Sentiment Based Market Prediction

Issue #1: Number of articles per day is not the same.

Solution #1: Normalize data

Concern: Normalizing will give more weight to each article, from days with fewer articles. The fewer the articles available for a particular day, the more weight each article carries.

Solution #2: Cap the number of daily articles to a maximum. This maximum is equal to the articles from the day with least amount of articles.

Concern: Not using the entire article data set available, i.e., you are discarding articles.

Decision: Normalize sentiment score using the average sentiment score for each day.

Observations:

From Aug 27th 2011 to November 30th 2013, the least amount of articles was 21 in a day.

Nov 27th, 2011: only 27 articles.

Dec 25th, 2011: only 22 articles.