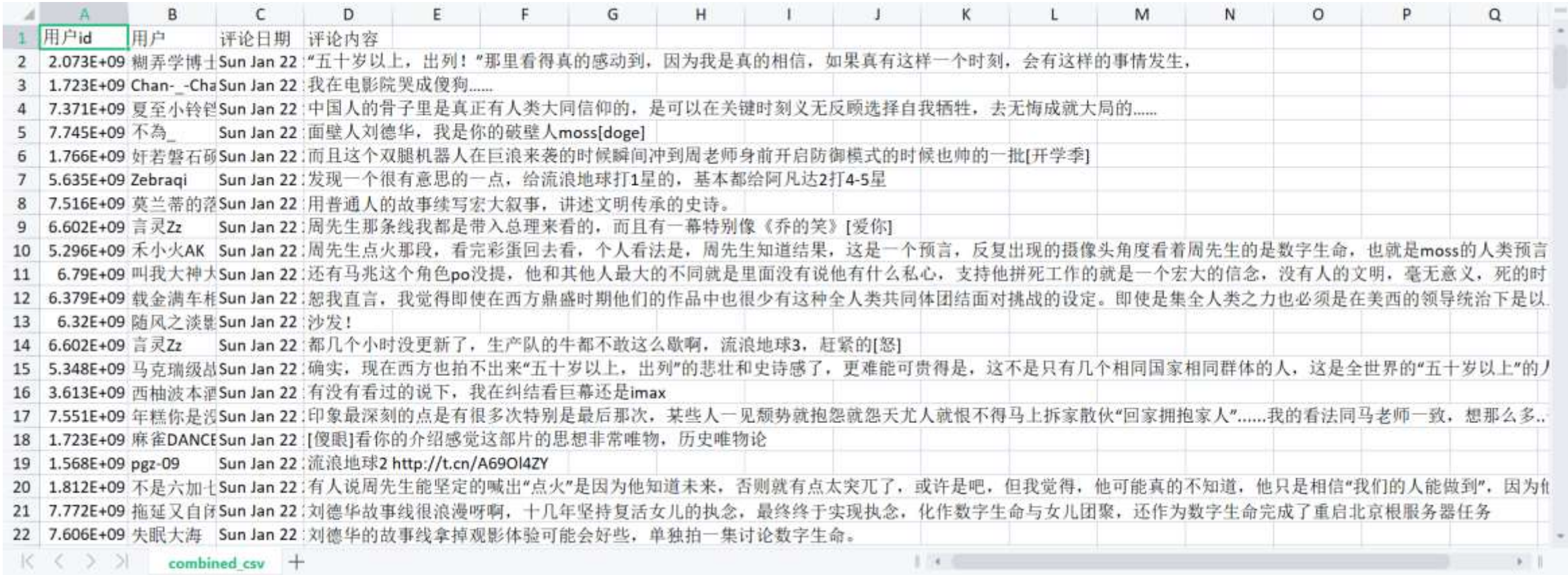
对新浪微博的影评文本进行情感分析

关于电影《流浪地球2》的评论文本

- 数据来源：新浪微博热门電影博主、导演郭帆微博
- 数据量：约5000条有效評論
- 分析步骤：爬虫—存为csv文件—分词—提取关键词—统计词频—制作词云—情感分析（snownlp)—主题聚類（构建LDA模型）
- 基于各种机器学习对微博文本进行情感分析
  - Bayes
  - SVM
  - xgboost
  - lstm
  - Bert

部分爬虫结果：

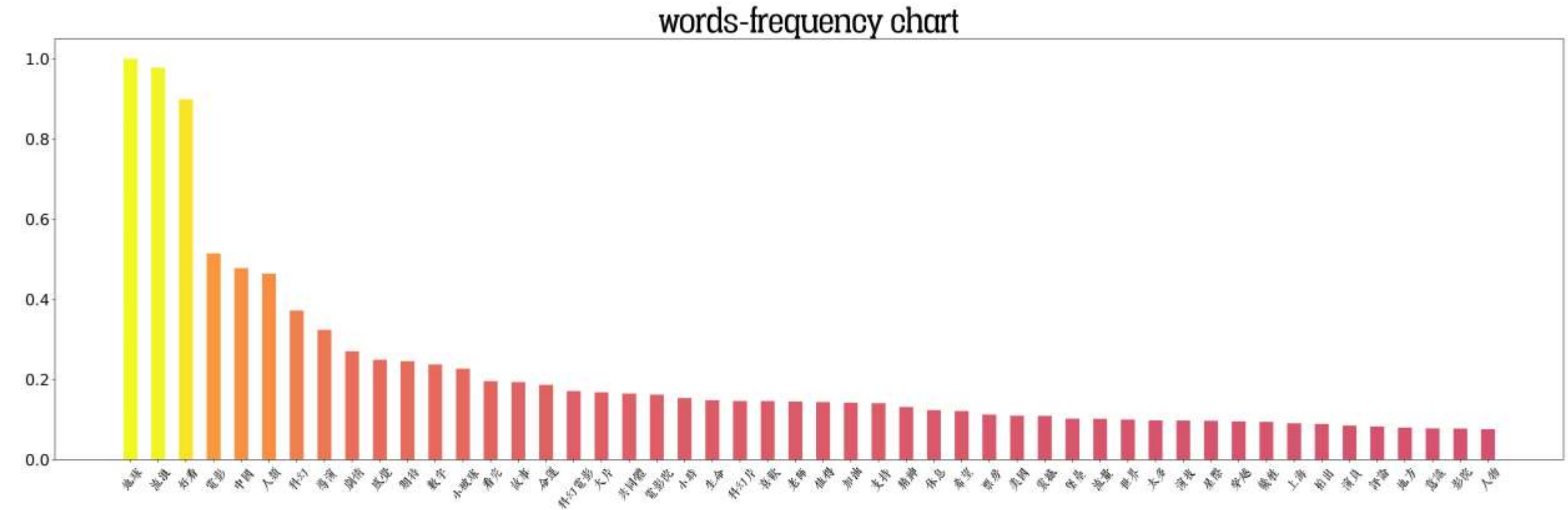


	A	B	C	D
1	用户id	用户	评论日期	评论内容
2	2.073E+09	糊弄学博士	Sun Jan 22	“五十岁以上，出列！”那里看得真的感动到，因为我是真的相信，如果真有这样一个时刻，会有这样的事情发生，
3	1.723E+09	Chan-_-Cha	Sun Jan 22	我在电影院哭成傻狗……
4	7.371E+09	夏至小铃铛	Sun Jan 22	中国人的骨子里是真正有人类大同信仰的，是可以在关键时刻义无反顾选择自我牺牲，去无悔成就大局的……
5	7.745E+09	不為_	Sun Jan 22	面壁人刘德华，我是你的破壁人moss[doge]
6	1.766E+09	奸若磐石砾	Sun Jan 22	而且这个双腿机器人在巨浪来袭的时候瞬间冲到周老师身前开启防御模式的时候也帅的一批[开学季]
7	5.635E+09	Zebraqi	Sun Jan 22	发现一个很有意思的一点，给流浪地球打1星的，基本都给阿凡达2打4-5星
8	7.516E+09	莫兰蒂的落	Sun Jan 22	用普通人的故事续写宏大叙事，讲述文明传承的史诗。
9	6.602E+09	言灵Zz	Sun Jan 22	周先生那条线我都是带入总理来看的，而且有一幕特别像《乔的笑》[爱你]
10	5.296E+09	禾小火AK	Sun Jan 22	周先生点火那段，看完彩蛋回去看，个人看法是，周先生知道结果，这是一个预言，反复出现的摄像头角度看着周先生的是数字生命，也就是moss的人类预言
11	6.79E+09	叫我大神大	Sun Jan 22	还有马兆这个角色po没提，他和其他人最大的不同就是里面没有说他有什么私心，支持他拼死工作的就是一个宏大的信念，没有人的文明，毫无意义，死的时
12	6.379E+09	载金满车桦	Sun Jan 22	恕我直言，我觉得即使在西方鼎盛时期他们的作品中也很少有这种全人类共同体团结面对挑战的设定。即使是集全人类之力也必须是在美西的领导统治下是以
13	6.32E+09	随风之淡景	Sun Jan 22	沙发！
14	6.602E+09	言灵Zz	Sun Jan 22	都几个小时没更新了，生产队的牛都不敢这么敬啊，流浪地球3，赶紧的[怒]
15	5.348E+09	马克瑞级战	Sun Jan 22	确实，现在西方也拍不出来“五十岁以上，出列”的悲壮和史诗感了，更难能可贵得是，这不是只有几个相同国家相同群体的人，这是全世界的“五十岁以上”的人
16	3.613E+09	西柚波本酒	Sun Jan 22	有没有看过的说下，我在纠结看巨幕还是imax
17	7.551E+09	年糕你是没	Sun Jan 22	印象最深刻的点是有很多次特别是最后那次，某些人一见颓势就抱怨就怨天尤人就恨不得马上拆家散伙“回家拥抱家人”……我的看法同马老师一致，想那么多..
18	1.723E+09	麻雀DANCE	Sun Jan 22	[傻眼]看你的介绍感觉这部片的思想非常唯物，历史唯物论
19	1.568E+09	pgz-09	Sun Jan 22	流浪地球2 http://t.cn/A69Ol4ZY
20	1.812E+09	不是六加七	Sun Jan 22	有人说周先生能坚定的喊出“点火”是因为他知道未来，否则就有点太突兀了，或许是吧，但我觉得，他可能真的不知道，他只是相信“我们的人能做到”，因为他
21	7.772E+09	拖延又自闭	Sun Jan 22	刘德华故事线很浪漫呀啊，十几年坚持复活女儿的执念，最终终于实现执念，化作数字生命与女儿团聚，还作为数字生命完成了重启北京根服务器任务
22	7.606E+09	失眠大海	Sun Jan 22	刘德华的故事线拿掉观影体验可能会好些，单独拍一集讨论数字生命。

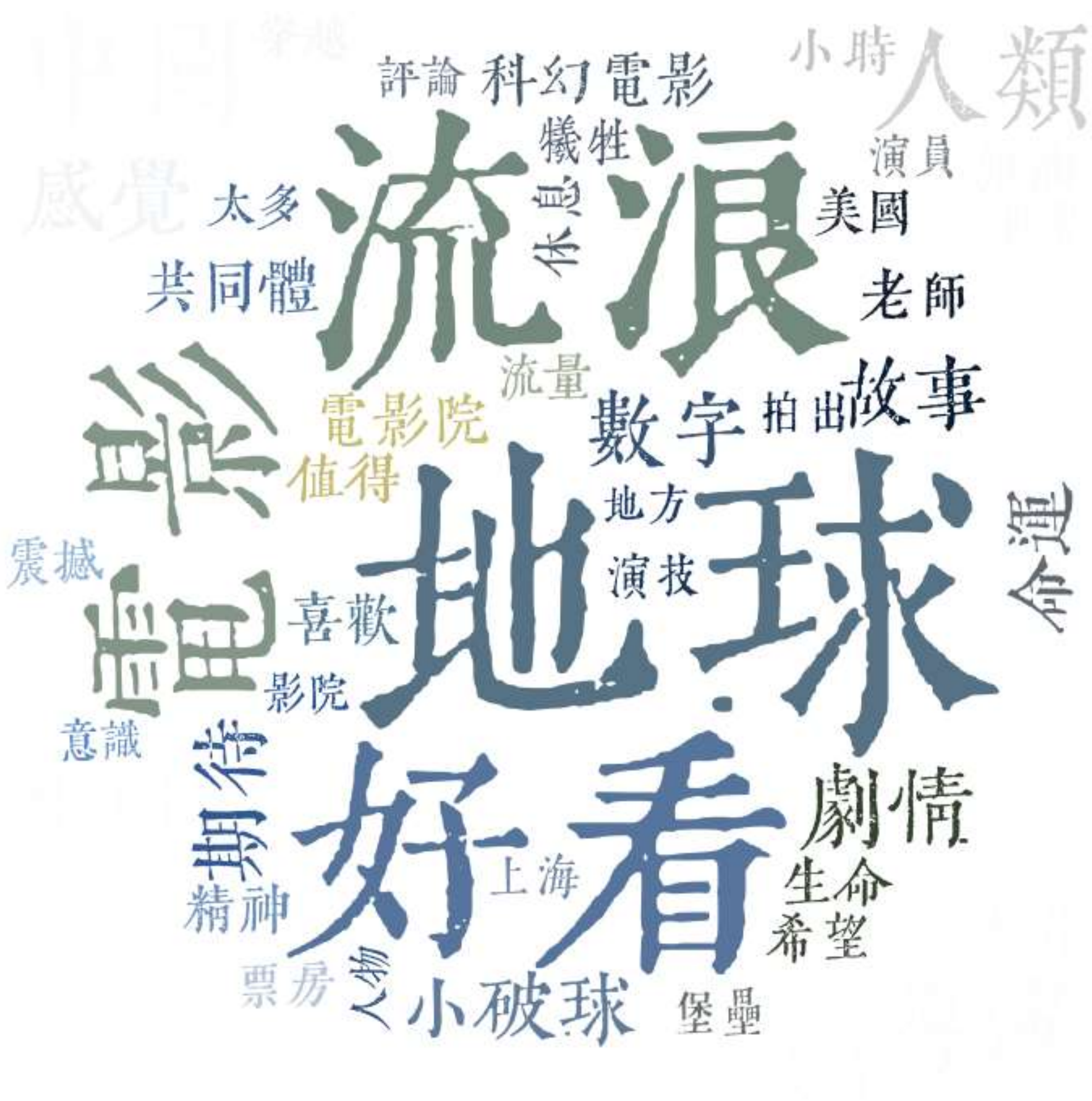
combined_csv

数据可视化结果：

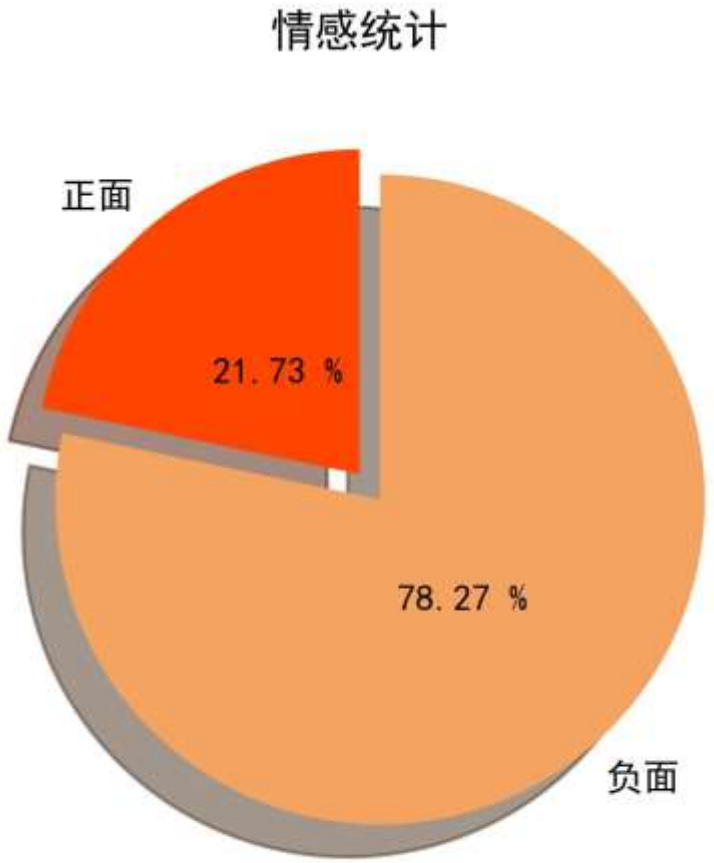
词频图



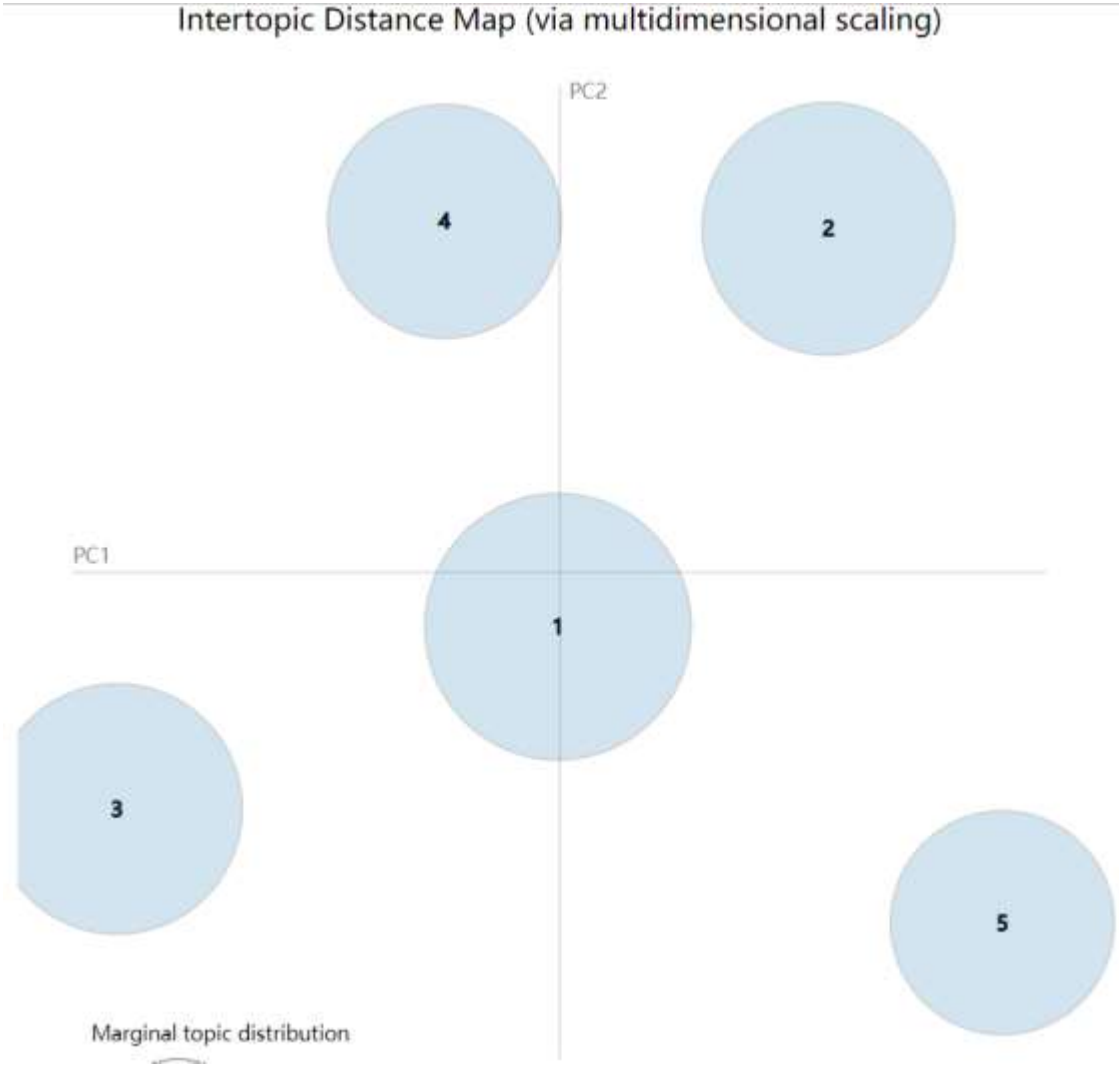
词云图

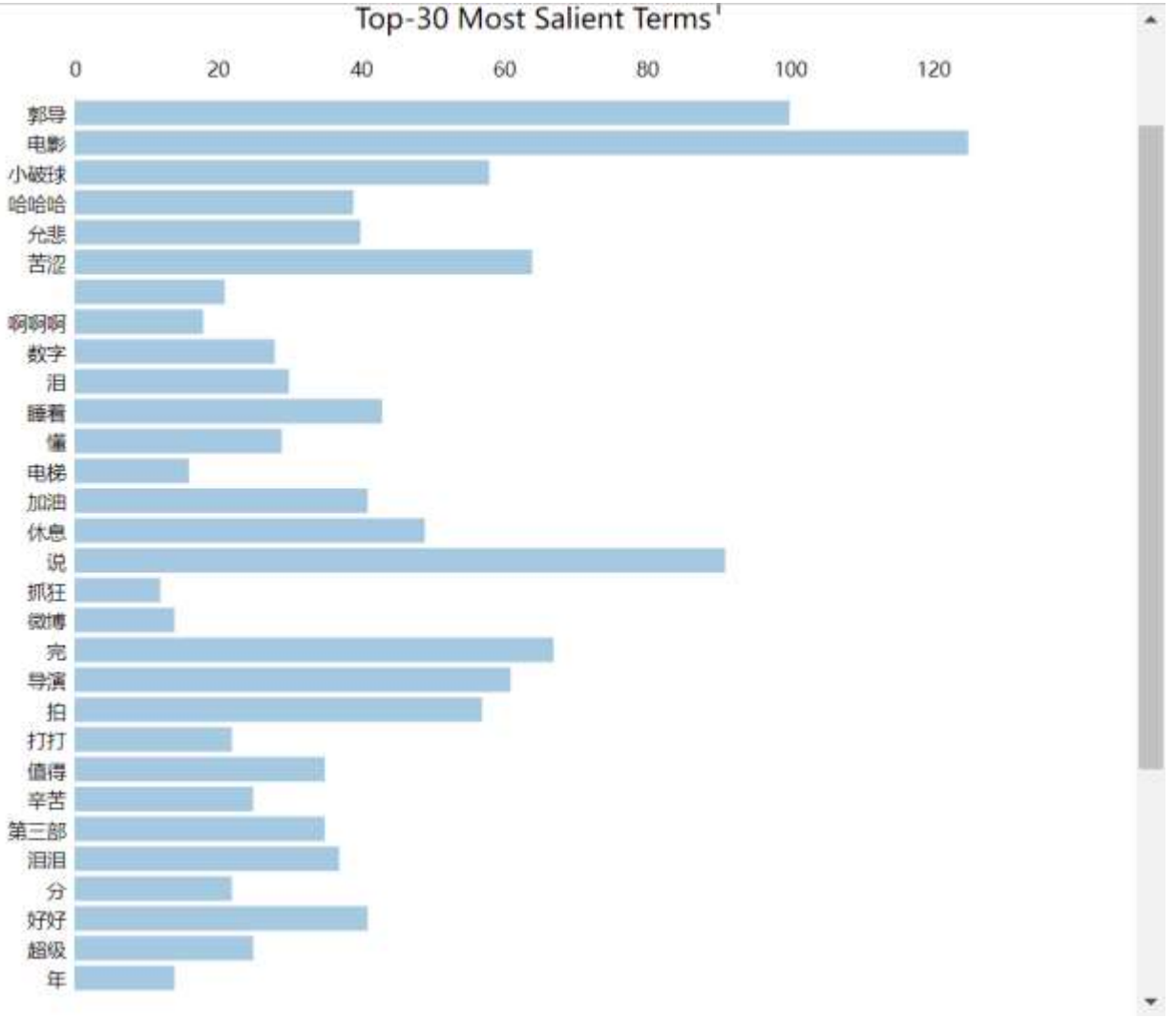

流浪
地球
好看
電影
期待
劇情
故事
數字
科幻電影
評論
犧牲
休息
太多
共同體
美國
老師
演員
小時
人類
感覺
流量
電影院
值得
拍出
地方
演技
命運
震撼
喜歡
影院
意識
精神
上海
生命
希望
票房
人物
小破球
堡壘

情感统计
（使用snownlp）



LDA聚类展示


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
4
2
1
3
5
Marginal topic distribution


Top-30 Most Salient Terms
0
20
40
60
80
100
120
郭导
电影
小破球
哈哈哈
允悲
苦涩
啊啊啊
数字
泪
睡着
懂
电梯
加油
休息
说
抓狂
微博
完
导演
拍
打打
值得
辛苦
第三部
泪泪
分
好好
超级
年

各种机器学习在测试集上的测试结果

```
from sklearn.naive_bayes import MultinomialNB

clf = MultinomialNB()
clf.fit(X_train, y_train)

# 在测试集上用模型预测结果
y_pred = clf.predict(X_test)
```

...

```
# 测试集效果检验
from sklearn import metrics

print(metrics.classification_report(y_test, y_pred))
print("准确率:", metrics.accuracy_score(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.84      0.38      0.52        69
           1       0.75      0.96      0.84       131

    accuracy                           0.76       200
   macro avg       0.79      0.67      0.68       200
weighted avg       0.78      0.76      0.73       200

准确率: 0.76
```

Bayes

```
from sklearn import svm

clf = svm.SVC()
clf.fit(X_train, y_train)
```

```
# 在测试集上用模型预测结果
y_pred = clf.predict(X_test)
# 测试集效果检验
from sklearn import metrics

print(metrics.classification_report(y_test, y_pred))
print("准确率:", metrics.accuracy_score(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.97      0.46      0.63        69
           1       0.78      0.99      0.87       131

    accuracy                           0.81       200
   macro avg       0.87      0.73      0.75       200
weighted avg       0.84      0.81      0.79       200

准确率: 0.81
```

SVM

```
dmatrix = xgb.DMatrix(X_test)
y_pred = model.predict(dmatrix)

from sklearn import metrics

auc_score = metrics.roc_auc_score(y_test, y_pred)
y_pred = list(map(lambda x:'1' if x > 0.5 else '0', y_pred))

# 测试集效果检验
print(metrics.classification_report(y_test, y_pred))
print("准确率:", metrics.accuracy_score(y_test, y_pred))
print("AUC:", auc_score)
```

```
              precision    recall  f1-score   support

           0       0.80      0.51      0.62        69
           1       0.78      0.93      0.85       131

    accuracy                           0.79       200
   macro avg       0.79      0.72      0.73       200
weighted avg       0.79      0.79      0.77       200

准确率: 0.785
AUC: 0.8027436663347716
```

xgboost

lstm

进行迭代训练：

```
epoch:1, step:10, loss:0.6848745346069336
epoch:1, step:20, loss:0.6560856103897095
              precision    recall  f1-score   support

         0.0       0.00      0.00      0.00        69
         1.0       0.65      0.99      0.79       131

    accuracy                           0.65       200
   macro avg       0.33      0.50      0.39       200
weighted avg       0.43      0.65      0.52       200

准确率: 0.65
AUC: 0.6552162849872774
saved model:  C:/Users/11040/Desktop/weibonlp-master/sentiments/model/lstm_1.model
epoch:2, step:10, loss:0.6238754987716675
epoch:2, step:20, loss:0.6135467886924744
              precision    recall  f1-score   support

         0.0       1.00      0.03      0.06        69
         1.0       0.66      1.00      0.80       131

    accuracy                           0.67       200
   macro avg       0.83      0.51      0.43       200
weighted avg       0.78      0.67      0.54       200

准确率: 0.665
AUC: 0.7511339749972342
saved model:  C:/Users/11040/Desktop/weibonlp-master/sentiments/model/lstm_2.model
```

```
epoch:3, step:10, loss:0.5807081460952759
epoch:3, step:20, loss:0.5680990815162659
              precision    recall  f1-score   support

         0.0       1.00      0.14      0.25        69
         1.0       0.69      1.00      0.82       131

    accuracy                           0.70       200
   macro avg       0.84      0.57      0.53       200
weighted avg       0.80      0.70      0.62       200

准确率: 0.705
AUC: 0.7937271822104215
saved model:  C:/Users/11040/Desktop/weibonlp-master/sentiments/model/lstm_3.model
epoch:4, step:10, loss:0.4826539158821106
epoch:4, step:20, loss:0.5218141674995422
              precision    recall  f1-score   support

         0.0       0.82      0.41      0.54        69
         1.0       0.75      0.95      0.84       131

    accuracy                           0.77       200
   macro avg       0.79      0.68      0.69       200
weighted avg       0.78      0.77      0.74       200

准确率: 0.765
AUC: 0.8365416528377032
saved model:  C:/Users/11040/Desktop/weibonlp-master/sentiments/model/lstm_4.model
epoch:5, step:10, loss:0.4440646767616272
epoch:5, step:20, loss:0.4457719922065735
              precision    recall  f1-score   support

         0.0       0.81      0.43      0.57        69
         1.0       0.76      0.95      0.84       131

    accuracy                           0.77       200
   macro avg       0.79      0.69      0.70       200
weighted avg       0.78      0.77      0.75       200

准确率: 0.77
AUC: 0.8354353357672308
saved model:  C:/Users/11040/Desktop/weibonlp-master/sentiments/model/lstm_5.model
```

bert

进行迭代训练：

```
epoch:1, step:10, loss:0.6797645092010498
epoch:1, step:20, loss:0.6257932782173157
epoch:1, step:30, loss:0.5806602835655212
epoch:1, step:40, loss:0.5842814445495605
epoch:1, step:50, loss:0.5907979011535645
epoch:1, step:60, loss:0.5796312689781189
epoch:1, step:70, loss:0.5663250088691711
              precision    recall  f1-score   support

           0       0.69      0.13      0.23        67
           1       0.68      0.97      0.80       125

    accuracy                           0.68       192
   macro avg       0.68      0.55      0.51       192
weighted avg       0.68      0.68      0.60       192

准确率: 0.6770833333333334
AUC: 0.7130149253731344
saved model:  C:/Users/11040/Desktop/weibonlp-master/sentiments/model/bert_dnn_1.model

准确率: 0.7239583333333334
AUC: 0.816077441077441
saved model:  C:/Users/11040/Desktop/weibonlp-master/sentiments/model/bert_dnn_2.model

准确率: 0.71875
AUC: 0.8238037552998183
saved model:  C:/Users/11040/Desktop/weibonlp-master/sentiments/model/bert_dnn_3.model

准确率: 0.734375
AUC: 0.8477082596912925
saved model:  C:/Users/11040/Desktop/weibonlp-master/sentiments/model/bert_dnn_4.model

准确率: 0.7395833333333334
AUC: 0.8528225806451613
saved model:  C:/Users/11040/Desktop/weibonlp-master/sentiments/model/bert_dnn_5.model

准确率: 0.7447916666666666
AUC: 0.83721923828125
saved model:  C:/Users/11040/Desktop/weibonlp-master/sentiments/model/bert_dnn_6.model

准确率: 0.75
AUC: 0.851070226070226
saved model:  C:/Users/11040/Desktop/weibonlp-master/sentiments/model/bert_dnn_7.model

准确率: 0.75
AUC: 0.8539104477611941
saved model:  C:/Users/11040/Desktop/weibonlp-master/sentiments/model/bert_dnn_8.model
```