1.主题和情感的描述性统计分析

1）各主题所占百分比

10003 17.952597

10000 12.153303

10005 9.783157

10006 9.379728

10008 8.169440

10001 7.009581

10002 6.505295

10015 5.950580

10013 5.244579

10016 2.924861

10010 2.672718

10019 2.571861

10017 2.370146

10018 1.916288

10014 1.462431

10012 1.260716

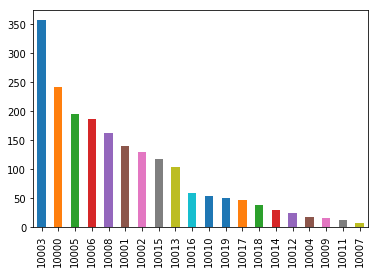
10004 0.857287

10009 0.806858

10011 0.655572

10007 0.353001

2）主题频率直方图



发现各主题的比例不一致，递减。

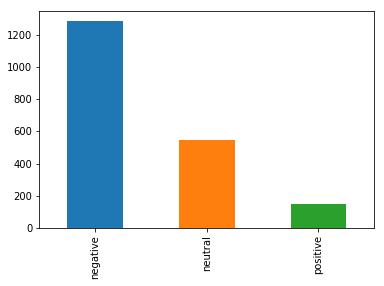
3）不同的情感所占百分比

negative 0.648008

neutral 0.276349

positive 0.075643

4）情感频率直方图



发现消极的情感占多数，中立次之，积极的很少。

2.词汇表数量不一样导致结果不一样。

1）决策树情感分析时：

选取频率最高的200个词做词汇表时的训练结果，采用的特征是bag\_of\_words

accuracy: 0.8703983862834089

precision\_micro: 0.8703983862834089

precision\_macro: 0.8561215362786693

recall\_micro: 0.8703983862834089

recall\_macro: 0.7589135669225987

f1\_micro: 0.8703983862834089

f1\_macro: 0.7961503606909495

precision recall f1-score support

negative 0.89 0.95 0.92 1285

neutral 0.82 0.78 0.80 548

positive 0.86 0.55 0.67 150

micro avg 0.87 0.87 0.87 1983

macro avg 0.86 0.76 0.80 1983

weighted avg 0.87 0.87 0.87 1983

选取频率最高的100个词做词汇表时的训练结果，采用的特征是bag\_of\_words

accuracy: 0.8451840645486637

precision\_micro: 0.8451840645486637

precision\_macro: 0.8224236750373786

recall\_micro: 0.8451840645486637

recall\_macro: 0.7276683676206526

f1\_micro: 0.8451840645486637

f1\_macro: 0.7626076315977963

precision recall f1-score support

negative 0.87 0.92 0.90 1285

neutral 0.78 0.75 0.76 548

positive 0.82 0.51 0.63 150

micro avg 0.85 0.85 0.85 1983

macro avg 0.82 0.73 0.76 1983

weighted avg 0.84 0.85 0.84 1983

选取频率最高的300个词做词汇表时的训练结果，采用的特征是bag\_of\_words

accuracy: 0.880988401412002

precision\_micro: 0.880988401412002

precision\_macro: 0.8416003929269414

recall\_micro: 0.880988401412002

recall\_macro: 0.7804547763987113

f1\_micro: 0.880988401412002

f1\_macro: 0.8060486437322979

precision recall f1-score support

negative 0.91 0.95 0.93 1285

neutral 0.83 0.81 0.82 548

positive 0.79 0.59 0.67 150

micro avg 0.88 0.88 0.88 1983

macro avg 0.84 0.78 0.81 1983

weighted avg 0.88 0.88 0.88 1983

选取频率最高的200个词做词汇表时的训练结果，采用的特征是tfidf

accuracy: 0.8920827029752899

precision\_micro: 0.8920827029752899

precision\_macro: 0.8749313670401747

recall\_micro: 0.8920827029752899

recall\_macro: 0.7920043486356078

f1\_micro: 0.8920827029752899

f1\_macro: 0.822679284518265

precision recall f1-score support

negative 0.92 0.94 0.93 1285

neutral 0.82 0.86 0.84 548

positive 0.88 0.57 0.69 150

micro avg 0.89 0.89 0.89 1983

macro avg 0.87 0.79 0.82 1983

weighted avg 0.89 0.89 0.89 1983

选取频率最高的100个词做词汇表时的训练结果，采用的特征是tfidf

accuracy: 0.8875441250630358

precision\_micro: 0.8875441250630358

precision\_macro: 0.8937356367457768

recall\_micro: 0.8875441250630358

recall\_macro: 0.7696503419264644

f1\_micro: 0.8875441250630358

f1\_macro: 0.8120017081868335

precision recall f1-score support

negative 0.90 0.96 0.93 1285

neutral 0.84 0.83 0.83 548

positive 0.94 0.53 0.68 150

micro avg 0.89 0.89 0.89 1983

macro avg 0.89 0.77 0.81 1983

weighted avg 0.89 0.89 0.88 1983

选取频率最高的300个词做词汇表时的训练结果，采用的特征是tfidf

accuracy: 0.9011598587997983

precision\_micro: 0.9011598587997983

precision\_macro: 0.8796556768064345

recall\_micro: 0.9011598587997983

recall\_macro: 0.814687848119641

f1\_micro: 0.9011598587997982

f1\_macro: 0.8407936976622331

precision recall f1-score support

negative 0.93 0.95 0.94 1285

neutral 0.84 0.86 0.85 548

positive 0.87 0.63 0.73 150

micro avg 0.90 0.90 0.90 1983

macro avg 0.88 0.81 0.84 1983

weighted avg 0.90 0.90 0.90 1983

综上所述，可以看出选取频率最高的300个词的训练结果最好

2）决策树主题分类时：

选取频率最高的200个词做词汇表时的训练结果，采用的特征是bag\_of\_words

accuracy: 0.6031265758951084

precision\_micro: 0.6031265758951084

precision\_macro: 0.5734277357239714

recall\_micro: 0.6031265758951084

recall\_macro: 0.44632208370303256

f1\_micro: 0.6031265758951084

f1\_macro: 0.4693973419140304

precision recall f1-score support

10000 0.58 0.83 0.68 241

10001 0.60 0.54 0.57 139

10002 0.59 0.75 0.66 129

10003 0.56 0.61 0.59 356

10004 0.58 0.41 0.48 17

10005 0.65 0.73 0.69 194

10006 0.48 0.53 0.51 186

10007 0.00 0.00 0.00 7

10008 0.76 0.68 0.72 162

10009 0.60 0.19 0.29 16

10010 0.56 0.62 0.59 53

10011 0.40 0.15 0.22 13

10012 0.67 0.16 0.26 25

10013 0.61 0.54 0.57 104

10014 0.58 0.24 0.34 29

10015 0.87 0.75 0.81 118

10016 0.67 0.07 0.12 58

10017 0.58 0.40 0.48 47

10018 0.45 0.45 0.45 38

10019 0.68 0.25 0.37 51

micro avg 0.60 0.60 0.60 1983

macro avg 0.57 0.45 0.47 1983

weighted avg 0.61 0.60 0.59 1983

选取频率最高的300个词做词汇表时的训练结果，采用的特征是bag\_of\_words

accuracy: 0.6369137670196672

precision\_micro: 0.6369137670196672

precision\_macro: 0.5873839373008333

recall\_micro: 0.6369137670196672

recall\_macro: 0.5174207580483938

f1\_micro: 0.6369137670196672

f1\_macro: 0.5315371090606179

precision recall f1-score support

10000 0.62 0.82 0.70 241

10001 0.63 0.57 0.60 139

10002 0.59 0.81 0.68 129

10003 0.59 0.64 0.61 356

10004 0.64 0.53 0.58 17

10005 0.70 0.73 0.71 194

10006 0.57 0.53 0.55 186

10007 0.00 0.00 0.00 7

10008 0.81 0.77 0.78 162

10009 0.71 0.75 0.73 16

10010 0.55 0.68 0.61 53

10011 0.50 0.15 0.24 13

10012 0.57 0.32 0.41 25

10013 0.61 0.56 0.58 104

10014 0.40 0.14 0.21 29

10015 0.93 0.75 0.83 118

10016 0.57 0.21 0.30 58

10017 0.65 0.36 0.47 47

10018 0.46 0.63 0.53 38

10019 0.66 0.41 0.51 51

micro avg 0.64 0.64 0.64 1983

macro avg 0.59 0.52 0.53 1983

weighted avg 0.64 0.64 0.63 1983

选取频率最高的500个词做词汇表时的训练结果，采用的特征是bag\_of\_words

accuracy: 0.6555723651033787

precision\_micro: 0.6555723651033787

precision\_macro: 0.6501710199620426

recall\_micro: 0.6555723651033787

recall\_macro: 0.5633031102881098

f1\_micro: 0.6555723651033787

f1\_macro: 0.5803509685724422

precision recall f1-score support

10000 0.64 0.80 0.71 241

10001 0.56 0.59 0.57 139

10002 0.66 0.84 0.74 129

10003 0.60 0.67 0.63 356

10004 0.67 0.59 0.62 17

10005 0.69 0.73 0.71 194

10006 0.59 0.49 0.54 186

10007 0.50 0.14 0.22 7

10008 0.77 0.78 0.78 162

10009 0.61 0.88 0.72 16

10010 0.55 0.68 0.61 53

10011 0.67 0.31 0.42 13

10012 0.62 0.40 0.49 25

10013 0.78 0.63 0.70 104

10014 0.53 0.28 0.36 29

10015 0.88 0.78 0.83 118

10016 0.75 0.21 0.32 58

10017 0.70 0.34 0.46 47

10018 0.56 0.61 0.58 38

10019 0.68 0.53 0.59 51

micro avg 0.66 0.66 0.66 1983

macro avg 0.65 0.56 0.58 1983

weighted avg 0.66 0.66 0.65 1983

选取频率最高的200个词做词汇表时的训练结果，采用的特征是tfidf

accuracy: 0.5688350983358548

precision\_micro: 0.5688350983358548

precision\_macro: 0.5339808133887667

recall\_micro: 0.5688350983358548

recall\_macro: 0.4110502895718648

f1\_micro: 0.5688350983358548

f1\_macro: 0.4349599502981462

precision recall f1-score support

10000 0.49 0.76 0.60 241

10001 0.48 0.59 0.53 139

10002 0.61 0.69 0.65 129

10003 0.54 0.62 0.58 356

10004 0.56 0.53 0.55 17

10005 0.71 0.68 0.69 194

10006 0.41 0.42 0.41 186

10007 0.50 0.14 0.22 7

10008 0.72 0.67 0.69 162

10009 0.60 0.19 0.29 16

10010 0.66 0.43 0.52 53

10011 0.00 0.00 0.00 13

10012 0.00 0.00 0.00 25

10013 0.51 0.53 0.52 104

10014 0.62 0.17 0.27 29

10015 0.87 0.78 0.82 118

10016 0.60 0.10 0.18 58

10017 0.52 0.26 0.34 47

10018 0.57 0.42 0.48 38

10019 0.71 0.24 0.35 51

micro avg 0.57 0.57 0.57 1983

macro avg 0.53 0.41 0.43 1983

weighted avg 0.57 0.57 0.55 1983

选取频率最高的300个词做词汇表时的训练结果，采用的特征是tfidf

accuracy: 0.5975794251134644

precision\_micro: 0.5975794251134644

precision\_macro: 0.564731647251428

recall\_micro: 0.5975794251134644

recall\_macro: 0.4604470654022178

f1\_micro: 0.5975794251134644

f1\_macro: 0.48639271111753163

precision recall f1-score support

10000 0.55 0.72 0.62 241

10001 0.57 0.55 0.56 139

10002 0.61 0.71 0.65 129

10003 0.54 0.66 0.59 356

10004 0.71 0.59 0.65 17

10005 0.66 0.70 0.68 194

10006 0.46 0.52 0.49 186

10007 0.00 0.00 0.00 7

10008 0.77 0.75 0.76 162

10009 0.77 0.62 0.69 16

10010 0.67 0.49 0.57 53

10011 0.00 0.00 0.00 13

10012 0.64 0.28 0.39 25

10013 0.58 0.47 0.52 104

10014 0.67 0.14 0.23 29

10015 0.87 0.81 0.84 118

10016 0.57 0.14 0.22 58

10017 0.55 0.26 0.35 47

10018 0.46 0.45 0.45 38

10019 0.66 0.37 0.48 51

micro avg 0.60 0.60 0.60 1983

macro avg 0.56 0.46 0.49 1983

weighted avg 0.60 0.60 0.59 1983

综上所述，可以看出选取频率最高的500个词的训练结果最好

3.与vader情感分析的baseline进行比较：

Vader情感分析的结果：

accuracy: 0.0015128593040847202

precision\_micro: 0.0015128593040847202

precision\_macro: 0.15

recall\_micro: 0.0015128593040847202

recall\_macro: 0.0005836575875486381

f1\_micro: 0.0015128593040847202

f1\_macro: 0.0011627906976744186

precision recall f1-score support

negative 0.60 0.00 0.00 1285

neural 0.00 0.00 0.00 0

neutral 0.00 0.00 0.00 548

positive 0.00 0.00 0.00 150

avg / total 0.39 0.00 0.00 1983

4.决策树情感分析时，使用频率最高的300个词做词汇库，采用tfidf作为特征。

去掉了stopwords的结果如下：

accuracy: 0.9011598587997983

precision\_micro: 0.9011598587997983

precision\_macro: 0.8796556768064345

recall\_micro: 0.9011598587997983

recall\_macro: 0.814687848119641

f1\_micro: 0.9011598587997982

f1\_macro: 0.8407936976622331

precision recall f1-score support

negative 0.93 0.95 0.94 1285

neutral 0.84 0.86 0.85 548

positive 0.87 0.63 0.73 150

micro avg 0.90 0.90 0.90 1983

macro avg 0.88 0.81 0.84 1983

weighted avg 0.90 0.90 0.90 1983

没去掉stopwords的结果如下

accuracy: 0.8935955622793746

precision\_micro: 0.8935955622793746

precision\_macro: 0.8402497679319642

recall\_micro: 0.8935955622793746

recall\_macro: 0.7920848204846614

f1\_micro: 0.8935955622793746

f1\_macro: 0.811654562806574

precision recall f1-score support

negative 0.94 0.95 0.94 1285

neutral 0.82 0.86 0.84 548

positive 0.76 0.57 0.65 150

micro avg 0.89 0.89 0.89 1983

macro avg 0.84 0.79 0.81 1983

weighted avg 0.89 0.89 0.89 1983

可以看出不去掉stopwords准确度、精确度、召回率和F1分数等都有略微的下降。

没提取词干（stemmer）时的结果：

accuracy: 0.8890569843671206

precision\_micro: 0.8890569843671206

precision\_macro: 0.8636160595147345

recall\_micro: 0.8890569843671206

recall\_macro: 0.788745901458276

f1\_micro: 0.8890569843671206

f1\_macro: 0.8147711191838818

precision recall f1-score support

negative 0.93 0.93 0.93 1285

neutral 0.80 0.88 0.84 548

positive 0.86 0.55 0.67 150

micro avg 0.89 0.89 0.89 1983

macro avg 0.86 0.79 0.81 1983

weighted avg 0.89 0.89 0.89 1983

可以看出没提取词干准确度、精确度、召回率和F1分数等都有略微的下降。

5.去中立的影响

决策树去中立（N=200）

accuracy: 0.9581881533101045

precision\_micro: 0.9581881533101045

precision\_macro: 0.9232140329202219

recall\_micro: 0.9581881533101045

recall\_macro: 0.8412191958495461

f1\_micro: 0.9581881533101045

f1\_macro: 0.8765301088714766

precision recall f1-score support

negative 0.97 0.99 0.98 1285

positive 0.88 0.69 0.78 150

micro avg 0.96 0.96 0.96 1435

macro avg 0.92 0.84 0.88 1435

weighted avg 0.96 0.96 0.96 1435伯努利贝叶斯去中立

accuracy: 0.8968641114982578

precision\_micro: 0.8968641114982578

precision\_macro: 0.9483600837404047

recall\_micro: 0.8968641114982578

recall\_macro: 0.5066666666666667

f1\_micro: 0.8968641114982578

f1\_macro: 0.4859319933387553

precision recall f1-score support

negative 0.90 1.00 0.95 1285

positive 1.00 0.01 0.03 150

micro avg 0.90 0.90 0.90 1435

macro avg 0.95 0.51 0.49 1435

weighted avg 0.91 0.90 0.85 1435

多项式贝叶斯去中立：

accuracy: 0.9609756097560975

precision\_micro: 0.9609756097560975

precision\_macro: 0.9608872806436903

recall\_micro: 0.9609756097560975

recall\_macro: 0.8251102464332036

f1\_micro: 0.9609756097560975

f1\_macro: 0.8781937017231135

precision recall f1-score support

negative 0.96 1.00 0.98 1285

positive 0.96 0.65 0.78 150

micro avg 0.96 0.96 0.96 1435

macro avg 0.96 0.83 0.88 1435

weighted avg 0.96 0.96 0.96 1435

6.

决策树模型，采用tfidf特征：

情感分析：

accuracy: 0.9011598587997983

precision\_micro: 0.9011598587997983

precision\_macro: 0.8796556768064345

recall\_micro: 0.9011598587997983

recall\_macro: 0.814687848119641

f1\_micro: 0.9011598587997982

f1\_macro: 0.8407936976622331

precision recall f1-score support

negative 0.93 0.95 0.94 1285

neutral 0.84 0.86 0.85 548

positive 0.87 0.63 0.73 150

micro avg 0.90 0.90 0.90 1983

macro avg 0.88 0.81 0.84 1983

weighted avg 0.90 0.90 0.90 1983

伯努利贝叶斯模型，采用词袋特征

情感分析

accuracy: 0.8472012102874432

precision\_micro: 0.8472012102874432

precision\_macro: 0.5746493751599827

recall\_micro: 0.8472012102874432

recall\_macro: 0.5777874028420764

f1\_micro: 0.8472012102874432

f1\_macro: 0.571839167911718

precision recall f1-score support

negative 0.84 0.99 0.91 1285

neutral 0.89 0.74 0.81 548

positive 0.00 0.00 0.00 150

micro avg 0.85 0.85 0.85 1983

macro avg 0.57 0.58 0.57 1983

weighted avg 0.79 0.85 0.81 1983

话题分类：

多项式贝叶斯模型，采用词袋特征

情感分析

accuracy: 0.9147755925365607

precision\_micro: 0.9147755925365607

precision\_macro: 0.9293915799974624

recall\_micro: 0.9147755925365607

recall\_macro: 0.8106125643380322

f1\_micro: 0.9147755925365607

f1\_macro: 0.8545915178765489

precision recall f1-score support

negative 0.91 0.97 0.94 1285

neutral 0.91 0.86 0.89 548

positive 0.97 0.59 0.74 150

micro avg 0.91 0.91 0.91 1983

macro avg 0.93 0.81 0.85 1983

weighted avg 0.92 0.91 0.91 1983

采用决策树、伯努利贝叶斯和多项式贝叶斯模型融合，采用软投票机制，效果如下：

accuracy: 0.9152798789712556

precision\_micro: 0.9152798789712556

precision\_macro: 0.9457001539297356

recall\_micro: 0.9152798789712556

recall\_macro: 0.7773736513075886

f1\_micro: 0.9152798789712556

f1\_macro: 0.8307645980567141

precision recall f1-score support

negative 0.90 0.99 0.95 1285

neutral 0.93 0.85 0.89 548

positive 1.00 0.49 0.65 150

micro avg 0.92 0.92 0.92 1983

macro avg 0.95 0.78 0.83 1983

weighted avg 0.92 0.92 0.91 1983

显然模型融合后的效果最好，相比于标准模型和baseline的模型，该方法采用了多个模型融合。

话题分类：

决策树模型，采用bag\_of\_words做特征，词汇表500

accuracy: 0.6555723651033787

precision\_micro: 0.6555723651033787

precision\_macro: 0.6501710199620426

recall\_micro: 0.6555723651033787

recall\_macro: 0.5633031102881098

f1\_micro: 0.6555723651033787

f1\_macro: 0.5803509685724422

precision recall f1-score support

10000 0.64 0.80 0.71 241

10001 0.56 0.59 0.57 139

10002 0.66 0.84 0.74 129

10003 0.60 0.67 0.63 356

10004 0.67 0.59 0.62 17

10005 0.69 0.73 0.71 194

10006 0.59 0.49 0.54 186

10007 0.50 0.14 0.22 7

10008 0.77 0.78 0.78 162

10009 0.61 0.88 0.72 16

10010 0.55 0.68 0.61 53

10011 0.67 0.31 0.42 13

10012 0.62 0.40 0.49 25

10013 0.78 0.63 0.70 104

10014 0.53 0.28 0.36 29

10015 0.88 0.78 0.83 118

10016 0.75 0.21 0.32 58

10017 0.70 0.34 0.46 47

10018 0.56 0.61 0.58 38

10019 0.68 0.53 0.59 51

micro avg 0.66 0.66 0.66 1983

macro avg 0.65 0.56 0.58 1983

weighted avg 0.66 0.66 0.65 1983

伯努利贝叶斯模型：

accuracy: 0.3625819465456379

precision\_micro: 0.3625819465456379

precision\_macro: 0.2905338075702433

recall\_micro: 0.3625819465456379

recall\_macro: 0.13838729691768364

f1\_micro: 0.36258194654563786

f1\_macro: 0.1284800103227357

precision recall f1-score support

10000 0.75 0.80 0.77 241

10001 0.00 0.00 0.00 139

10002 0.40 0.02 0.03 129

10003 0.23 0.99 0.38 356

10004 0.00 0.00 0.00 17

10005 0.76 0.31 0.44 194

10006 0.84 0.26 0.40 186

10007 0.00 0.00 0.00 7

10008 0.82 0.37 0.51 162

10009 0.00 0.00 0.00 16

10010 0.00 0.00 0.00 53

10011 0.00 0.00 0.00 13

10012 0.00 0.00 0.00 25

10013 1.00 0.01 0.02 104

10014 0.00 0.00 0.00 29

10015 1.00 0.01 0.02 118

10016 0.00 0.00 0.00 58

10017 0.00 0.00 0.00 47

10018 0.00 0.00 0.00 38

10019 0.00 0.00 0.00 51

micro avg 0.36 0.36 0.36 1983

macro avg 0.29 0.14 0.13 1983

weighted avg 0.49 0.36 0.29 1983

多项式贝叶斯：

accuracy: 0.7518910741301059

precision\_micro: 0.7518910741301059

precision\_macro: 0.8270547109058273

recall\_micro: 0.7518910741301059

recall\_macro: 0.5304473301064708

f1\_micro: 0.7518910741301059

f1\_macro: 0.586104033216541

precision recall f1-score support

10000 0.83 0.90 0.86 241

10001 0.82 0.63 0.72 139

10002 0.77 0.74 0.75 129

10003 0.58 0.87 0.70 356

10004 1.00 0.24 0.38 17

10005 0.73 0.89 0.80 194

10006 0.83 0.80 0.81 186

10007 1.00 0.14 0.25 7

10008 0.79 0.90 0.84 162

10009 0.00 0.00 0.00 16

10010 0.88 0.28 0.43 53

10011 1.00 0.08 0.14 13

10012 1.00 0.24 0.39 25

10013 0.86 0.86 0.86 104

10014 1.00 0.45 0.62 29

10015 0.87 0.88 0.88 118

10016 0.79 0.38 0.51 58

10017 1.00 0.34 0.51 47

10018 0.96 0.61 0.74 38

10019 0.83 0.39 0.53 51

micro avg 0.75 0.75 0.75 1983

macro avg 0.83 0.53 0.59 1983

weighted avg 0.78 0.75 0.74 1983

采用bagging采样方法训练10个多项式贝叶斯模型，进行模型融合，采用软投票机制，效果如下：

accuracy: 0.7473524962178517

precision\_micro: 0.7473524962178517

precision\_macro: 0.8289588810621377

recall\_micro: 0.7473524962178517

recall\_macro: 0.5205664560646708

f1\_micro: 0.7473524962178517

f1\_macro: 0.5761980380158894

precision recall f1-score support

10000 0.83 0.90 0.86 241

10001 0.83 0.61 0.70 139

10002 0.76 0.71 0.74 129

10003 0.57 0.88 0.69 356

10004 1.00 0.18 0.30 17

10005 0.74 0.88 0.80 194

10006 0.83 0.80 0.81 186

10007 1.00 0.14 0.25 7

10008 0.79 0.91 0.84 162

10009 0.00 0.00 0.00 16

10010 0.87 0.25 0.38 53

10011 1.00 0.08 0.14 13

10012 1.00 0.24 0.39 25

10013 0.85 0.85 0.85 104

10014 1.00 0.41 0.59 29

10015 0.88 0.87 0.88 118

10016 0.81 0.36 0.50 58

10017 1.00 0.34 0.51 47

10018 1.00 0.61 0.75 38

10019 0.83 0.39 0.53 51

micro avg 0.75 0.75 0.75 1983

macro avg 0.83 0.52 0.58 1983

weighted avg 0.78 0.75 0.73 1983