Overview

This document contains the setup for postgres database setup on cloud environment and prepared a script that executes every hour which will basically crawl over different sources and scrap the required data.

Here i listed out all the necessary steps to meet the requirements

- **SET UP AWS INSTANCE:**
  We create ubuntu free tier aws instance and login to the instance thereafter installed the requirements to the instance like (docker, docker-compose)
  Useful links :
  Launch aws instance
  Install docker
  Install docker-compose

- **DATABASE SETUP**
  We used a docker image of postgres to set up database service in the cloud. In order to achieve run postgres docker container in aws instance and mapped the data volume inside the instance.

  Steps:
  1. Clone this repository postgres-docker
  2. Go to the repository folder edit docker-compose.yml to set name and password to database
  3. Run command `docker-compose up -d --build`
  4. Postgres docker service will ready to use

- **Run Script every hour**
  The available python script in file crawldash.py is mapped to api endpoints using Flask and this api endpoints is set up in linux flock bash commands. This linux command is mapped to crontab of instance where time is defined one hour which results to trigger the api endpoint in every hour. This api basically use the functions from that file and then result from that file is stored to database checking if that title from data is already in the database or not