1000
- (For R): Package Rtweet/TwitteR
- (For Gephi): Twitter streaming importer plugin

Date of creating the Twitter account

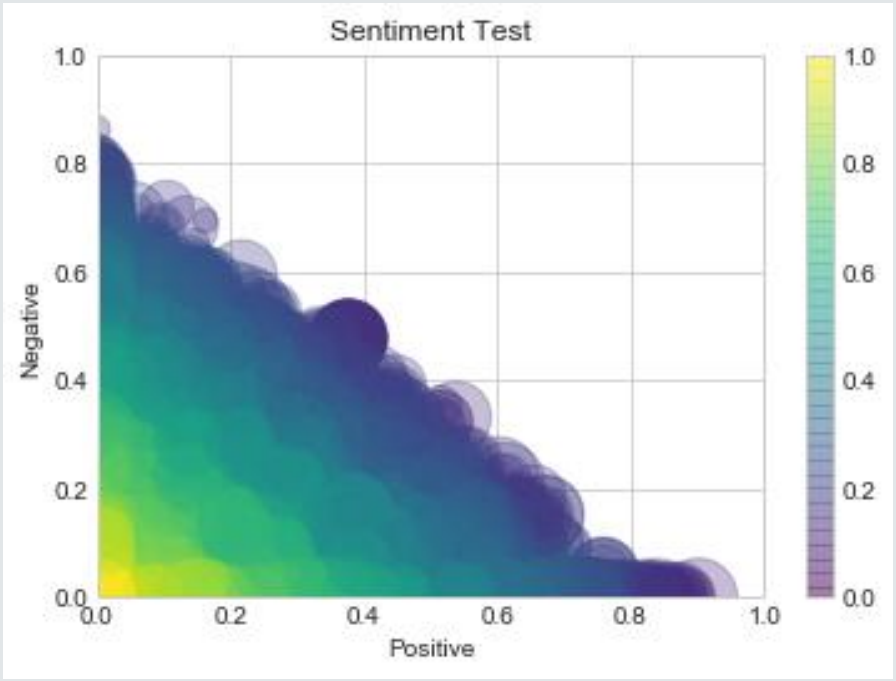
- Several peaks of creating the account

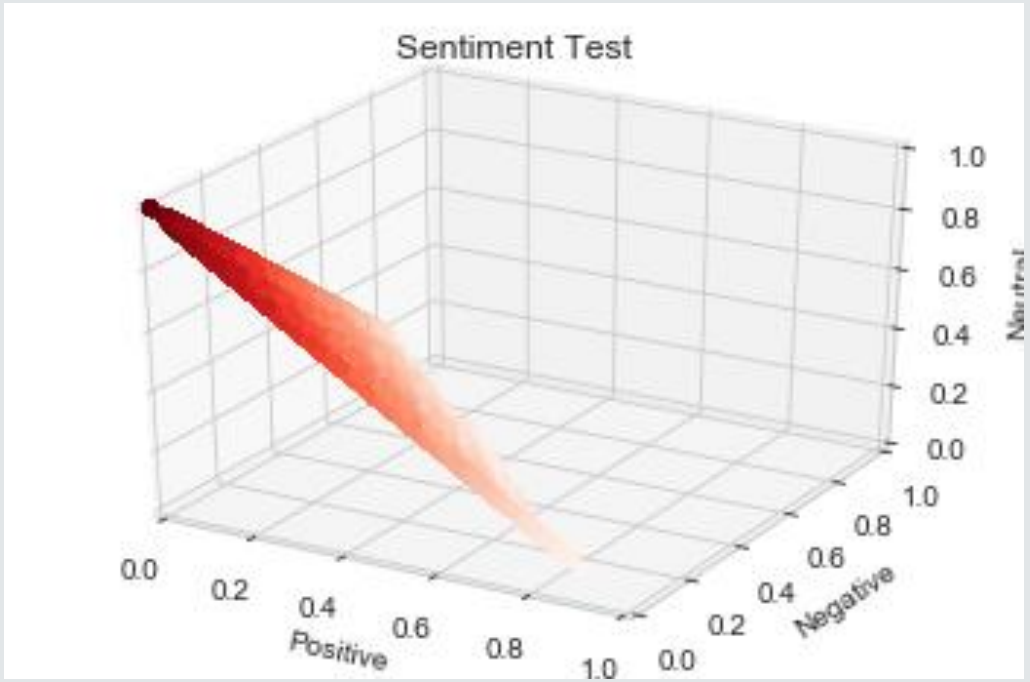
- Occupy Airport Incident
- "831" (Prince Edward Railway Station Incident)
- National Day
- Sieged PolyU & CUHK Incident



Sentiment classification

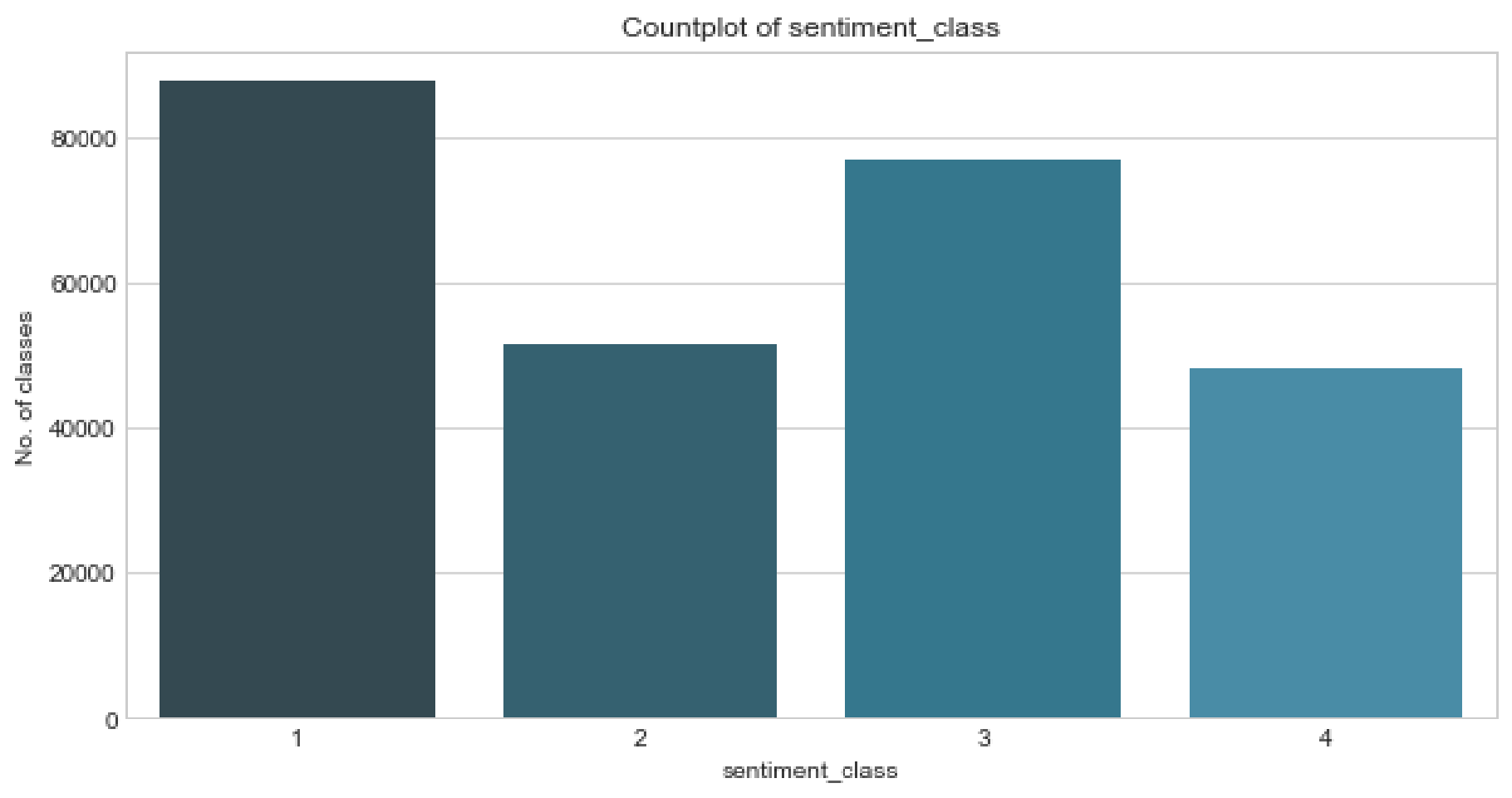
- Sentiment classification
  - NLTK Vader_Lexicon Library
  - Compound Score
- Word embedding
  - Skip-gram model
- Measure the Twitter user's sentiment



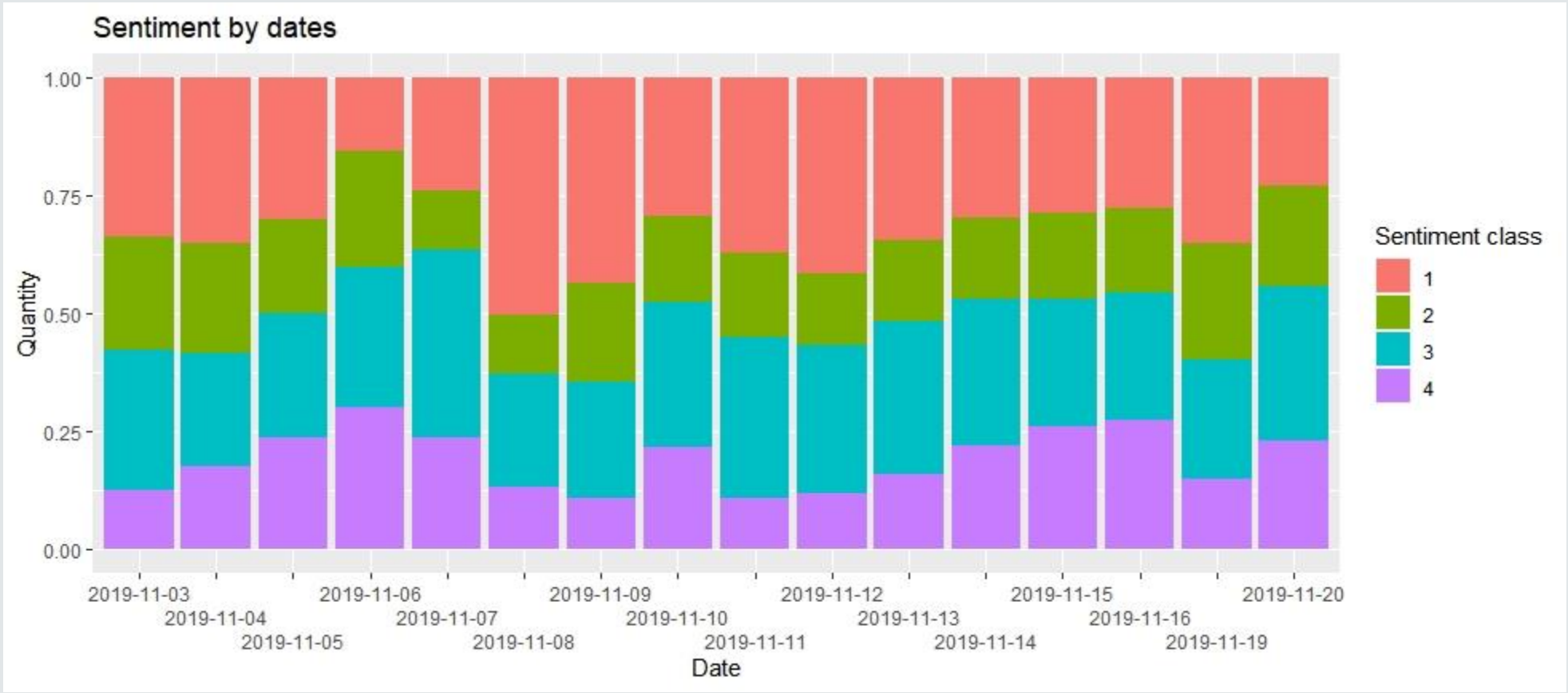


Sentiment Library
NLTK Vader_Lexicon Library

Sentiment Class



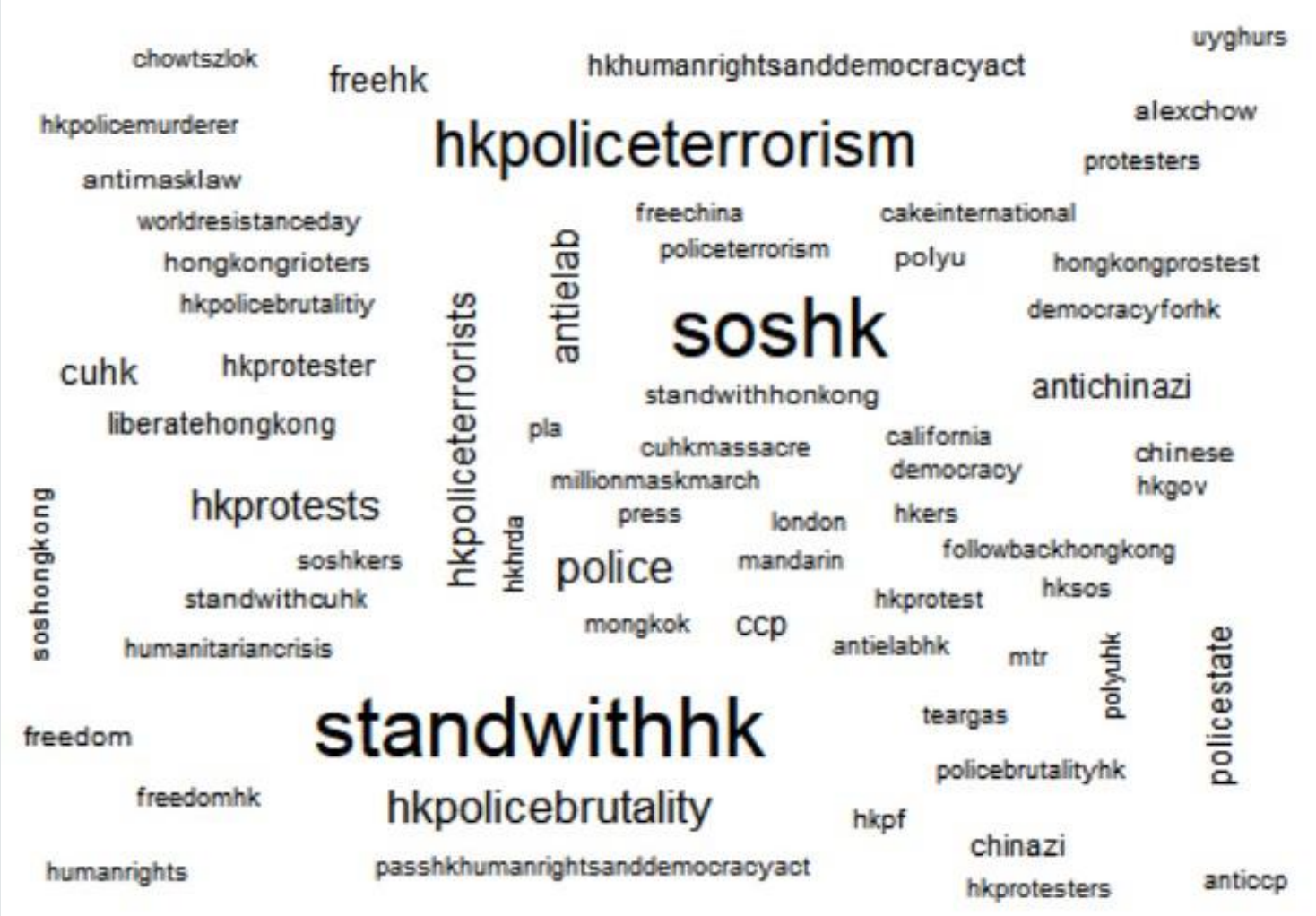
Proportion of sentiment by date



Word Embedding

- Skip-gram
- 'police'
  - 'riot', 0.5946824550628662
  - 'againsting', 0.517828643321991
  - 'force', 0.5115256905555725
  - 'bulletscuhk', 0.4930810332298279
- 'government'
  - 'afternoons', 0.6524575352668762
  - 'usstandwithhongkong', 0.6151518821716309
  - 'clarified', 0.5893750190734863

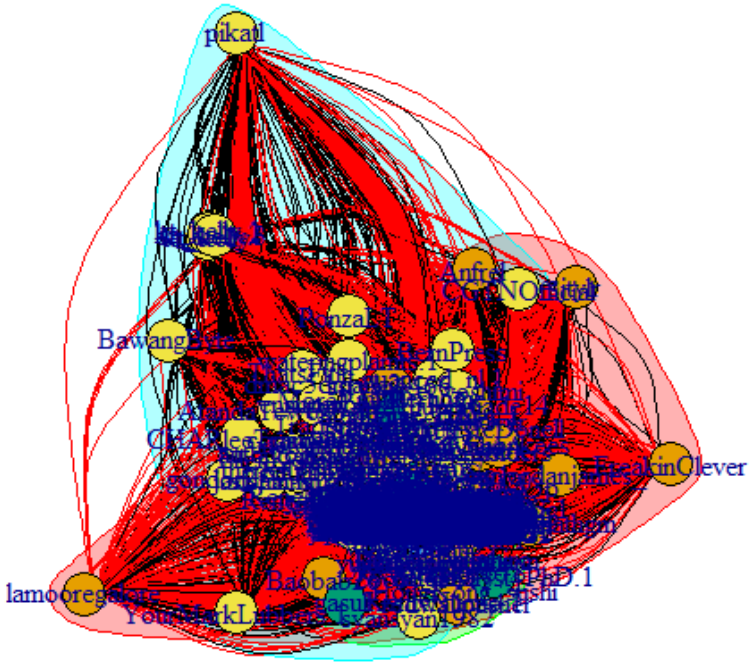
Word Cloud (hashtags)



Clustering by Shared audience

- Sample of 316 users
- No. Of common audience (follower) => weight = edges
- Louvain Community Detection method
- Modularity: measuring the density of connection within clusters compared to the density of connections between clusters (Blondel 2008)

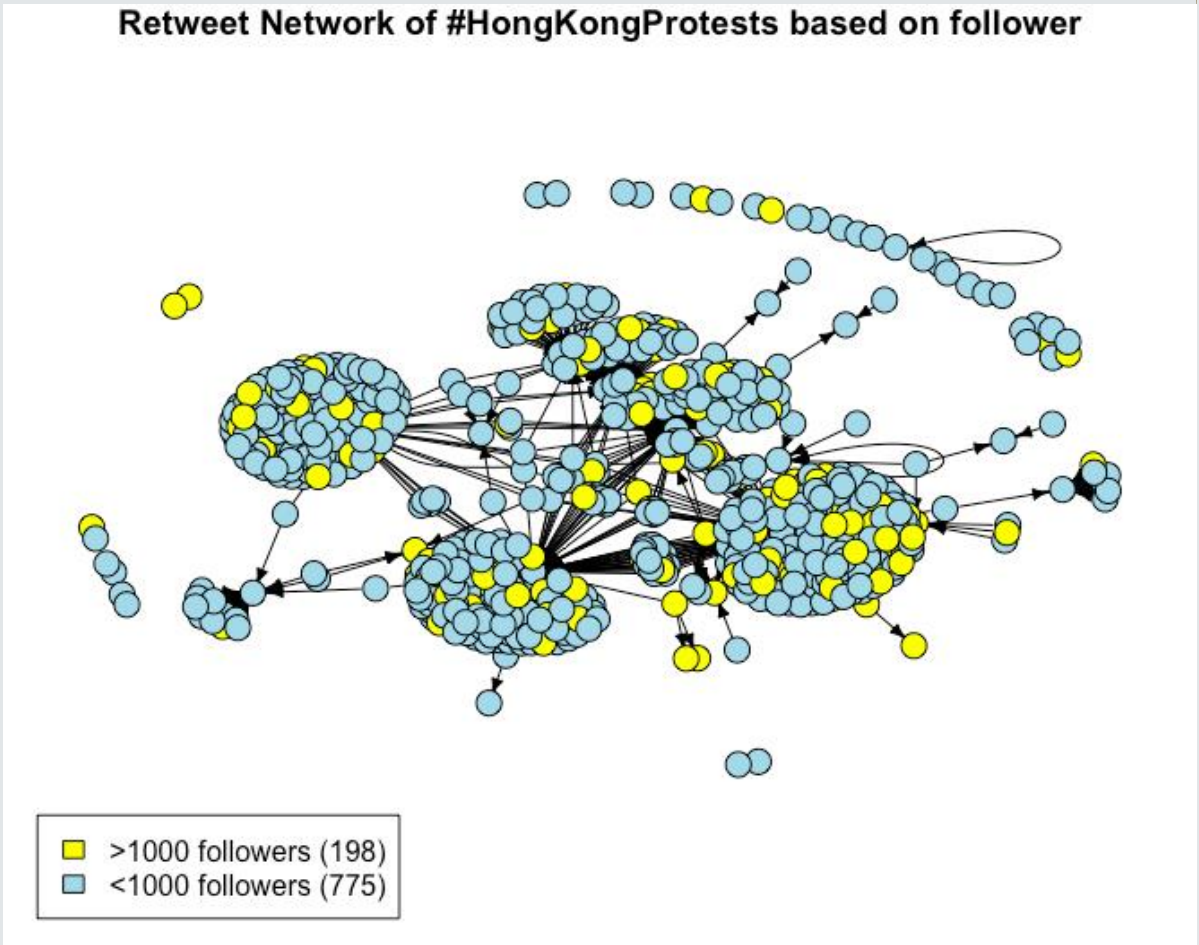
Clustering by Shared audience



- 6 distinct clusters of different size
- Largest cluster has 95 members
- Smallest cluster has only 2 members!

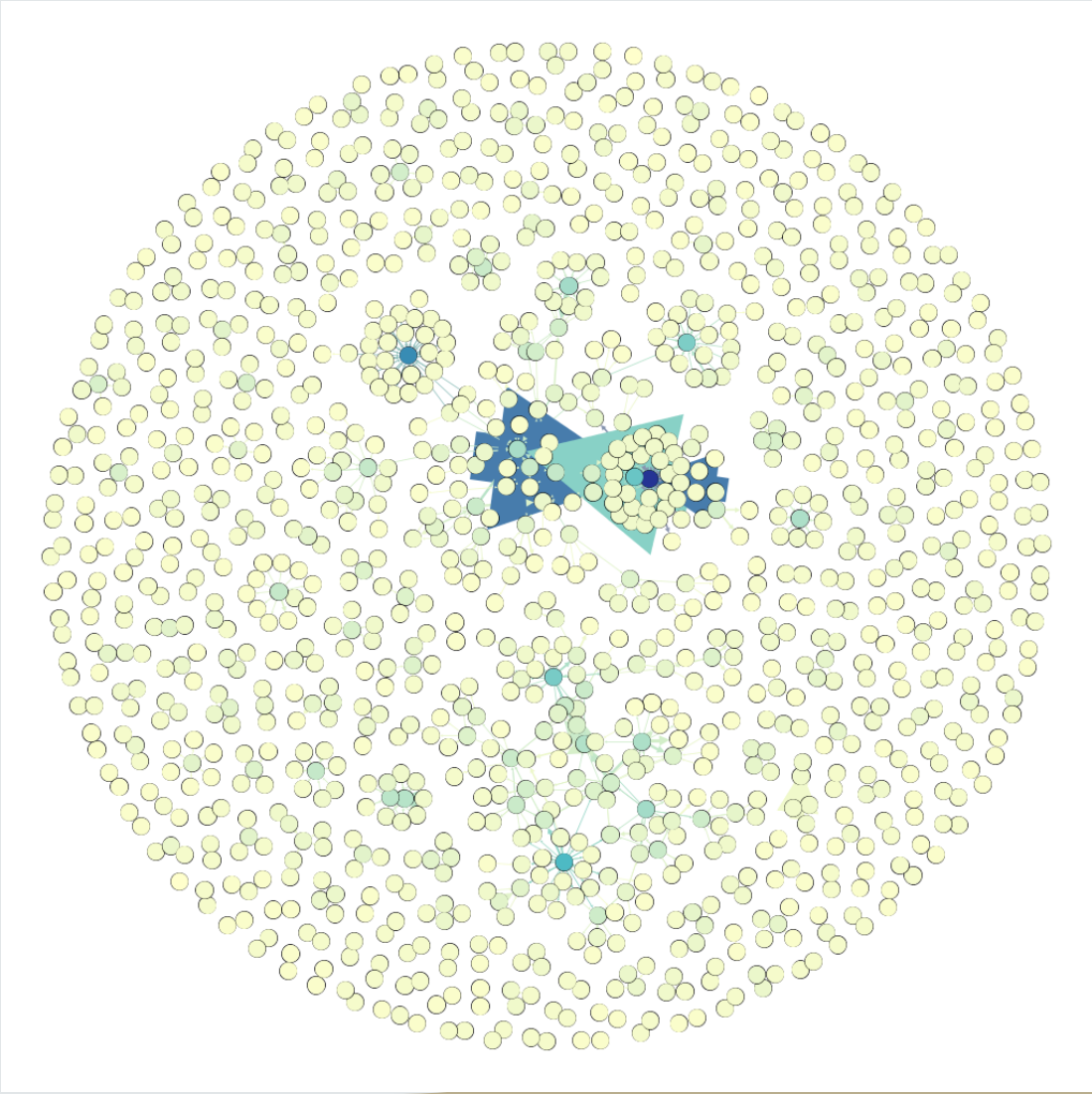
Retweet Network Analysis (Twitter API)

- Around 5 major cluster (N=971)
- Clusters are concentrated
- Only a few nodes are willing to retweet other cluster's post
- Yellow dots: influential users, which are in the center of the cluster, but also helping different cluster to connect



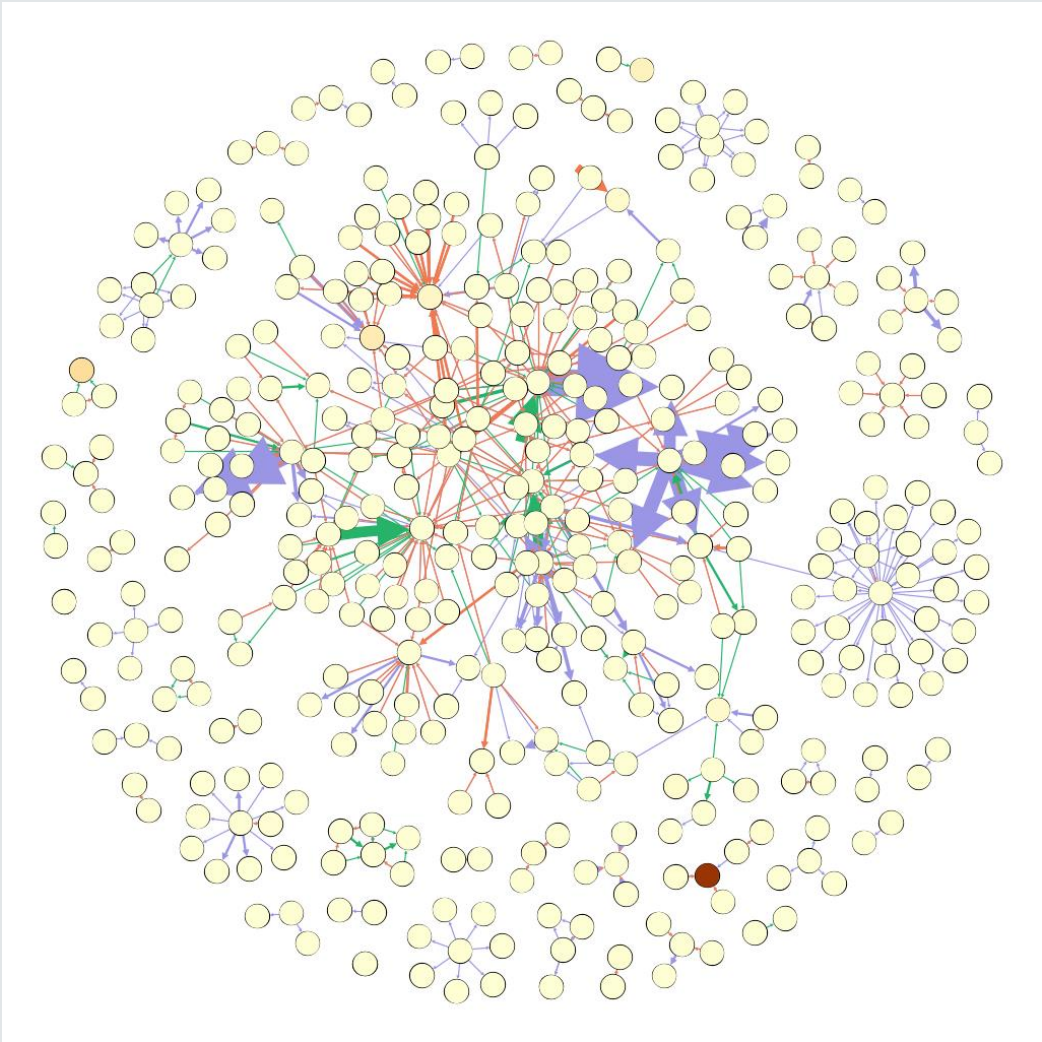
User Network: Twitter API (Gephi)

- Keyword: "China", "Hong Kong Protest", "Hong Kong", "Anti-ELAB"
- N=1216
- The node in the middle: @realdonaldtrump ; @llinwood
- Not exactly what we are looking for
- Refined result later on



User Network: Twitter API (Gephi)

- New Keywords: "721","831", "polyusiege","cuhksiege" + existing keywords
- N = 417
- Edge:
  - Purple = mention
  - Orange = retweet
  - Green = quote
- Color of node represent the follower count
  - The Red one is @bts_twt
  - The node in the middle: @quicktake (Bloomberg news), @eileenechang (activist/participant?), @studioincendo (web-based independent news media page)



Future Direction

- A better and more concise keyword for the twitter API
- Qualitative analysis on selected users who have great influence
- Larger sample for shared audience graph
- -> Explore some common characteristics between members within the communities.