# 数据下载

从tweet上搜索包含“trade war”或#tradewar的推文，时间范围20180320-20191015。得到148W条。从中过滤含有china的，还有1078418 （all\_tweets\_china.jl）条。

# 数据清洗

都小写；

去掉http,pic等超链接；

去掉hashtag的#和@someone的@

去掉所有cashtag, eg: $WTC

使用nltk.stem 对单词做词形还原lemmatizer from nltk.stem import WordNetLemmatizer

去掉内容重复的，一般为转发其他tweet；

去掉不足10个字的，认为太短的句子不利于情感分类，

还剩507053. （all\_tweets\_cleaned.jl）

张新的标注数据，单词出现次数

CENTRAL前20的词：('china', 3913)|('trade', 3610)|('war', 3382)|('us', 2067)|('trump', 1163)|('tariff', 712)|('tradewar', 483)|('$', 447)|('escal', 304)|('market', 261)|('say', 249)|('new', 229)|('economi', 215)|('chines', 193)|('could', 192)|('billion', 176)|('stock', 170)|('news', 166)|('import', 162)|('good', 161)

POSITIVE前20的词：('china', 1099)|('trade', 996)|('war', 903)|('us', 550)|('trump', 321)|('tariff', 136)|('tradewar', 130)|('$', 109)|('market', 82)|('win', 81)|('say', 77)|('economi', 69)|('chines', 67)|('escal', 63)|('&', 62)|('new', 62)|('talk', 59)|('end', 51)|('good', 49)|('could', 49)

NEGATIVE前20的词：('china', 4587)|('trade', 4473)|('war', 4192)|('us', 2247)|('trump', 1547)|('tariff', 596)|('tradewar', 447)|('market', 404)|('say', 339)|('economi', 330)|('$', 324)|('farmer', 308)|('escal', 304)|('stock', 277)|('could', 266)|('american', 256)|('start', 226)|('presid', 221)|('chines', 212)|('&', 202)

# LDA

需要去掉停用词，包括nltk english的所有停用词，以及自定义停用词china, us, trade, war,america, tradewar。

模型输出结果如下

"C:\Program Files\python\python.exe" D:/code/github/zxx/utils/nlp\_models/lda.py

k=5

begin to train

[(0,

'0.013\*"economy" + 0.012\*"economic" + 0.011\*"global" + 0.009\*"say" + '

'0.009\*"world" + 0.009\*"talk" + 0.008\*"state" + 0.006\*"impact" + '

'0.005\*"recession" + 0.005\*"united" + 0.005\*"damage" + 0.005\*"deal" + '

'0.005\*"currency" + 0.005\*"political" + 0.005\*"may" + 0.004\*"agreement" + '

'0.004\*"show" + 0.004\*"new" + 0.004\*"south" + 0.004\*"negotiation"'),

(1,

'0.050\*"market" + 0.020\*"stock" + 0.015\*"hong" + 0.014\*"kong" + '

'0.012\*"point" + 0.012\*"low" + 0.011\*"slow" + 0.011\*"end" + 0.010\*"wall" + '

'0.010\*"fall" + 0.009\*"global" + 0.009\*"investor" + 0.009\*"news" + '

'0.008\*"growth" + 0.007\*"amid" + 0.007\*"gain" + 0.007\*"open" + 0.006\*"high" '

'+ 0.006\*"fear" + 0.006\*"future"'),

(2,

'0.037\*"tariff" + 0.022\*"trump" + 0.018\*"dollar" + 0.014\*"chinese" + '

'0.012\*"billion" + 0.010\*"new" + 0.010\*"export" + 0.010\*"escalate" + '

'0.009\*"year" + 0.009\*"good" + 0.009\*"hit" + 0.008\*"say" + 0.007\*"price" + '

'0.007\*"import" + 0.007\*"company" + 0.007\*"oil" + 0.007\*"president" + '

'0.006\*"amid" + 0.005\*"economy" + 0.005\*"business"'),

(3,

'0.042\*"trump" + 0.017\*"deal" + 0.013\*"get" + 0.012\*"go" + 0.011\*"say" + '

'0.011\*"president" + 0.010\*"win" + 0.009\*"want" + 0.009\*"start" + '

'0.009\*"think" + 0.008\*"end" + 0.008\*"one" + 0.008\*"like" + 0.008\*"make" + '

'0.007\*"know" + 0.007\*"dont" + 0.006\*"back" + 0.006\*"need" + 0.006\*"take" + '

'0.005\*"good"'),

(4,

'0.030\*"farmer" + 0.023\*"trump" + 0.014\*"pay" + 0.012\*"make" + 0.012\*"lose" '

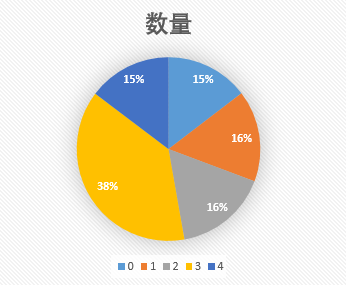
'+ 0.010\*"hurt" + 0.010\*"job" + 0.008\*"tax" + 0.008\*"money" + 0.008\*"get" + '

'0.007\*"buy" + 0.007\*"india" + 0.007\*"business" + 0.007\*"go" + 0.007\*"cant" '

'+ 0.006\*"many" + 0.006\*"company" + 0.006\*"win" + 0.006\*"biden" + '

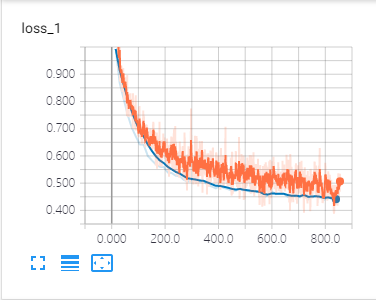
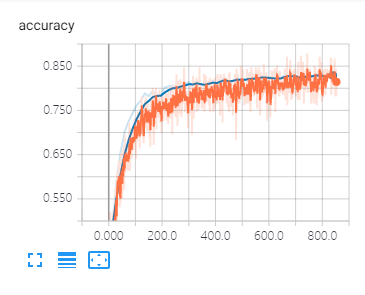
'0.006\*"cost"')]

|  |  |  |
| --- | --- | --- |
| 类别 | 解释 | 关键词 |
| 0 | 世界经济/中美对世界经济的影响 | economy, economic,global,world,impact,recession |
| 1 | 金融以及股票市场 | market,stock,hong kong, point,fall,investor,growth |
| 2 | 关税以及进出口 | tariff,trump,dollar,billion,export,escalate,good,price,  import,company,business |
| 4 | 农业，工作，美国内部 | trump,farmer,lose,hurt,job,win |
| 3 | 关心，希望和解 | trump, deal,say,win,end, |

主题对应tweet数量统计

|  |  |
| --- | --- |
| 类别 | 数量 |
| 0 | 57912 |
| 1 | 63875 |
| 2 | 65492 |
| 3 | 150970 |
| 4 | 58311 |

然后我们标注了4W多条推文，之后使用近几年提出的一个文本分类模型TextCNN做情感分类。下两个图是运行过程，大概准确率83%，还可以。



得到模型后，把模型在在标注的数据上预测，得到所有数据的情感。

分类统计如下

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 类别 | 总量 | CENTRAL | NEGATIVE | POSITIVE | 中性占比 | 负向占比 | 正向占比 |
| 0 | 57912 | 33961 | 17846 | 6105 | 58.64% | 30.82% | 10.54% |
| 1 | 63875 | 37068 | 21358 | 5449 | 58.03% | 33.44% | 8.53% |
| 2 | 65492 | 44084 | 14687 | 6721 | 67.31% | 22.43% | 10.26% |
| 3 | 150970 | 75751 | 51594 | 23625 | 50.18% | 34.18% | 15.65% |
| 4 | 58311 | 33263 | 17238 | 7810 | 57.04% | 29.56% | 13.39% |

大概可以看出，大部分推文都是陈述性的中性。具有情感态度的，几个类别都是负向多于正向。但是正向占比可以反应出信心的多少，类别1股市最低，类别3农业或者内生工作最高。

然后就是挑几条典型的推文，说一下。