Reddit Network Visualization AWS Neo4j GraphXR Instructions

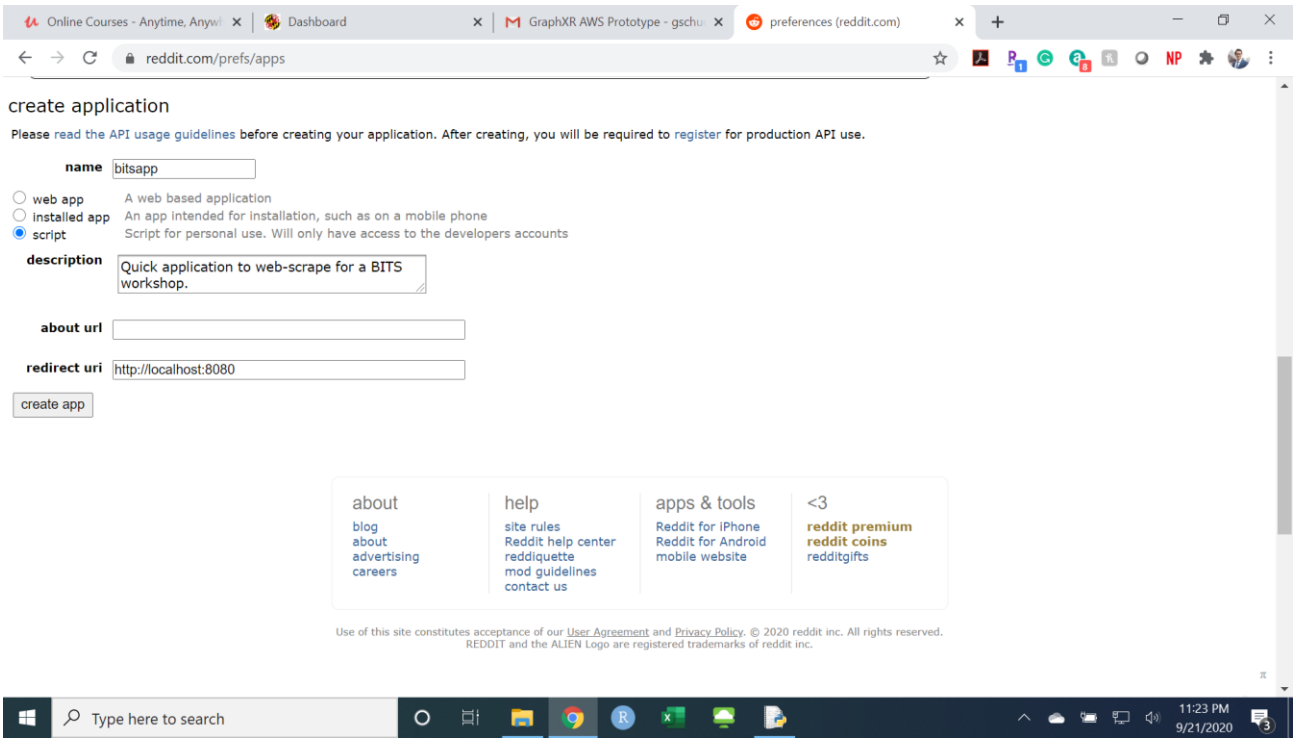
Contact: Graham Schuckman at gschuckm@terpmail.umd.edu

Recording Link: https://youtu.be/tPq3CdvlTL0

UMD BITS
BUSINESS & IT SOCIETY

Workshop Prerequisites:

1. AWS Account (Requires Credit Card Info)
    a. Note: You will not incur any charges for this workshop
    b. Note: If asked, you will need a personal and root user account
2. WinSCP (Windows) Or FileZilla Client (Mac)
3. GraphXR Account
4. Reddit Developer Account (ignore the warning about registering for production API usage)
5. Reddit Test App (see instructions in screenshot below)



6. Git Bash (required if not using terminal on Mac)
7. Workshop Folder (need to download the entire reddit folder, not just the contents)

Workshop Instructions:

1. Launch a default Linux 2 AMI instance on AWS with the user data pasted in from the attached docker_bash.txt file, and ports 7474 and 7687 open to custom TCP traffic in the security group.

2. Edit the reddit_api.py file and substitute the client_id and other credentials with your reddit dev app.

3. Unzip the attached files and copy the unzipped reddit folder into your EC2 instance using WinSCP, FileZilla, or another file transfer system. To use WinSCP, download it, pass in the DNS hostname of the EC2 instance, and use SSH under Advanced --> SSH --> Authentication with agent forwarding turned on. You will need to find the private key file by clicking the ... and then choosing "All Private Key Files" from the dropdown. It will likely ask you to convert your .pem to a .ppk which you will need to do. The username for WinSCP into your AWS instance should be ec2-user. Leave the password blank.
    a. Note: if using FileZilla, follow the instructions here (use ec2-user instead of ubuntu for the username) and enter password for the password.
4. SSH into your instance (click Connect and use EC2 Instance Connect in browser or use a terminal under and follow the SSH client instructions) and run the following commands:
    a. docker pull neo4j
        i. This command downloads the Neo4j image from the Docker Hub so it can be run as a container in the next command (think of it as downloading a song and then playing)
    b. docker run --name testneo4j -d -p7474:7474 -p7687:7687 -v $HOME/data:/data -v $HOME/logs:/logs -v $HOME/import:/var/lib/neo4j/import -v $HOME/plugins:/plugins --env NEO4J_AUTH=neo4j/password neo4j
        i. This command will run Neo4j as a container named “testneo4j” in detached mode (-d means in the background), open up the necessary ports between the container and the instance, mount the volumes to run the application, and sets the username as neo4j and password as password for the database login.
    c. cd reddit (if you accidentally cd somewhere else, type cd /home/ec2-user)
        i. This command changes the directory from the home directory to the reddit folder. Directories are basically folders in a tree-like structure, with root as the highest/top part.
    d. sudo pip3 install -r requirements.txt
        i. This command will use Python 3’s pip command to install the needed Python packages to interact with the Reddit API and Neo4j from a requirements.txt file.
    e. python3 reddit_api.py subredditname1 subredditname2 subredditname3 num_messages **(ex: python3 reddit_api.py umd umdcs terps 100)**
        i. This command will specify that we want to run our Python 3 file to pull data from chosen subreddits for up to 100 messages and store it as JSON in the posts folder.
    f. python3 import_reddit_messages.py
        i. This command will import the data we scraped from Reddit into our Neo4j container so that it can be viewed and analyzed.
5. Verify that the data was uploaded correctly by navigating to the public DNS of your instance followed by port 7474 in your browser, preferably Chrome **(ex: ec2-100-25-152-152.compute-1.amazonaws.com:7474)**
6. If it asks for a login, the username is neo4j and the password is password. Run the following command in the query area ($): **MATCH (n)-[r]-(m) RETURN ***

    a. This command will query all nodes and all relationships between nodes and return them.

7. Log into GraphXR, create a project and give it a name, then check the box that says Configure Neo4j Instance. Put in the public DNS of your instance and then port 7687 for the Bolt Port. The username should be neo4j and the password is password.

8. Under Connection Type, choose Through GraphXR server connection, then hit confirm. To view and interact with your data, click the Query icon on the left-hand bar (</>) and then paste in this query: MATCH (n)-[r]-(m) RETURN *

9. To run the query, hit Ctrl+Enter or the little right-facing arrow next to the query box.

10. TERMINATE YOUR INSTANCE IN THE AWS CONSOLE WHEN YOU ARE FINISHED!