CS 562: Project Proposal

Eric D. Stevens

January 31, 2019

Task

The purpose of this paper is to attempt to predict stock market movements by utilizing natural language processing techniques to analyze news headline data. The goal will be to use Keras TensorFlow to build a recurrent neural network that will evaluate the sentiment as well as extract meaning from the news clips and try to relate these factors to moves in the market. It will be interesting to see if a recurrent neural network will be able to draw correlations between the news and the stock market. If it can, it will be even more interesting to see if information about the features that were generated can be extracted.

Data

The data for this project is provided by Kaggle user arron7son (https://www.kaggle.com/aaron7sun) , and can be found can be found **here** (https://www.kaggle.com/aaron7sun/stocknews) . It consists of two distinct datasets. The first is normal Dow Jones Industrial Average ticker data such as open price, closing price, high, low, and volume, paired with a date for every day the market was open over a time period from August 2008 to July 2016. The second data consists of roughly 73,000 single sentence news wire headlines over the same time period. A small sample of this second set of data is below.

Date	News
6/22/08	b'"The Doomsday Code": Scary British documentary examines Christian end-timers and their growing political influence.'
6/22/08	b'AP - Everything seemingly is spinning out of control'
6/22/08	b'Why China is trying to colonise Africa'
6/22/08	b'Newsweek: The Booze Is Back in Baghdad; men are shaving their beards; women are wearing jeans and taking off their headscarves; couples are holding hands in public'
6/22/08	b'The ugly truth behind the Iraq and Afghanistan wars finally has emerged: These Wars Are About Oil, Not Democracy'

Date	News
6/22/08	b"Palestinians Angered Over Obama's Opinions on Jerusalem - Palestinian leaders reject his call for Jerusalem to be Israeli capitol"
6/22/08	b"Secret of the 'lost' tribe that wasn't: Tribal guardian admits the Amazon Indians' existence was already known, but he hoped the publicity would lift the threat of logging"
6/22/08	b'Iran says an attack by Israel is "Impossible" while ElBaradei claims it would turn the region "into a fireball"'
6/22/08	b'Italian Court Blocks Construction of U.S. Military Base'
6/22/08	b'NATO General: We need 6000 more troops to gain trust among the Afghan population or the situation will get worse'

Evaluation

There are several evaluation strategies that could be attempted. The leas complicated would be to generate a model that would simply output a given days closing price given the opening price and several days of news prior to the current day. This Idea could be extrapolated to modeling prices for different time periods by having a many to many type RNN. In this case the loss function would be simple. A loss signal cold be generated by taking the absolute value of the distance between the predicted price and the actual price.

Concerns

My concern about this proposal is that there may not be enough data to train a complex RNN. I do not have enough experience with either NLP or DNNs to evaluate the amount of data that would be needed to get meaningful results form this project. If advised not to proceed with RNNs due to a lake of data, an audible can be called and the project can utilize more traditional NLP techniques that require much less data, or, an effort can be made to find or generate a more substantial dataset.