# CATS

C Crawl Data

Movie M

A Activity

Book B

S Sentiment

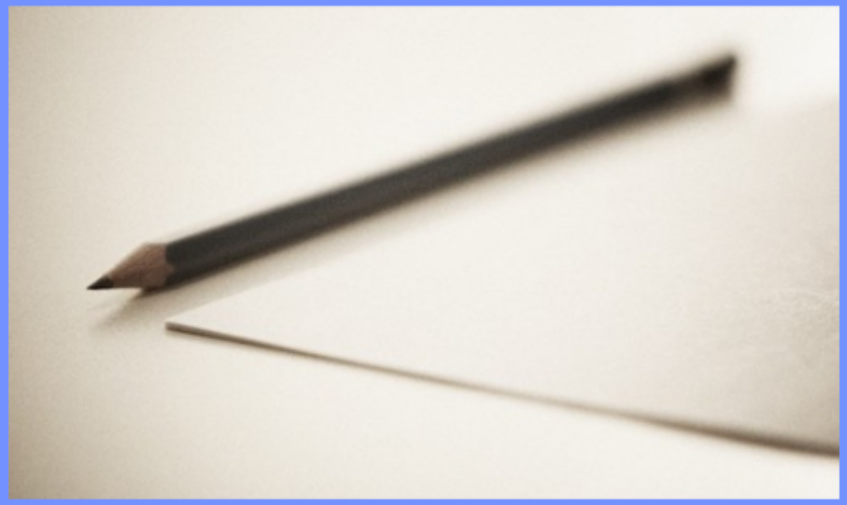
Province P

R Result & Appendix

# DOGS

# Data crawling

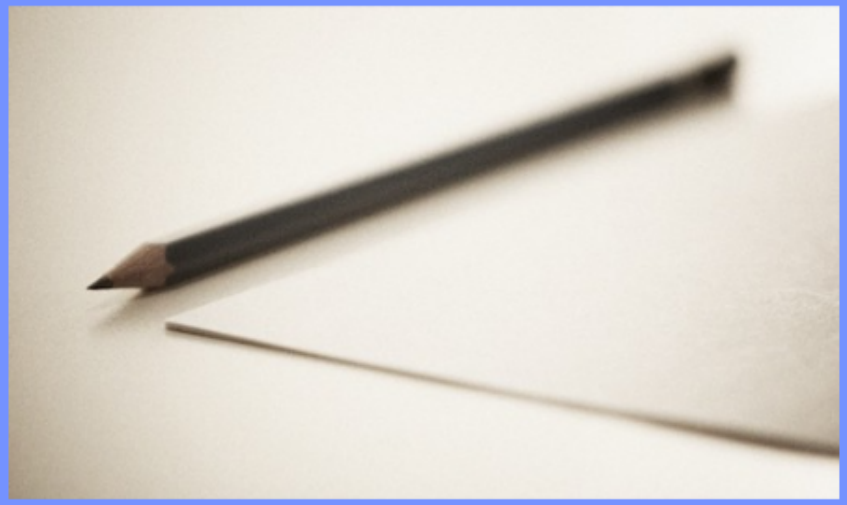
1. Two biggest groups in Douban - cat:337,238 ; dog:156,360



# Data crawling

1. Two biggest groups in Douban - cat:337,238 ; dog:156,360

2. Location Data: 10k for each group

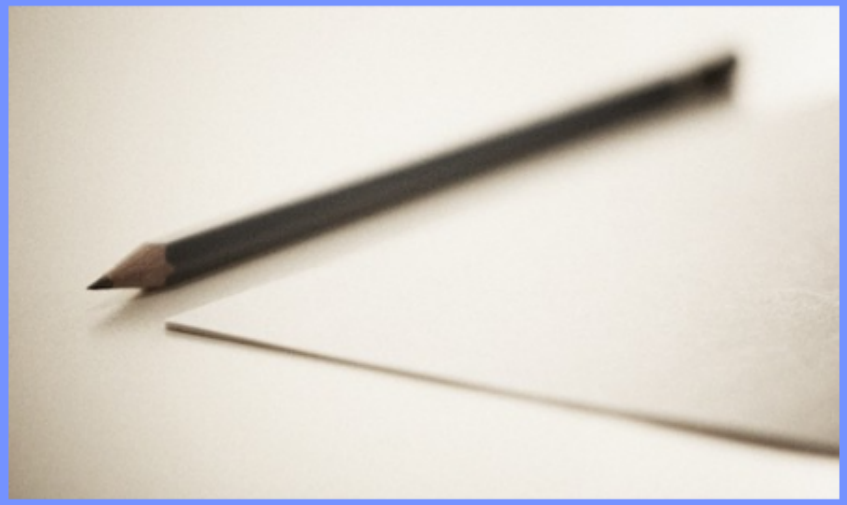


# Data crawling

1. Two biggest groups in Douban - cat:337,238 ; dog:156,360

2. Location Data: 10k for each group

3. Randomly pick 2k samples for each group to analyse their book and movie preference and activity



# CATS

C Crawl Data

Movie M

A Activity

Book B

S Sentiment

Province P

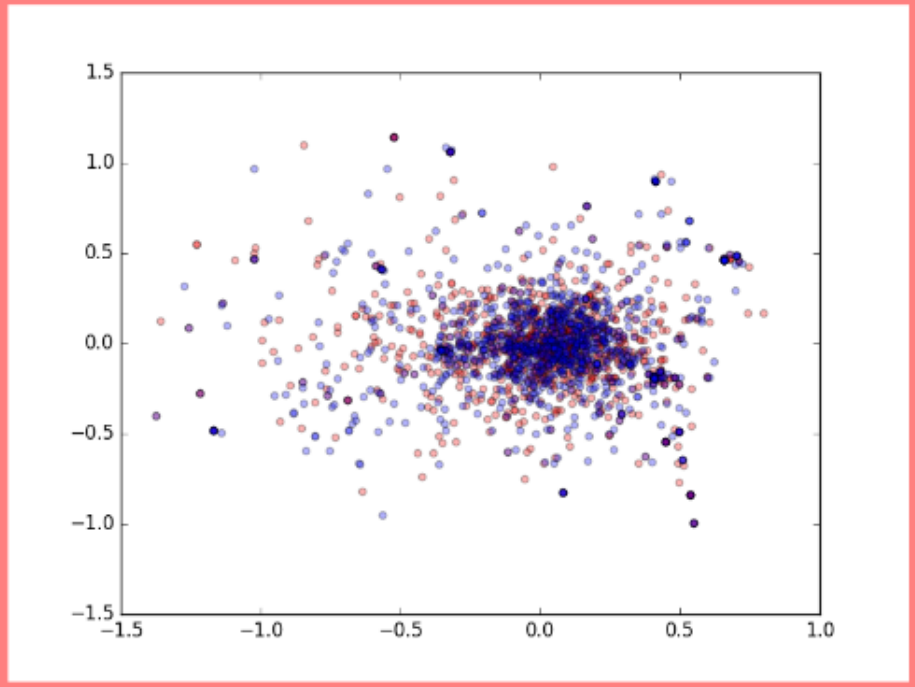
R Result & Appendix

# DOGS

# Movie

- 37 types in total
- 2k people each group - build corresponding feature vector for each people
- dimension reduction - SVD - dim(23) - 90%
- LR ; decision tree ; SVM ; naïve bayes ; KNN ; boosting

# Movie




A high overlap between them

# Movie

Are they different?

# Movie

Are they different? NOT

# Movie

Are they different? NOT

Result:

logistic regression score: 0.505
ldecision tree score: 0.541
SVM score: 0.533
Naive bayes score: 0.530
k nearest neighbour score: 0.523
boosting score: 0.513

# CATS

C Crawl Data

Movie M

A Activity

Book B

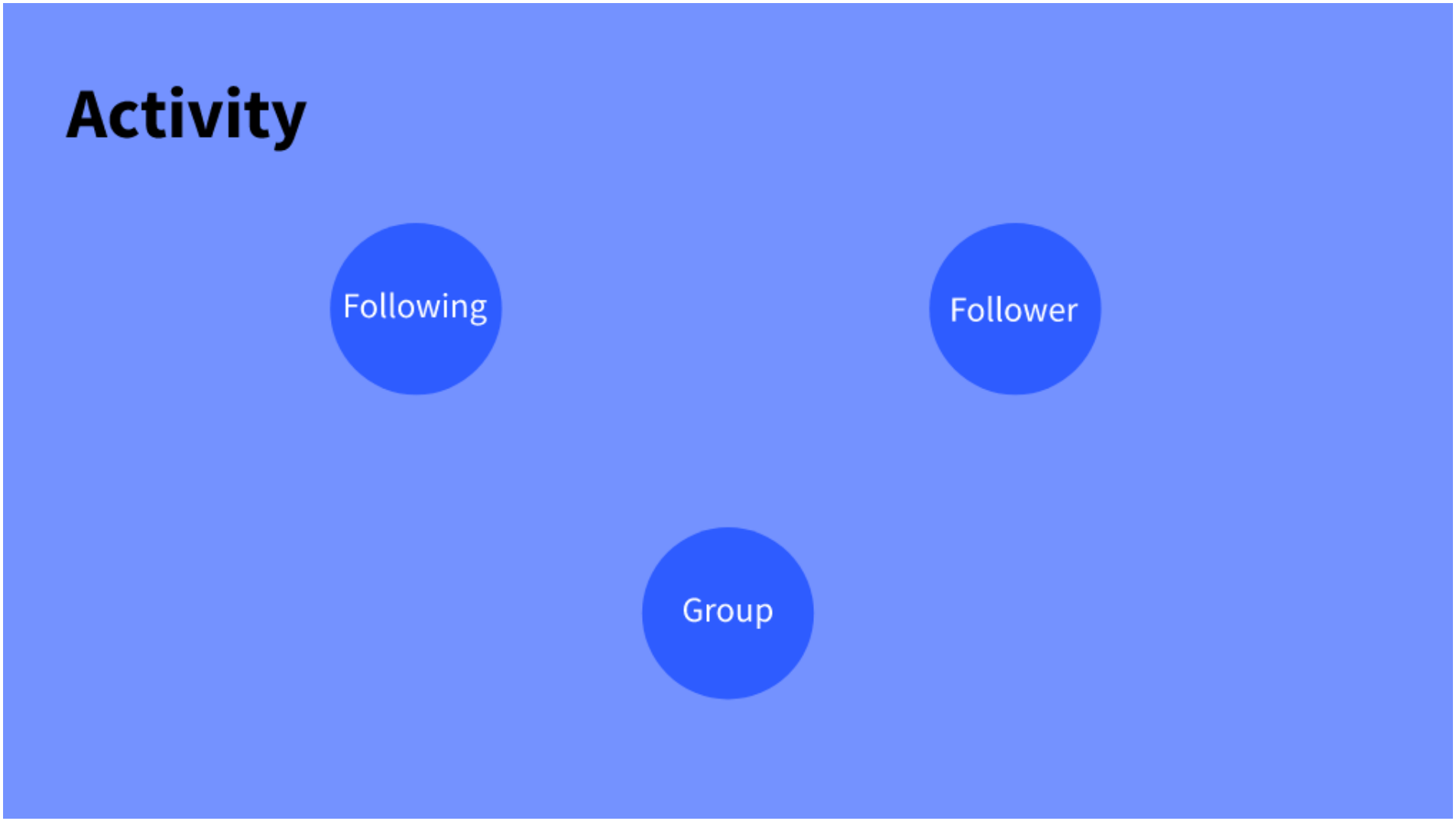
S Sentiment

Province P

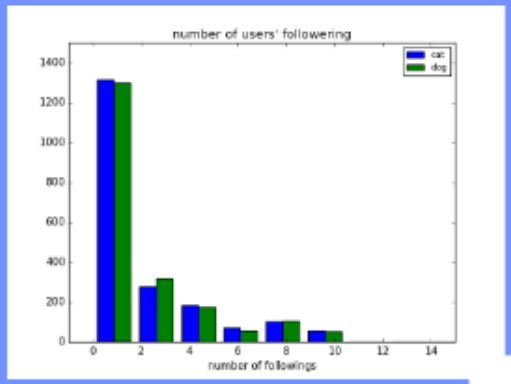
R Result & Appendix

# DOGS

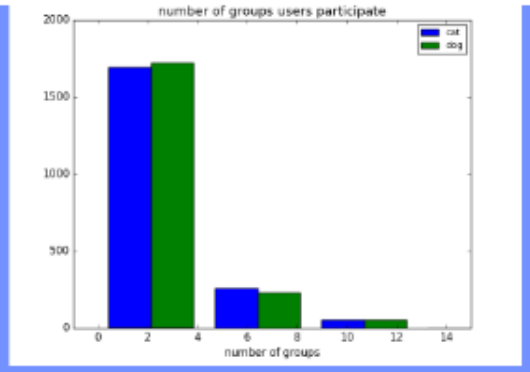
# Activity



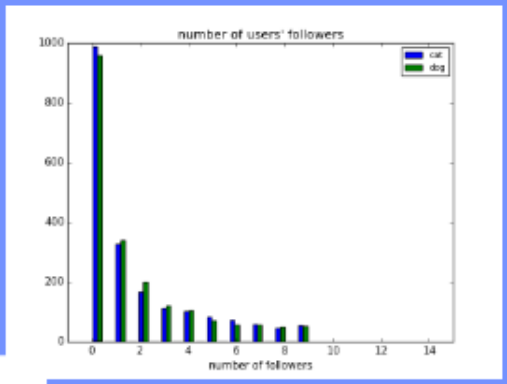
# Activity




Following




Group




Follower

# Activity

Are they different?

# Activity

Are they different? STILL NOT

# Activity

Are they different? STILL NOT

Result:

logistic regression score: 0.549
ldecision tree score: 0.503
SVM score: 0.528
Naive bayes score: 0.497
k nearest neighbour score: 0.537
boosting score: 0.555

# CATS

C Crawl Data

Movie M

A Activity

Book B

S Sentiment

Province P

R Result & Appendix

# DOGS

# Book

There is no special reading interests of people who like cats and dogs.

# **Book -** The process

Data source:

very sparse and no data for a lot of people

Data volume:

about 2k people for each group.

Implementation language:

Java

# CATS

C Crawl Data

Movie M

A Activity

Book B

S Sentiment

Province P

R Result & Appendix

# DOGS

# Sentiment

1. Pre-processing Data

- 1.1 Get data with cat and dog tags by Twitter API
- 1.2 Extract words from data and remove stop words
- 1.3 Grouped tweets together with same tags
- 1.4 Count TF-IDF

# Sentiment

1. Pre-processing Data

2. Count the sentiment of different groups

   - 2.1 Use dictionary to divided words into three categories: positive, negative and neutral
   - 2.2 Count the weight of positive and negative terms
   - 2.3 Normalize the result

| | Cat | Dog |
|---|---|---|
| Positive | 0.60 | 0.61 |
| Negative | 0.40 | 0.39 |

# Sentiment

1. Pre-processing Data

2. Count the sentiment of different groups

3. Visualize the result

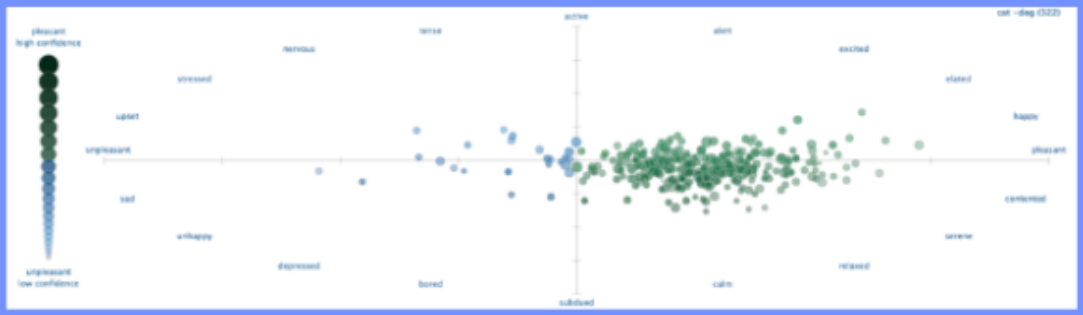
# Sentiment

1. Pre-processing Data

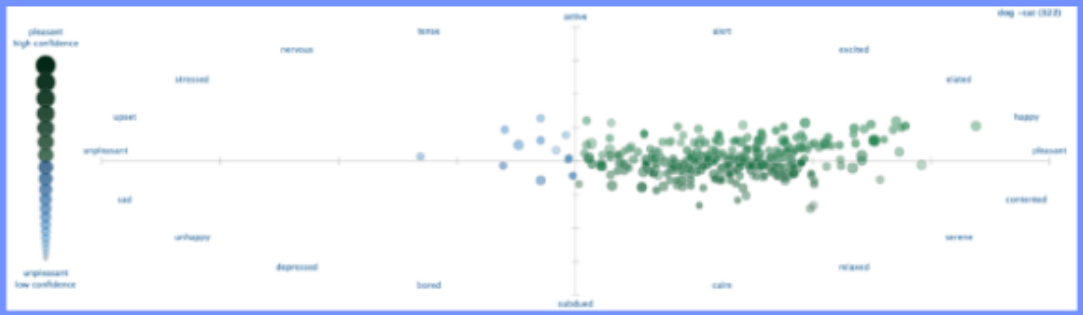
2. Count the sentiment of different groups

3. Visualize the result

4. Another visualization result



cat sentiment analysis



dog sentiment analysis

# CATS

C Crawl Data

Movie M

A Activity

Book B

S Sentiment

Province P

R Result & Appendix

# DOGS

# Province

1. Organize data

CatPeople.csv

```
id,location
146543035,上海
151632631,东营
65494746,武汉
173427242,合肥
77184085,北京
176576581,上海
142915595,厦门
138700775,上海
176575904,沈阳
78234944,重庆
175555322,广州
158999233,北京
175581870,苏州
47877728,杭州
176474997,郑州
146394266,北京
136709383,北京
58342889,杭州
140059271,北京
176529476,杭州
60913148,宽控玛依
121183577,北京
74216638,上海
52350230,北京
4721128,成都
160769112,上海
141582262,上海
68288742,广州
88129546,北京
89722783,杭州
44633627,北京
131001253,广州
141303965,石家庄
130000602,北京
```

cat_city_count.out

```
('北京', 13583)
('上海', 10307)
('广州', 5822)
('深圳', 4358)
('杭州', 3899)
('成都', 3689)
('南京', 2836)
('武汉', 2633)
('重庆', 2194)
('西安', 2049)
('苏州', 1625)
('天津', 1622)
('长沙', 1470)
('郑州', 1083)
('沈阳', 971)
('青岛', 944)
('厦门', 936)
('济南', 889)
('大连', 826)
('合肥', 768)
('宁波', 765)
('无锡', 738)
('哈尔滨', 712)
('福州', 692)
('昆明', 626)
('长春', 553)
('石家庄', 509)
('温州', 489)
('南宁', 481)
('佛山', 462)
```

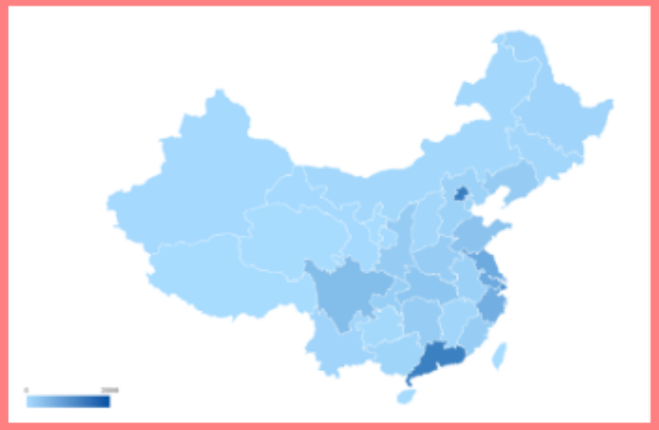
Cat_Province.data

```
{"name":"北京", "value": 13583},
{"name":"上海", "value": 10307},
{"name":"广东", "value": 13858},
{"name":"浙江", "value": 6587},
{"name":"四川", "value": 4581},
{"name":"江苏", "value": 7096},
{"name":"湖北", "value": 3329},
{"name":"重庆", "value": 2194},
{"name":"陕西", "value": 2480},
{"name":"天津", "value": 1622},
{"name":"湖南", "value": 2226},
{"name":"河南", "value": 2218},
{"name":"辽宁", "value": 2488},
{"name":"山东", "value": 3607},
{"name":"福建", "value": 2277},
{"name":"安徽", "value": 1664},
{"name":"黑龙江", "value": 1347},
{"name":"云南", "value": 1134},
{"name":"吉林", "value": 884},
{"name":"河北", "value": 1526},
{"name":"广西", "value": 1075},
{"name":"江西", "value": 976},
{"name":"山西", "value": 840},
{"name":"香港", "value": 417},
{"name":"贵州", "value": 550},
{"name":"甘肃", "value": 429},
{"name":"澳门", "value": 231},
{"name":"新疆", "value": 490},
{"name":"内蒙古", "value": 662},
{"name":"海南", "value": 312},
{"name":"宁夏", "value": 187},
{"name":"台湾", "value": 195},
{"name":"青海", "value": 144},
{"name":"西藏", "value": 100}
```

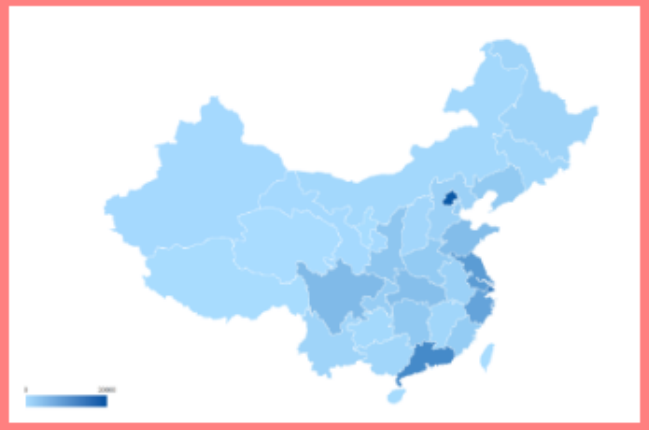
# Province

1. Organize data

2. Visualize data Demo: morikka.me:8000/tools/cats.html or dogs.html



Cat



Dog

# Province

1. Organize data

2. Visualize data

3. The difference

morikka.me:8000/tools/diff.html

# CATS

C Crawl Data

Movie M

A Activity

Book B

S Sentiment

Province P

R Result & Appendix

# DOGS

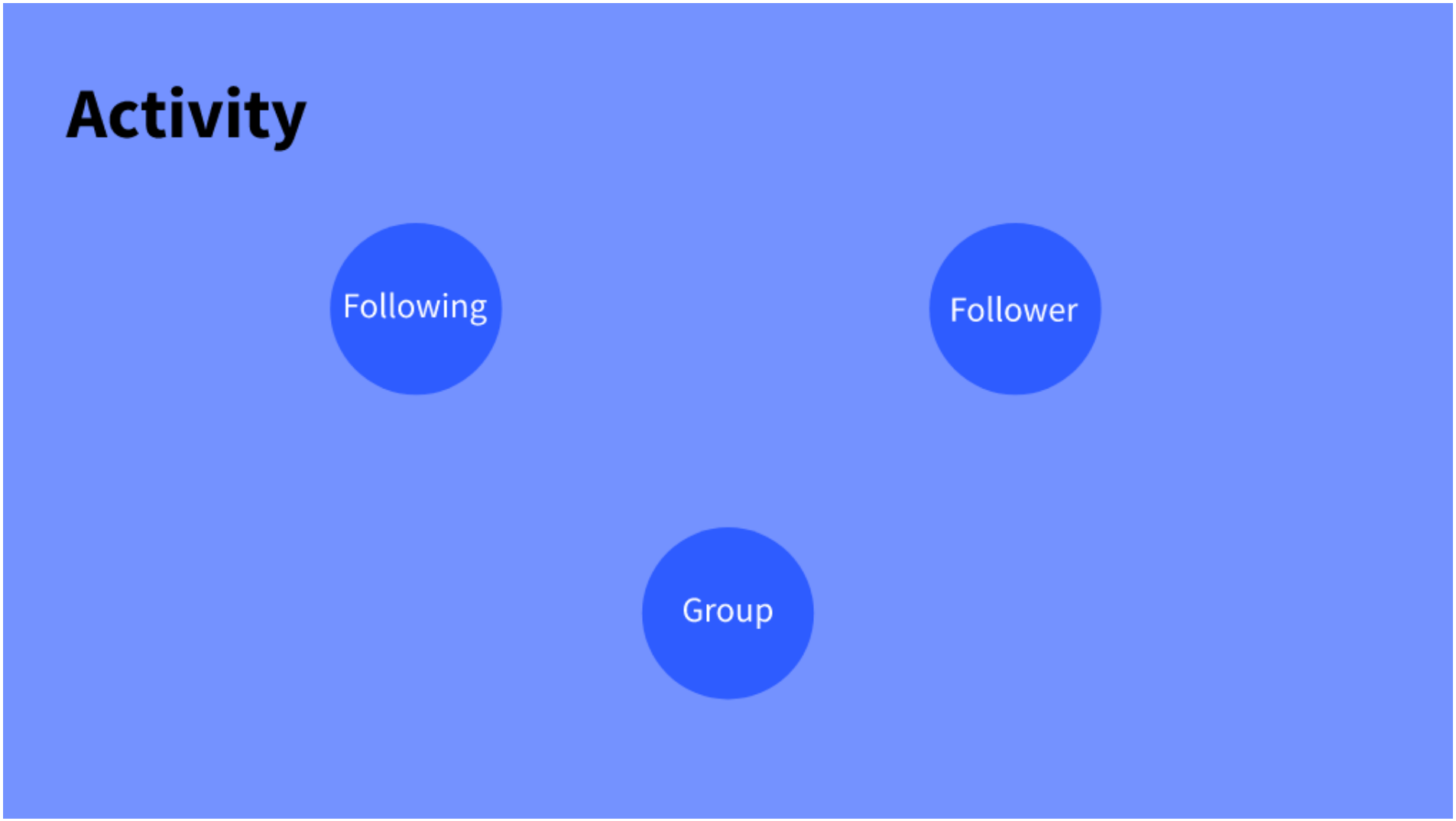
# Result

| Similar | Difference |
| --- | --- |
| Movie & Book choice | Cat is more popular online |
| Positive | Cat people are more negative |
| Friend activity | Cat people love Guangzhou<br>Dog people love Beijing |

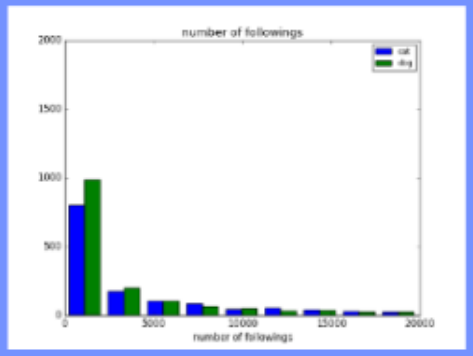
# Appendix

Community independence

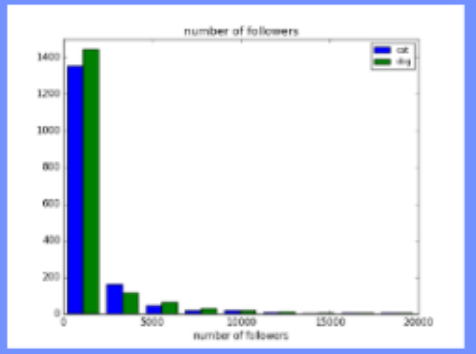
# Activity



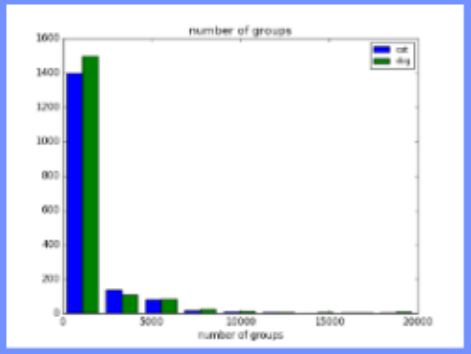
# Activity




Following




Follower




Group

# CATS

C Crawl Data

Movie M

A Activity

Book B

S Sentiment

Province P

R Result & Appendix

# DOGS