BT4222 Project - Text Analysis for Stock Returns Prediction

Data Files Documentation

Prepared by: Team 03
Jasmine Seah (A0171424M) | Kaustubh Jagtap (A0168820B) | Nicklaus Ong Jing Xue (A0170687U)
Sung Zheng Jie (A0168188M) | Vienna Wong (A0172316L)

This documentation details the file structure and general naming convention for our project.

Description for each file is also described.

Files/Folders	**Subfolders/Description**		
Raw Data	Data extracted from various textual sources		
	Price	price_labels.csv	Contains 10 years worth of DJIA stocks prices
	10K SEC	CIK_date.html	
	10Q SEC	CIK_date.html	
	8K SEC	CIK_date.txt (text document)	Contains events reported by the company in .txt file
	8K SEC Types	CIK_date.csv (event type)	Extracted event types are stored in .csv file
	Financial News	ticker_date.csv	
	Tweets	ticker_tweets.csv	Contains 10 years worth of daily tweets from the company
		Other Sources	Contains images for word cloud
	Amazon Product AAPL	aapl_public.tsv	AWS Public Dataset
		aapl_amazon.csv	Scrapy Scraping of Amazon Website

<table>
<tr><td rowspan="3"></td><td></td><td>aapl_products_dates.xslx</td><td>Apple Products and its release dates retrieve from online sources</td></tr>
<tr><td rowspan="2">Metalearner</td><td>cik_mapper.csv</td><td>Contains mapping of sector each DJIA company belongs to</td></tr>
<tr><td>technical_indicators_RAW.csv</td><td>Contains 10 years worth of technical indicators for each DJIA ticker</td></tr>
<tr><td>Processed Data</td><td colspan="3"><u>Data cleaned for analytics to be formed</u></td></tr>
<tr><td>10K SEC</td><td>CIK_df.txt</td><td rowspan="5">Each folder contains every individual ticker processed dataframe (total: 25).

To perform analytics, concatenate all individual CIK_ or ticker_df.pkl to get the combined dataframe.</td></tr>
<tr><td>10Q SEC</td><td>CIK_df.txt</td></tr>
<tr><td>8K SEC</td><td>CIK_df.pkl</td></tr>
<tr><td>Financial News</td><td>ticker_df.pkl</td></tr>
<tr><td>Tweets</td><td>ticker_df.pkl</td></tr>
<tr><td></td><td>Amazon Product AAPL</td><td>aapl_df.pkl</td><td>Folder only contains the AAPL processed dataframe.</td></tr>
<tr><td>Predictions</td><td colspan="3"><u>Buy-sell signals for meta-learner</u>
Each folder contains 25 individual train and test dataset for meta-learner model
- Train (2010 - 2017)
- Test (2018 - 2019)</td></tr>
<tr><td rowspan="2"></td><td rowspan="2">10K10Q</td><td colspan="2">ticker_train_10k10Q.csv</td></tr>
<tr><td colspan="2">ticker_test_10k10Q.csv</td></tr>
<tr><td rowspan="2"></td><td rowspan="2">8K SEC</td><td colspan="2">ticker_train_8K.csv</td></tr>
<tr><td colspan="2">ticker_test_8K.csv</td></tr>
<tr><td></td><td>Financial News</td><td colspan="2">ticker_train_news.csv</td></tr>
</table>

		ticker_test_news.csv	
	Tweets	ticker_train_tweets.csv	
		ticker_test_tweets.csv	
	Amazon Product AAPL	AAPL_train_AmazonReviews.csv	
		AAPL_test_AmazonReviews.csv	
	Metalearner	meta_train.csv	
		meta_test.csv	
		meta_aapl_train.csv	
		meta_aapl_test.csv	
		df_test_AAPL_predictions.csv	
		25_tickers_prediction.csv	
Analytics	<u>The files names follow the steps in the Machine Learning Life Cycles:</u> (1) Data Extraction (2) Data Processing (3) EDA and Feature Engineering (4) Modelling *Example Directory:* - *Analytics > 10K10Q > extraction.py* - *Analytics > 8K > model_building.ipynb*		
	(1) 10K10Q (2) 8K SEC (3) Financial News (4) Tweets (5) Amazon Product AAPL	extraction.py	Data extraction algo
		cleaning.py	Data cleaning and preprocessing algorithm
		model_building.ipynb	EDA, features engineering process, and bundle of prediction algos
	Tweets	techniques.py	Contains text processing functions
		corporaForSpellCorrection.txt slang.txt	Corpus used for cleaning Tweets. Referenced by cleaning.py

	Metalearner	join_sources.ipynb	Concatenate all the predictions of each ticker from all the datasets *Note: It is an OFFLINE version of our notebook which is hosted on Colab. It reads data from Google Drive directly and is NOT SUITABLE to be ran locally.*
		final_metalearner.ipynb	EDA, features engineering process, and bundle of prediction algos
		aapl_metalearner.ipynb	EDA, features engineering process, and bundle of prediction algos
	Backtest	Backtest.ipynb	Backtest on the meta-learner predictions to see the results of our model
		Backtest-Apple.ipynb	Backtest on the Apple meta-learner predictions to see the results of our model

Directory Examples	*Raw Data/10K/CIK/CIK_date.html*
	Raw Data/Amazon Product AAPL/apple.xlsx