News Headline Scraper Dataset

Group Members

Jasmit Kakkar (jkakkar@scu.edu)
Tanmay Agrawal (tagrawal@scu.edu)
Yuzheng Wu (ywu4@scu.edu)

The main goal of this project is to create a central database of news headlines and their appropriate timestamps. We will achieve this by creating spider programs that crawl major news websites and extract the data from them. The main points of information we want are headlines and timestamps. These headlines will have to be processed to apply lemmatization and basic NLP so we can extract just the important data. They will feed that data into a centralized database where we will perform some analysis.

The analysis we will perform will create a graphical representation of the most common headline tags that appear across all the websites in our dataset.

Motivation

To learn more about the news headline cycle and see trends in how the news headlines evolve over time. Once we can extract the raw data we can see what keywords are being talked about the most and which keywords are most relevant at that point in time.

Technologies

Technology: Spider(Scrapy - python), database(sql), inter-server communication, virtual machines(amazon ec2), docker(on the top of ec2, for quick deployment)

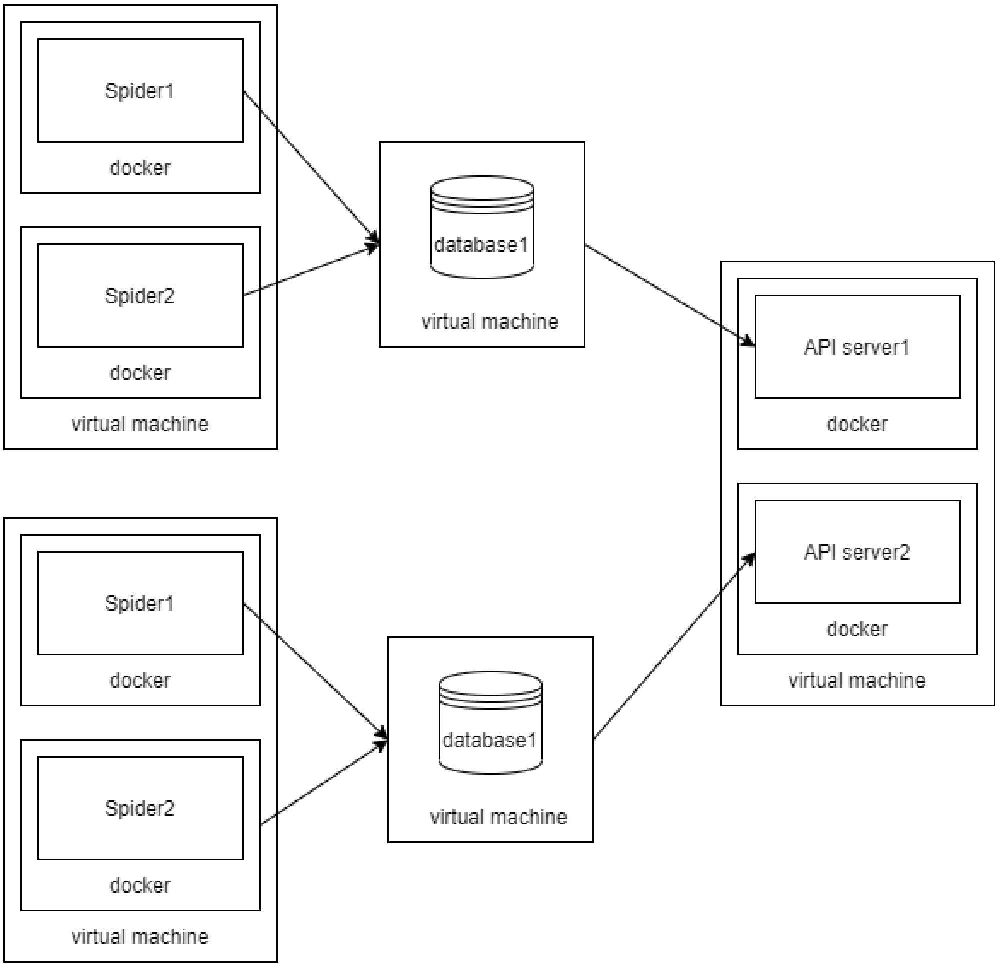
Architecture

instances(machines): the basic unit of our project would be:

- spider machine * n
- database * 1
- accessible api *1

It could be scaled up if necessary. So that more spiders could be added to improve the speed of scraping or increase the number of websites being worked on. More databases could be added for better data availability or to spread tables on different machines. More api servers could be added if the usage of the website gets more popular.

Each group of spiders working on the same webpage will be assigned a sql database. They will establish connections with the database and write to it. The API server will establish a connection with the database, and when people request for a piece of data, the api server will query from the database.



Tasks

Create spiders using python - (Scrapy) - Jasmit Kakkar

- Create script to scan specific website and grab headline/timestamp
- Generalize the script so you can deploy this across multiple machines and different websites

Create environment for spiders to run in (VM/container) - Yuzheng Wu

- Configure the vm
- Deploy the vm/container
- Write scripts for fast deployment

Connect the spiders to feed and create to a central dataset via SQL -Yuzheng Wu

- Write SQL sentences that will be used
- Establish connection between SQL and spiders

Create the api for people to access the data - Tanmay

- Create a RESTful api for data access
- Write the SQL queries to access the data