February 16, 2025

1 Introduction

This project involves collecting and analyzing data from three major sources: Reddit API, Yahoo Finance, and the World Health Organization (WHO) COVID-19 database. The collected data is processed and stored for further analysis.

2 Overview of the Topic

The primary focus of this project is to analyze healthcare trends using diverse data sources. The motivation behind selecting this topic is to understand how online discussions, financial trends, and pandemic statistics intersect to provide insights into public health concerns and economic impacts. The expected data includes user-generated discussions on telemedicine, historical stock prices of healthcare companies, and real-time COVID-19 case statistics.

3 Data Sources

3.1 Reddit API

The Reddit API is used to fetch posts from relevant subreddits such as `r/Telemedicine` and `r/AskDocs`. The extracted data includes post titles, text content, author information, upvotes, posting date, and subreddit name. This information is retrieved using the PRAW (Python Reddit API Wrapper) library and stored in a structured CSV format for analysis.

3.2 Yahoo Finance

Yahoo Finance provides historical stock price data for various healthcare-related companies and ETFs. The dataset includes closing prices for stocks such as Pfizer (PFE), Johnson Johnson (JNJ), Moderna (MRNA), UnitedHealth (UNH), and healthcare-focused ETFs like XLV, IYH, and VHT. The financial data is retrieved using the `yfinance` library and stored as a CSV file for further study.

3.3 WHO COVID-19 Data

The World Health Organization (WHO) provides COVID-19 statistics, which are extracted from their official dashboard using the `requests` library. The data includes case numbers and other relevant statistics, which are scraped and structured into a tabular format for analysis.

4 Data Collection Process

For each data source, specific steps were taken:

- **Reddit API**: Used PRAW to authenticate and fetch posts. Challenges included API rate limits and filtering out irrelevant content.
- **Yahoo Finance**: Retrieved historical stock data using `yfinance`. Ensuring data completeness over different timeframes was a challenge.
- **WHO COVID-19 Data**: Web scraped data from the WHO dashboard. Handling dynamic content loading was a technical hurdle.

5 AI Application

Using the collected data, an AI-driven analysis tool can be developed to track trends in healthcare discussions, correlate financial movements with public health developments, and predict future patterns in disease spread and stock performance.

6 Terms of Service and Privacy Concerns

- **Reddit API**: Restrictions on storing and redistributing user-generated content.
- **Yahoo Finance**: Limitations on extensive automated data extraction and redistribution.
- **WHO Data**: Ethical concerns regarding data use, especially for sensitive health information.

7 Impact of Multiple Data Sources

Collecting from multiple sources enhances data richness but introduces challenges in merging and aligning datasets. Differences in data granularity, reporting frequency, and possible discrepancies in reported values must be managed through normalization and cross-validation techniques.

8 Storing and Combining Data

To efficiently store and integrate data from multiple sources, a structured approach is necessary. A relational database such as PostgreSQL or MySQL can be used, where each data source is stored in separate tables with common attributes like date or keywords to facilitate merging. Alternatively, a NoSQL database like MongoDB can be employed to handle semi-structured data, making it easier to store diverse formats.

For integration, a data processing pipeline using Python's Pandas library can clean and normalize data, ensuring consistency across sources. Joining data based on common identifiers, such as dates for stock prices and COVID-19 statistics, allows for meaningful correlations. Additionally, cloud storage solutions like Google BigQuery or AWS S3 can be used for scalability and accessibility, enabling real-time data analysis and AI-driven insights.