Directory Structure

To simplify the review process of our work, the file structure is outlined below:

File 1: Data_Retrieval.ipynb

This file contains code to retrieve data from target sources, process sentiment scores for headlines and feature engineer additional financial indicators. This file also completes the data storage process in the PostgreSQL database. Please note, due to the local nature of our database, running of this file will be unavailable without alterations to database details. If you would like to run code for the model building process this is possible as all datasets used were saved outside the database as csv files.

File 2: Intermediary_Data_Processing.ipynb

This file is responsible for gathering specific data from the project database, completing minor processing to produce final, useable datasets for target companies. These datasets are stored in CSV files to be used in model training. Technological issues in accessing TensorFlow modules and Spark services resulted in this file being split from the final model training file. In the future, combining these files would be a primary goal to eliminate the need to use CSV files in any form. Datasets could be pulled directly from the database without the need for intermediary steps to further simplify the modelling process.

File 3: Model_Building.ipynb

This file is responsible for reading CSV files and performing all steps in the model training and testing process.

File 4: application.py

The application file is operable from the command line and serves as an extension of our work to depict a possible use case for our work. This file is basic in nature and missing key input controls, however, it effectively demonstrates the application of our model.