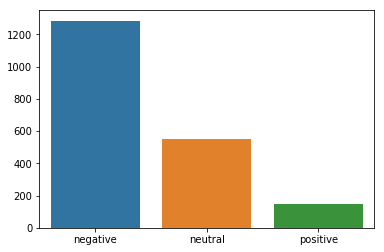
一、描述性统计分析

1.情感的柱状图



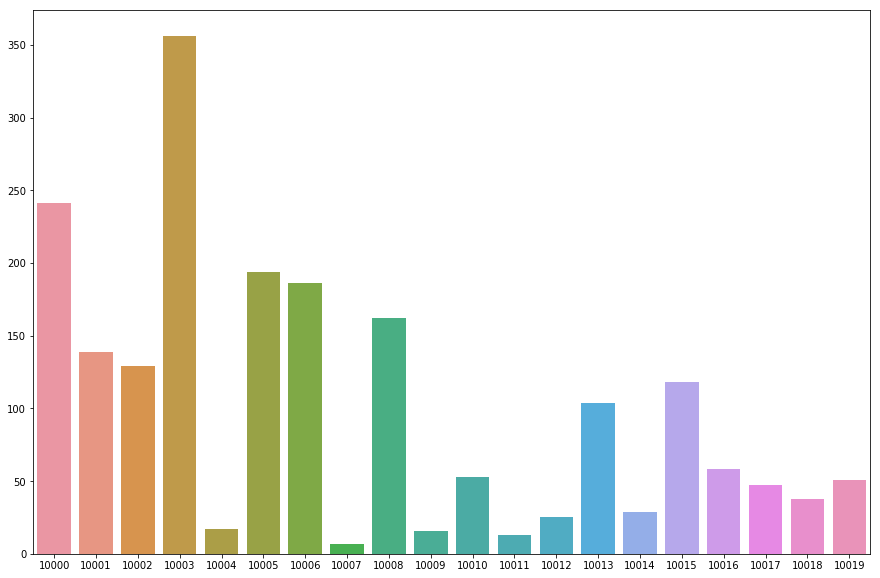
情感的统计

negative 1285

neutral 548

positive 150

主题的柱状图



主题的统计

10003 356

10000 241

10005 194

10006 186

10008 162

10001 139

10002 129

10015 118

10013 104

10016 58

10010 53

10019 51

10017 47

10018 38

10014 29

10012 25

10004 17

10009 16

10011 13

10007 7

二、

调整vocabulary的数量200和300之间的比较

当用决策树做情感分析时使用频率最高的前200个词做特征的模型结果如下：

accuracy: 0.8729198184568835

precision\_micro: 0.8729198184568835

precision\_macro: 0.8433979859599834

recall\_micro: 0.8729198184568835

recall\_macro: 0.7869465509134344

f1\_micro: 0.8729198184568835

f1\_macro: 0.8104118815525316

classification\_report precision recall f1-score support

negative 0.90 0.92 0.91 1285

neutral 0.81 0.82 0.82 548

positive 0.81 0.61 0.70 150

micro avg 0.87 0.87 0.87 1983

macro avg 0.84 0.79 0.81 1983

weighted avg 0.87 0.87 0.87 1983

当用决策树做情感分析时使用频率最高的前300个词做特征的模型结果如下：

accuracy: 0.875441250630358

precision\_micro: 0.875441250630358

precision\_macro: 0.8436722977124337

recall\_micro: 0.875441250630358

recall\_macro: 0.7984065311583843

f1\_micro: 0.8754412506303582

f1\_macro: 0.8180113649133552

classification\_report precision recall f1-score support

negative 0.91 0.92 0.92 1285

neutral 0.82 0.82 0.82 548

positive 0.81 0.65 0.72 150

micro avg 0.88 0.88 0.88 1983

macro avg 0.84 0.80 0.82 1983

weighted avg 0.87 0.88 0.87 1983

300的词汇量效果比200的稍微好一点，但是随着实验的进行，如果用大量词汇，将会导致过拟合。

三、baseline vader模型的实验效果

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

效果显然是我们训练的模型好。我把vader输出的复杂感情看作中立了。

四、没有提取词干的决策树用来做情感分析的结果如下：

accuracy: 0.8603126575895108

precision\_micro: 0.8603126575895108

precision\_macro: 0.8126509330116677

recall\_micro: 0.8603126575895108

recall\_macro: 0.734488498837103

f1\_micro: 0.8603126575895108

f1\_macro: 0.7628167517253038

classification\_report precision recall f1-score support

negative 0.89 0.93 0.91 1285

neutral 0.80 0.81 0.80 548

positive 0.74 0.47 0.57 150

micro avg 0.86 0.86 0.86 1983

macro avg 0.81 0.73 0.76 1983

weighted avg 0.86 0.86 0.86 1983

很明显，效果是稍微下降的，因为词干在英语中非常重要，不同的单词形式可能表达的是一样的意思。

五、去掉中立情感以后的决策树效果：

accuracy: 0.9519163763066202

precision\_micro: 0.9519163763066202

precision\_macro: 0.9011711067405686

recall\_micro: 0.9519163763066202

recall\_macro: 0.8259403372243839

f1\_micro: 0.9519163763066202

f1\_macro: 0.8584830827148267

classification\_report precision recall f1-score support

negative 0.96 0.99 0.97 1285

positive 0.84 0.67 0.74 150

micro avg 0.95 0.95 0.95 1435

macro avg 0.90 0.83 0.86 1435

weighted avg 0.95 0.95 0.95 1435

显然中立比较难区分，去掉中立之后的效果好。

六、以下是几个模型的结果：

DT

accuracy: 0.5668179525970751

precision\_micro: 0.5668179525970751

precision\_macro: 0.5185071259687877

recall\_micro: 0.5668179525970751

recall\_macro: 0.40508149368227364

f1\_micro: 0.5668179525970751

f1\_macro: 0.41782372320241434

classification\_report precision recall f1-score support

10000 0.55 0.74 0.63 241

10001 0.53 0.50 0.52 139

10002 0.58 0.74 0.65 129

10003 0.52 0.62 0.57 356

10004 0.58 0.41 0.48 17

10005 0.64 0.69 0.66 194

10006 0.43 0.54 0.48 186

10007 0.00 0.00 0.00 7

10008 0.63 0.66 0.65 162

10009 0.39 0.56 0.46 16

10010 0.59 0.49 0.54 53

10011 0.00 0.00 0.00 13

10012 0.50 0.04 0.07 25

10013 0.61 0.40 0.49 104

10014 0.75 0.10 0.18 29

10015 0.88 0.78 0.83 118

10016 0.48 0.26 0.34 58

10017 0.65 0.28 0.39 47

10018 0.47 0.21 0.29 38

10019 0.57 0.08 0.14 51

micro avg 0.57 0.57 0.57 1983

macro avg 0.52 0.41 0.42 1983

weighted avg 0.57 0.57 0.55 1983

BNB

accuracy: 0.8466969238527483

precision\_micro: 0.8466969238527483

precision\_macro: 0.8394764357355845

recall\_micro: 0.8466969238527483

recall\_macro: 0.5846815366020682

f1\_micro: 0.8466969238527483

f1\_macro: 0.5872770349245341

classification\_report precision recall f1-score support

negative 0.84 0.99 0.91 1285

neutral 0.88 0.74 0.80 548

positive 0.80 0.03 0.05 150

micro avg 0.85 0.85 0.85 1983

macro avg 0.84 0.58 0.59 1983

weighted avg 0.85 0.85 0.81 1983

accuracy: 0.37367624810892586

precision\_micro: 0.37367624810892586

precision\_macro: 0.31815859844331973

recall\_micro: 0.37367624810892586

recall\_macro: 0.14406480317899167

f1\_micro: 0.37367624810892586

f1\_macro: 0.13459092955259025

classification\_report precision recall f1-score support

10000 0.73 0.81 0.77 241

10001 0.50 0.01 0.01 139

10002 0.40 0.02 0.03 129

10003 0.24 1.00 0.39 356

10004 0.00 0.00 0.00 17

10005 0.80 0.32 0.46 194

10006 0.86 0.32 0.47 186

10007 0.00 0.00 0.00 7

10008 0.84 0.38 0.53 162

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 0.37 0.37 0.37 1983

macro avg 0.32 0.14 0.13 1983

weighted avg 0.53 0.37 0.30 1983

MNB

accuracy: 0.9067070095814422

precision\_micro: 0.9067070095814422

precision\_macro: 0.9322025941613523

recall\_micro: 0.9067070095814422

recall\_macro: 0.7767590357263452

f1\_micro: 0.9067070095814422

f1\_macro: 0.82779100836526

classification\_report precision recall f1-score support

negative 0.90 0.98 0.94 1285

neutral 0.91 0.84 0.87 548

positive 0.99 0.51 0.67 150

micro avg 0.91 0.91 0.91 1983

macro avg 0.93 0.78 0.83 1983

weighted avg 0.91 0.91 0.90 1983

MNB

accuracy: 0.697428139183056

precision\_micro: 0.697428139183056

precision\_macro: 0.7242338509748872

recall\_micro: 0.697428139183056

recall\_macro: 0.4399538854335182

f1\_micro: 0.697428139183056

f1\_macro: 0.48136053715740745

classification\_report precision recall f1-score support

10000 0.83 0.88 0.85 241

10001 0.83 0.50 0.63 139

10002 0.80 0.61 0.69 129

10003 0.48 0.89 0.63 356

10004 1.00 0.06 0.11 17

10005 0.71 0.87 0.78 194

10006 0.79 0.76 0.78 186

10007 0.00 0.00 0.00 7

10008 0.79 0.86 0.82 162

10009 0.00 0.00 0.00 16

10010 0.92 0.23 0.36 53

10011 0.00 0.00 0.00 13

10012 1.00 0.08 0.15 25

10013 0.84 0.78 0.81 104

10014 1.00 0.28 0.43 29

10015 0.86 0.80 0.83 118

10016 0.81 0.22 0.35 58

10017 1.00 0.19 0.32 47

10018 1.00 0.53 0.69 38

10019 0.81 0.25 0.39 51

micro avg 0.70 0.70 0.70 1983

macro avg 0.72 0.44 0.48 1983

weighted avg 0.75 0.70 0.67 1983

LinearSVC

sentiment

accuracy: 0.9949571356530509

precision\_micro: 0.9949571356530509

precision\_macro: 0.9901578760250344

recall\_micro: 0.9949571356530509

recall\_macro: 0.9913871137745716

f1\_micro: 0.9949571356530509

f1\_macro: 0.9907677226406332

classification\_report precision recall f1-score support

negative 1.00 1.00 1.00 1285

neutral 0.99 0.99 0.99 548

positive 0.98 0.99 0.98 150

micro avg 0.99 0.99 0.99 1983

macro avg 0.99 0.99 0.99 1983

weighted avg 0.99 0.99 0.99 1983

linearSVC

topics

accuracy: 0.848714069591528

precision\_micro: 0.848714069591528

precision\_macro: 0.8534853008198551

recall\_micro: 0.848714069591528

recall\_macro: 0.8257101608161548

f1\_micro: 0.848714069591528

f1\_macro: 0.8374987921296881

classification\_report precision recall f1-score support

10000 0.92 0.91 0.91 241

10001 0.83 0.79 0.81 139

10002 0.77 0.87 0.81 129

10003 0.84 0.80 0.82 356

10004 0.93 0.82 0.87 17

10005 0.82 0.88 0.85 194

10006 0.89 0.84 0.86 186

10007 1.00 0.86 0.92 7

10008 0.85 0.93 0.89 162

10009 0.85 0.69 0.76 16

10010 0.72 0.72 0.72 53

10011 0.83 0.77 0.80 13

10012 0.74 0.80 0.77 25

10013 0.84 0.91 0.88 104

10014 0.96 0.86 0.91 29

10015 0.89 0.93 0.91 118

10016 0.78 0.66 0.71 58

10017 0.90 0.79 0.84 47

10018 0.90 0.92 0.91 38

10019 0.81 0.76 0.79 51

micro avg 0.85 0.85 0.85 1983

macro avg 0.85 0.83 0.84 1983

weighted avg 0.85 0.85 0.85 1983

linearSVC采用svm分类器，达到了最好的效果。