1532385

Exercise 1 Report

The code can be run with or without a POS tagger. If it is used then it will take 4min longer. Hence it is left out by default. If it is to be used then uncomment the 2 lines of code and comment the original lem_text.

Results

Without POS:

Runtime ~ 40s

```
Number of types (vocabulary size): 124045
Number of tokens: 5692756

The top 25 trigrams are:
 [(('one', 'of', 'the'), 2434), (('on', 'share', 'of'), 2095), (('on', 'the', 'stock'), 1567), (('a', 'well', 'a'), 1423),
(('in', 'research', 'report'), 1415), (('in', 'research', 'note'), 1373), (('the', 'united', 'state'), 1223), (('for',
'the', 'quarter'), 1221), (('average', 'price', 'of'), 1193), (('research', 'report', 'on'), 1177), (('research',
'note', 'on'), 1138), (('share', 'of', 'the'), 1132), (('the', 'end', 'of'), 1130), (('in', 'report', 'on'), 1124),
(('earnings', 'per', 'share'), 1121), (('cell', 'phone', 'plan'), 1073), (('phone', 'plan', 'detail'), 1070),
(('according', 'to', 'the'), 1066), (('of', 'the', 'company'), 1057), (('buy', 'rating', 'to'), 1016), (('appeared',
'first', 'on'), 995), (('moving', 'average', 'price'), 995), (('day', 'moving', 'average'), 993), (('price', 'target',
'on'), 981), (('part', 'of', 'the'), 935)]

Number of positive words in corpus: 170754
Number of negative words in corpus: 129731

Number of stories with more positive words:  10826
Number of stories with more negative words:  6394
```

With POS:

Runtime ~ 4min 55s

```
 Number of types (vocabulary size): 122475
 Number of tokens: 5692756

 The top 25 trigrams are:
 [(('one', 'of', 'the'), 2434), (('on', 'share', 'of'), 2095), (('on', 'the', 'stock'), 1567), (('as', 'well', 'a'), 1417),
(('in', 'research', 'report'), 1415), (('in', 'research', 'note'), 1373), (('be', 'able', 'to'), 1267), (('the', 'united',
'state'), 1223), (('for', 'the', 'quarter'), 1221), (('average', 'price', 'of'), 1193), (('research', 'report', 'on'),
1177), (('research', 'note', 'on'), 1138), (('the', 'end', 'of'), 1135), (('share', 'of', 'the'), 1133), (('in', 'report',
'on'), 1124), (('earnings', 'per', 'share'), 1121), (('cell', 'phone', 'plan'), 1073), (('phone', 'plan', 'detail'),
1070), (('accord', 'to', 'the'), 1066), (('buy', 'rating', 'to'), 1016), (('of', 'the', 'company'), 1002), (('appear',
'first', 'on'), 994), (('day', 'move', 'average'), 993), (('price', 'target', 'on'), 981), (('be', 'one', 'of'), 970)]
```

1532385

Number of positive words in corpus: 176423
Number of negative words in corpus: 143129

Number of stories with more positive words: 10436
Number of stories with more negative words: 6893

Comparison with and without POS tag

Using the POS tagger increases runtime by approximately 4mins.

Using the POS tagger results in a larger vocabulary size (more types found) but the number of tokens remains the same.

Also by using the POS tagger the code finds more positive and negative words. This leads to a decrease in the number of stores with more positive words by 390. However, this causes an increase of the number of stories with more negative words by 499.