Requirements Document

Introduction

Project Name:

Research Project, Effects of News Media sentiment and Stock price fluctuations.

Problem Statement:

This project will develop machine learning models and techniques to further the research done by the Leeds School of Business in applying machine learning to finance. Specifically, this project will apply NLP models to gauge sentiment in business TV corpora and correlate the sentiment to movements in stock prices. Additionally, this process will attempt to measure the effects of different news sources or political segments and explore how news media has changed over time.

While the main goal of the project is to correlate sentiment from the TV corpora with changes in stock prices, there are several other potential topics of interest. One of these is investigating the correlation of politics and economics, specifically how political news affects the stock market. Additionally, investigating and predicting stock reactions to business interviews on a minute to minute basis would be another topic of interest. Finally, time allowing, utilizing optical character recognition on the Moody's manuals would be an additional area of research.

Project Scope:

Throughout the time working on this project we will help further the research of the correlation between sentiment in business TV corpora and movements in stock prices. If positive results are made, our project will deliver effective methods to process TV corpora. Also, this research could be published by our sponsor: Diego Garcia, Chair of the Finance Division at CU Boulder which will detail the relationships we have discovered. Anyone who keeps up with financial research could be affected by this in learning to what extent outside forces affect the stock market. Our area of expertise is software systems. We will create a software system that is robust, scalable, maintainable, and flexible for many use cases. Provide specific detail on what the project will deliver. Provide information on the departments and the functional areas that will be impacted by the project.

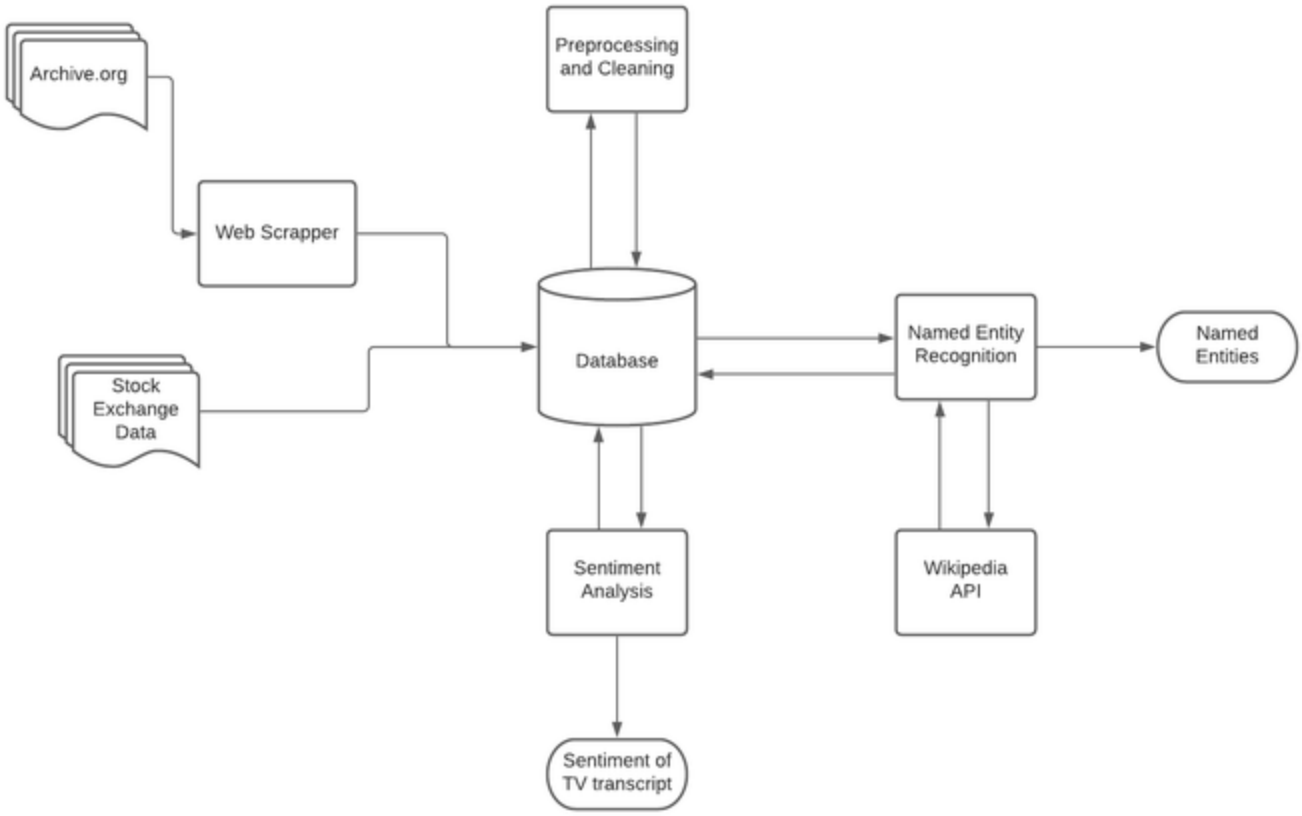
Functional Objectives

- Have database that stores all relevant data
- Collect and organize data from many sources
- Clean and preprocess data for analysis and future processing
- Find sentiment within financial text
- Identify named entities in financial text
- Collect data to describe named entities
- Analyze data collected and correlate it with stock market changes
- Produce results useful for research objectives
- Mature database and data processing pipeline to maintain up-to-date metrics
- Database should be reliable and accessible to any users who have permission
- Improve run time by running code on CU compute cluster

Non-Functional Objectives

- Program should run on any OS through Docker
- Create database structure to store scraped web data from multiple sources
- Implement web scrappers to extract data for processing
- Use NLP processing to extract quantitative contextual information
- Use Wikipedia API to get page for named entities
- Parse information from Wikipedia page
- Run database and code on CU compute cluster using Singularity

Data Flow Diagram



Audience and User Interface:

- Access program from Github repo (https://github.com/TheBuffsOfWallStreet/NLP-FIN-LAB) and follow README for setup
- Interact with program through terminal
- Perfect for researchers that have interest in building database of TV corpora and deriving their own insights

User Case

Use Case Name	Create Database of Shows
Summary	The user can use a program that will create database entries for shows and then run a download program which will fill the index's with transcripts and important metadata
Flow	1. Run main program 2. Type "i" (build index) and press enter 3. Select network of interest 4. Input how many shows user wants to build database index for 5. Repeat steps 2-5 for as many shows and networks that are available to download 6. Type "d" (download episodes) and press enter
Alternative Flows	User does not have to do step 6 if they only want to build the index for the database

Use Case Name	Detect Duplicates
Summary	The user can run a program that will iterate over all the shows and will flag shows that are similar to ones that came before it
Flow	1. Run main program 2. Type "dd" and press enter 3. Wait until progress bar completes

Preconditions	User has built an index for shows and has downloaded shows

Use Case Name	**Get Sentiment in Text**
Summary	The user can run a program that will return the sentiment for a given input text.
Flow	1. Run main program 2. Program asks for text 3. Input text to be analyzed 4. Return positive and negative sentiment values
Preconditions	NLP libraries have been installed
Alternative Flows	User could use wrapper program to run sentiment analysis on many texts from the database.

Use Case Name	**Name Entity Recognition**
Summary	The user can run a program that will find the named entities within input text.
Flow	1. Run program 2. Ask for text 3. Input text that which will be used to find the named entities 4. Return a list of named entities that are in the text
Preconditions	Database contains financial text and NLP libraries have been installed
Alternative Flows	User could use wrapper program to named entity recognition on many texts from the database.
Exception Flows	No named entities are found

Use Case Name	**Get Wikipedia Information**
Summary	The user can get information from Wikipedia given an input company or person name.
Flow	1. Run program 2. Program asks for name of company or person 3. Input name 4. Searches Wikipedia for relevant pages and/or suggestions 5. If no pages found for that name return error 6. Choose the best result from the search 7. Return Wikipedia page 8. Parse and extract information from Wikipedia page 9. Return extracted data
Preconditions	Internet connection and Wikipedia API installed
Alternative Flows	User could use wrapper program to collect Wikipedia information about many named entities
Exception Flows	No Wikipedia page found from input name