Mid-Term Project

Summary

Dataset details:

I have used Daily News for Stock Market Prediction - Using 8 years daily news headlines to predict stock market movement for my Mid-term project. I have downloaded Combined\_News\_DJIA.csv dataset from <https://www.kaggle.com/aaron7sun/stocknews#Combined_News_DJIA.csv> . This dataset contains 27columns and 1989 rows. The first column is "Date", the second is "Label", and the following ones are news headlines ranging from "Top1" to "Top25".

After reading **description of the dataset, I have decided to replace headers with easily recognizable names. From my understanding of the description I have created a Combined\_News\_DJIA\_updated\_headers.csv file with updated headers which I have used for replacing headers.**

After printing the first row in human readable format, I have decided to

* Change the date to standard American time format,
* replace '0', '1' values in 'Stock\_market\_movement' column to DJIA Adj Close value decreased or increased or stayed the same and
* Remove the extra ’b’ from the news headlines.

While looking for the outliers and missing data I have observed seven ‘None’ values in the dataset. When I was searching for the ‘None’ values in the data what I have observed is that they are actually empty spaces which have been converted on to ‘None’.

I couldn’t find any duplicate rows. But what if the news columns have duplicate rows? I have decided to find the duplicate news as well. In addition to seven ‘None’ values, I have found that Top\_news\_headline3, “Putin: Who gave NATO right to kill Gaddafi?” has been repeated twice. As the duplicates may not harm or impact the Label value much alone, I have decided to keep the data as it is.

I have done fuzzy matching on 'Top\_news\_headline1' of row 1 using different functions and compared the results. Overall it was not a straight forward process moving from step1 to step2 but I had to move back and forth as I got different questions while completing the process.

References:

<https://stackoverflow.com/questions/41918836/how-do-i-get-rid-of-the-b-prefix-in-a-string-in-python/43680634#43680634>

<https://www.geeksforgeeks.org/python-removing-duplicate-dicts-in-list/>