

Requirement Phase

Team: Mount Harvard

Members: Jenna Beutler, Seth Grossman, Miao Hua, Samuel Statton, Nhi Vo

Dataset: Tweets related to Coronavirus

[https://www.kaggle.com/smid80/coronavirus-covid19-tweets#2020-03-00%20Coronavirus%20Tweets%20\(pre%202020-03-12\).CSV](https://www.kaggle.com/smid80/coronavirus-covid19-tweets#2020-03-00%20Coronavirus%20Tweets%20(pre%202020-03-12).CSV)

Business Case: To investigate the changing topics and sentiment in tweets from around the world between March 12th and 30th.

Functional Requirements:

The following functional requirements need to be in place to achieve the end product as described in the Project Requirement document:

- Data Ingestion
 - Loading the data in AWS via EMR master node
- Coding
 - Collaboration amongst group members via Google Colab, Github and AWS to write, review, and deploy code
 - Enable horizontal scaling
- Data Visualization
 - Using Pyspark to perform basic aggregate functions and visualize the data
- Text Analysis
 - Utilizing Pyspark text analysis tools to gain insights into each tweet
- Time Analysis
 - Change in sentiment, number of tweets, arising topics throughout time
- Business Applications
 - Translate insights from the data into recommendations for running business during a global event like this
- Deliverables
 - April 12: Requirements Document, GitHub Classroom
 - April 25: Design and Development Test
 - April 28: Deployment & Presentation

Non-Functional Requirements:

The following non-functional requirements have been considered by the team as factors that will influence how the project is conducted but deemed non-functional:

- Collaboration

- Using Google Colab to share code and ideas before posting to GitHub
- Communication
 - WhatsApp is our primary text communication app while Slack will be utilized when we have questions for Peter.
 - Zoom will also be used for checkpoints and to share screens for code examples