Natural Language Processing on Stock Market News Data

We consider a dataset consisting two channels of data provided in this dataset.Using 8 years daily news headlines to predict stock market movement.

1. News data: Historical news headlines from Reddit World News Channel . They are ranked by reddit users' votes, and only the top 25 headlines are considered for a single date. (Range: 2008-06-08 to 2016-07-01)
2. Stock data: Dow Jones Industrial Average (DJIA) is used to "prove the concept". (Range: 2008-08-08 to 2016-07-01)

Understanding how NLP works

Through forms of AI, like neural networks and deep learning models, NLP can extract, examine and utilize patterns in stored data, and the value of results provided to the user *improve* with each new search.

The algorithm looks like :

- Read text written in our human languages
- Interpret its meaning through its own complex filters, and
- Translate it back to us providing exactly what we asked for

Thanks to big data and cloud computing, NLP has taken some major leaps forward. What used to be insurmountable mountains of data, is now a casual hike for machine learning. NLP has now gained the ability to understand our ambiguous human languages making natural language search through troves of data a walk in the park.

After reading the data and seeing the data structure we use preprocessing on the data as follows:

The process involved:

- Converting the headline to lowercase letters.
- Splitting the sentence into a list of words.
- Removing punctuation and meaningless words.
- Transforming that list into a table of counts.

What started as a relatively "messy" sentence has now become an neatly organized table.

We will get distinctive words like

```
['the', 'commander', 'of', 'navy', 'air', 'reconnaissance', 'squadron',
'that', 'provides', 'the', 'president', 'and', 'the', 'defense',
'secretary', 'the', 'airborne', 'ability', 'to', 'command', 'the',
'nation', 'nuclear', 'weapons', 'has', 'been', 'relieved', 'of', 'duty']
```

And we also get a count of them, Our resulting table contains counts for 31,675 different words.

Our most positive words don't seem particularly interesting, however there are some negative sounding words within our bottom 10, such as "sanctions," "low," and "hacking." Maybe the saying "no news is good news" is true here?

Advanced Modeling The technique we just used is known as a bag-of-words model. We essentially placed all of our headlines into a "bag" and counted the words as we pulled them out. However, most people would agree that a single word doesn't always have enough meaning by itself. Obviously, we need to consider the rest of the words in the sentence as well.

We will refer to Appendix Section to take a look at charts and tables.

This is where the n-gram model comes in. In this model, n represents the length of a sequence of words to be counted. This means our bag-of-words model was the same as an n-gram model where n = 1..We will witness what happens when we run an n-gram model where n = 2.

Most of the positive bigrams are unremarkable, while a few of the negative ones like "bin laden" and "threatens to" could be considered to carry some negative meaning.