Documentation

Premise: Here the background was to make a crawler that would do the various website monitoring. One of the main goals was to incorporate MySQL RDS to our script. We run the script after a specific period of time to get the updated results.

Tools Used:

1.Docker container
2.MySQL
3.Python
4.LXML,TREE,DOM
5.File System and sys config
6.Unit tests

Architecture:
Mainly consists of four parts:
1.Python script
2. Json files for XPATH reference
3. Docker
4. MySQL connector script

We followed the architecture that would allow us to fetch multiple websites.

Websites: Scrapped the data from the following websites:

1. 7news
2. 9news
3. News.com.au

Outputs: The outputs were stored in the MySQL database in the Docker container with openport :80

Display: The command line was used to display the Updated results.

Common Observations:

1. There were heavy news update in the morning
2. Slowest update in the late night
3. 9news has the most updates

Problems faced:

1. Connection time-out(Put the argument in the make_connection function to increase time)
2. Programs stopped(Put the Try: Catch: block)

3. Database was not updating(did the Connection.commit() function execution)

Future Improvements:

1. Efficient Use of MYSQL select function that would be time optimizing
2. Include more website to the lists
3. Scraping to the individual category page of the each websites
4. Deleting the Chunking big mysql database

Usages: The scraped data could be used in the analytics of websites about their upload time, contents, monitoring and many more