News Collection and Analysis System

Group Members

1. Jinqian Wang
2. Kaushik
3. Revanth

PROG8380/Web Design Capstone

Sec X

Submitted to

Zaman ANK, PhD

# News Collection and Analysis System

## **Introduction**

Our project is about news collection and analysis. There are three main parts of this system. Data collection system, data analysis system, and data demonstration page. The system will collect news from mainstream media of all over the world automatically by the timetable which is created by admins. Then, the analysis system will analyze these data, like extracting keywords, making tags for articles and pictures. Finally, the demonstration system will show all news by date, categories, and keywords with graphs.

Now, let us introduce more detail about these systems.

The first one is the news collection system. This system will be developed by python. We will develop a python web API server to set the basic configuration for collection information. Admin can decide the timetable of gather, which information will be collected, and where the data come from. The page of configuration is created by MEAN. The system will follow the rules of configuration and collect the data automatically. After collection, the data will be sent to the data analysis system.

The second is about the data analysis system. After collection, this system will get all the data and start to analyze it. In this system, AI will be the main assistant to manage this huge data. Our AI will help admins to extract the keywords of the news and tap the labels to the news and sort the news into different categories that are made by admins. After analysis, these data will be saved in the database.

The third is about the web pages for the whole project. For the whole project, there are two main parts, one is the front pages to show the news and graphs, another is the back pages for admins to control the system. We will use MEAN tech to build these pages.

We use different program languages to do this project. We use node.js and express to build our website, use python to collect data, use MySQL as the main database and use MongoDB as an auxiliary database. The main purpose to do this is that we want the system can have a good extension. Data collection system not only collect news. It can collect the data depend on the requirements from uses. The data analysis system can be used for other information.

## **Design and Data Flow**

Website Technology:

Express.js, node.js, Angular.js, D3.js, MongoDB, MySQL

Data Collection System Technology:

Flash(python framework), Scrapy(python framework)

Data analysis System Technology:

TensorFlow(python)

Front pages:

Home page, news list page, graph pages, search page, image page.

Back pages:

Collection configuration page, admin page, news management page, category page, label page, image management page.

## **Technology Description**

We use MEAN tech because all group members know how to use MEAN technology. Python is one of the popular programming languages for data collection. It has a lot of libraries which can help us to achieve our goals easier. The TensorFlow framework can help us complete machine learning rapidly. For databases, we will MySQL and MongoDB because MySQL is good for complex data and MongoDB can promote our big data response speed.

## **Market Survey**

Please list top-ten technologies asking in most recent job postings. We do not need the definition or description of the technology in this document. The goal of doing a market survey is to encourage you to choose a project and work on a project that will help you find a job after graduation.

## **Conclusion**

The conclusion of your proposal goes here that includes a rough idea of the project completion timeline and related comments.

## **GitHub Link**

https://github.com/echoaiyaya/InfoCollectionSystem