# 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100W条。从中过滤含有china的，还有597709 （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

过滤掉为空的，还剩597333 （all\_tweets\_cleaned.jl）

去重，去掉不足7个字的，还剩396560.

# LDA

## 3类

[(0,

'0.017\*"global" + 0.014\*"economi" + 0.012\*"talk" + 0.011\*"say" + '

'0.010\*"market" + 0.009\*"amid" + 0.008\*"econom" + 0.008\*"end" + '

'0.008\*"growth" + 0.007\*"news"'),

(1,

'0.031\*"trump" + 0.011\*"deal" + 0.009\*"farmer" + 0.006\*"get" + 0.006\*"win" + '

'0.005\*"make" + 0.005\*"like" + 0.005\*"go" + 0.005\*"one" + 0.005\*"start"'),

(2,

'0.063\*"tariff" + 0.037\*"trump" + 0.022\*"escal" + 0.020\*"stock" + '

'0.019\*"billion" + 0.017\*"new" + 0.016\*"hit" + 0.015\*"market" + '

'0.015\*"import" + 0.014\*"donald"')]

Perplexity: -8.334492397433264

## 4类

[(0,

'0.073\*"tariff" + 0.043\*"trump" + 0.022\*"billion" + 0.022\*"new" + '

'0.017\*"chines" + 0.017\*"import" + 0.017\*"escal" + 0.015\*"hit" + '

'0.012\*"good" + 0.011\*"export"'),

(1,

'0.023\*"economi" + 0.020\*"global" + 0.019\*"could" + 0.018\*"compani" + '

'0.014\*"state" + 0.014\*"say" + 0.014\*"busi" + 0.012\*"econom" + '

'0.011\*"impact" + 0.011\*"damag"'),

(2,

'0.041\*"market" + 0.019\*"amid" + 0.019\*"stock" + 0.014\*"year" + 0.012\*"kong" '

'+ 0.012\*"hong" + 0.011\*"end" + 0.011\*"oil" + 0.010\*"price" + 0.010\*"news"'),

(3,

'0.031\*"trump" + 0.012\*"deal" + 0.009\*"farmer" + 0.006\*"get" + 0.006\*"win" + '

'0.006\*"make" + 0.006\*"like" + 0.005\*"go" + 0.005\*"one" + 0.005\*"start"')]

Perplexity: -8.410683042263114

## 5类

[(0,

'0.061\*"tariff" + 0.032\*"trump" + 0.022\*"escal" + 0.019\*"billion" + '

'0.016\*"export" + 0.015\*"hit" + 0.014\*"chines" + 0.014\*"import" + '

'0.014\*"new" + 0.011\*"good"'),

(1,

'0.039\*"end" + 0.035\*"talk" + 0.015\*"reach" + 0.015\*"impact" + 0.012\*"say" + '

'0.011\*"elect" + 0.008\*"warn" + 0.008\*"currenc" + 0.008\*"firm" + '

'0.007\*"econom"'),

(2,

'0.043\*"market" + 0.027\*"global" + 0.021\*"amid" + 0.020\*"stock" + '

'0.016\*"economi" + 0.013\*"growth" + 0.012\*"year" + 0.011\*"investig" + '

'0.011\*"ever" + 0.010\*"point"'),

(3,

'0.037\*"trump" + 0.013\*"deal" + 0.010\*"farmer" + 0.007\*"get" + 0.007\*"win" + '

'0.006\*"make" + 0.006\*"economi" + 0.006\*"say" + 0.006\*"like" + '

'0.006\*"presid"'),

(4,

'0.034\*"news" + 0.021\*"kong" + 0.021\*"hong" + 0.020\*"lost" + 0.019\*"oil" + '

'0.018\*"agre" + 0.014\*"gold" + 0.012\*"fed" + 0.011\*"set" + 0.011\*"cnbc"')]

Perplexity: -8.550309520911895

第二轮

[(0,

'0.051\*"tariff" + 0.045\*"trump" + 0.016\*"chinese" + 0.015\*"new" + '

'0.015\*"billion" + 0.012\*"$" + 0.011\*"president" + 0.010\*"donald" + '

'0.009\*"state" + 0.009\*"talk" + 0.009\*"export" + 0.008\*"good" + 0.008\*"say" '

'+ 0.007\*"united" + 0.007\*"beijing"'),

(1,

'0.026\*"market" + 0.015\*"global" + 0.014\*"economy" + 0.012\*"stock" + '

'0.010\*"year" + 0.009\*"amid" + 0.009\*"growth" + 0.007\*"oil" + '

'0.007\*"economic" + 0.006\*"price" + 0.006\*"week" + 0.006\*"news" + '

'0.005\*"month" + 0.005\*"brexit" + 0.005\*"india"'),

(2,

'0.017\*"recession" + 0.013\*"say" + 0.009\*"iran" + 0.008\*"economic" + '

'0.007\*"deal" + 0.007\*"economy" + 0.007\*"world" + 0.006\*"talk" + '

'0.006\*"korea" + 0.006\*"end" + 0.005\*"democrat" + 0.005\*"news" + '

'0.005\*"north" + 0.005\*"maybe" + 0.005\*"failed"'),

(3,

'0.031\*"trump" + 0.011\*"deal" + 0.011\*"farmer" + 0.006\*"one" + 0.006\*"like" '

'+ 0.006\*"get" + 0.005\*"country" + 0.005\*"going" + 0.005\*"want" + '

'0.005\*"make" + 0.005\*"economy" + 0.004\*"win" + 0.004\*"think" + '

'0.004\*"president" + 0.004\*"need"')]

[(0,

'0.017\*"economy" + 0.015\*"say" + 0.013\*"state" + 0.013\*"economic" + '

'0.013\*"world" + 0.012\*"trump" + 0.012\*"global" + 0.009\*"president" + '

'0.009\*"united" + 0.008\*"two" + 0.008\*"chinese" + 0.008\*"election" + '

'0.008\*"talk" + 0.007\*"agreement" + 0.007\*"recession"'),

(1,

'0.036\*"market" + 0.017\*"stock" + 0.012\*"amid" + 0.012\*"year" + '

'0.011\*"global" + 0.010\*"kong" + 0.010\*"hong" + 0.008\*"price" + '

'0.008\*"growth" + 0.008\*"point" + 0.007\*"week" + 0.007\*"investor" + '

'0.006\*"rate" + 0.006\*"gold" + 0.006\*"news"'),

(2,

'0.055\*"tariff" + 0.038\*"trump" + 0.017\*"chinese" + 0.016\*"billion" + '

'0.014\*"new" + 0.013\*"$" + 0.011\*"good" + 0.010\*"export" + 0.008\*"company" + '

'0.008\*"import" + 0.007\*"business" + 0.007\*"news" + 0.007\*"president" + '

'0.006\*"ivanka" + 0.006\*"economy"'),

(3,

'0.033\*"trump" + 0.017\*"farmer" + 0.007\*"like" + 0.006\*"one" + 0.006\*"get" + '

'0.006\*"win" + 0.006\*"going" + 0.006\*"country" + 0.006\*"people" + '

'0.006\*"don\'t" + 0.005\*"think" + 0.005\*"started" + 0.005\*"job" + '

'0.005\*"economy" + 0.005\*"know"'),

(4,

'0.030\*"deal" + 0.026\*"trump" + 0.007\*"oil" + 0.006\*"talk" + 0.006\*"end" + '

'0.006\*"want" + 0.005\*"iran" + 0.005\*"losing" + 0.005\*"ongoing" + '

'0.005\*"get" + 0.005\*"2020" + 0.004\*"usa" + 0.004\*"investigate" + '

'0.004\*"paying" + 0.004\*"done"')]

count 396698.000000

mean 0.787715

std 0.756685

min 0.000000

10% 0.000000

20% 0.125000

30% 0.250000

40% 0.375000

50% 0.625000

60% 0.750000

70% 1.000000

80% 1.375000

90% 1.875000

max 9.500000

Name: pos, dtype: float64

count 396698.000000

mean 0.755981

std 0.759188

min 0.000000

10% 0.000000

20% 0.125000

30% 0.250000

40% 0.375000

50% 0.500000

60% 0.750000

70% 1.000000

80% 1.250000

90% 1.750000

max 9.625000

Name: neg, dtype: float64

count 396698.000000

mean 12.734992

std 6.344646

min 0.000000

10% 5.250000

20% 6.750000

30% 8.000000

40% 9.625000

50% 11.625000

60% 14.000000

70% 16.500000

80% 19.125000

90% 22.000000

max 37.500000

