Overview

The Dow Jones Industrial Average (Dow), is a stock market index that measures the stock performance of 30 large companies listed on stock exchanges in the US. Dow data can be downloaded from yahoo finance.

This dataset:

Includes Dow movement from 2008-08 to 2016-01 along with top news headlines for the corresponding date.

Project Components:

1. Part 1: Full EDA and data prep
    a. Apply data cleansing as needed
    b. Feature creation
        i. Use any lexicon-based method (eg: text-blob, vader, etc) to determine the overall sentiment of the news headlines by date.
        ii. Create numeric features from the text which can be applied in your machine learning models. Examples include: sentiment polarity scores, text statistics (character and word counts), etc.
2. Part 2: NLP (Unsupervised) Model
    a. Perform topic modeling on the news headlines for the date(s) with the max and min increase based on Label.
        i. You essentially create 2 corpora, one with the max Label value and other with the min Label value. The idea is to understand the news topics which may have contributed to these increases/ decreases.
3. Part 3: ML (Supervised) Model for prediction or classification

    Some examples:

    a. Classification: Predict whether the Dow Index will increase or decrease based on the sentiment of news headlines.
        i. You can change the Label field to be a binary field (0, 1) where 0 means increase and 1 means decrease. (Note: Label is the difference between the Adjusted Close from the prior day. Hence positive means the index is up.

b. Prediction: Create a regression model to predict the Label based on numeric features, including sentiment polarity scores

Requirements:

1. INTRODUCTION: analysis objective and the ML approach applied as well as why you selected it.

2. EDA/DATA PREP: explore data issues that may require cleansing, data wrangling/munging, etc.. Include visualizations, statistical analysis, etc.. to better understand the data such as how it's distributed and correlations. You must include analysis/text that explains the meaning of each output.

3. FEATURE SELECTION/ MODELING: determine features to be applied and create a model.

4. PERFORMANCE ASSESSMENT: assess model performance using appropriate metrics.

5. CONCLUSION: a conclusion summarizing the analysis and the results. Were you able to meet your analysis objective as described in your introduction?

6. REFERENCES