The Final Project: Proposal

Team AcubeG: Amogh Kallihal, Anchal Jain, Abhishek Sundar Raman, Gayatri Ganapathy

1. List 3 questions that you intend to answer (1 point)

1. Whether we can generate stable stock portfolios using various clustering methods for a quarter/period of time?
2. Is it possible to do asset management using stock portfolios obtained through clustering methods?
3. Can we use NLP methods to see whether the news about the companies has any impact on stocks in the same period of time(let's say a month, quarter, a year or any socio-political event/news that affects the prices of the stocks)?

2. List all the datasets you intend to use (1 point)

1. Financial news data about the companies from various news API's.
2. OHLCV stock data streamed through API's.

3. Give us a rough idea of how you plan to use the datasets to answer these questions. (2 points)

- Data Collection:
  Financial OHLCV stock data would be accessed through Quandl or AlphaVantage API and news data from different news sources using APIs. Eg. Yahoo News, Bloomberg, News API.

- Data Exploration:
  EDA would be required to understand the OHLCV data and what they represent over a period of time for the stocks. EDA task would assist us to finalize on the parameters and identify the relationships in the input data. Through this analysis, we will decide the final features to be used as input to the clustering model.

- Data Cleaning:

The data obtained from the News API is in the form of text data which requires needs to be preprocessed to be converted into feature vectors utilized in sentiment analysis. We also need to filter the OHLCV data to be considered for the respective time frame of our analysis.

- Data Integration:
  News data from different sources for different companies will be integrated to generate clusters based on news sentiment. Further, OHLCV data needs to be integrated with news data for creating clusters based on sentiment and historical data.

- Data Analysis:
  Various clustering techniques would be employed to arrive at a portfolio of the stocks. Cluster evaluation metrics would be used to arrive at optimal clusters. The effect of news data on the stock would also be analyzed using NLP sentiment method.

- Data Product:
  The product would contain visualizations of the data and also a report of the clustering techniques utilized to arrive at the trends and common stocks portfolio generation through data.

4. Think about that once your project is complete, what impacts it can make. Pick up the greatest one and write it down. (1 point)

The end product would be an insightful report and visualizations for a financial advisor/investor to make informed decisions about stock portfolios built using clustering techniques and assist asset management.

5. Tools and technologies to be explored:

Data Collection	Quandl or AlphaVantage API and news data from different news sources using APIs
Data Integration	Kafka Streaming
Data Analysis	NLTK, Clustering methods like K-means, DBSCAN, etc.
Data Product	Report and Visualization through Tableau/Web portal