**Subreddit Ranking**

**Files**

# subreddit\_ranking\_with\_luigi.py

Luigi task scheduler script to connect API, download data and processing the data.

# subreddit\_twice\_day.py

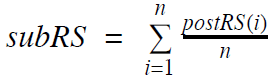
Python script to schedule the process twice in a day.

**Problem statement**

Thegoal of this project is to design and develop a solution that tracks and ranks the trending topics in Reddit over time.

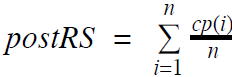
**Solution**

Given ranking criteria for subreddit is



Where **n** is number of posts and **postRS** is the post score.

Ranking criteria for the posts are also given



Here **n** is the number of comments and **cp** is the comment score.

From these ranking criteria and the requirement, flow design is prepared.

**Design**

From the requirement needs to ranks the top subreddits and its posts as well as the comments. So, from reddit we get the trending posts in the categories of top, hot and controversial. With help of “Python Reddit API Wrapper”.

The design process is given below.

**Reddit Account Creation**

For the data downloading from the reddits needs an official reddit account. With help of this “[reddit: the front page of the internet](https://www.reddit.com/)” Link creates reddit account by giving the credentials.

**Reddit API Connection Establishment**

From reddit data downloading is require certain steps.

Go through this “https://www.reddit.com/prefs/apps/” Link and gave the credential in developer option.

After giving the credentials we get a window with secret ID and secret key.

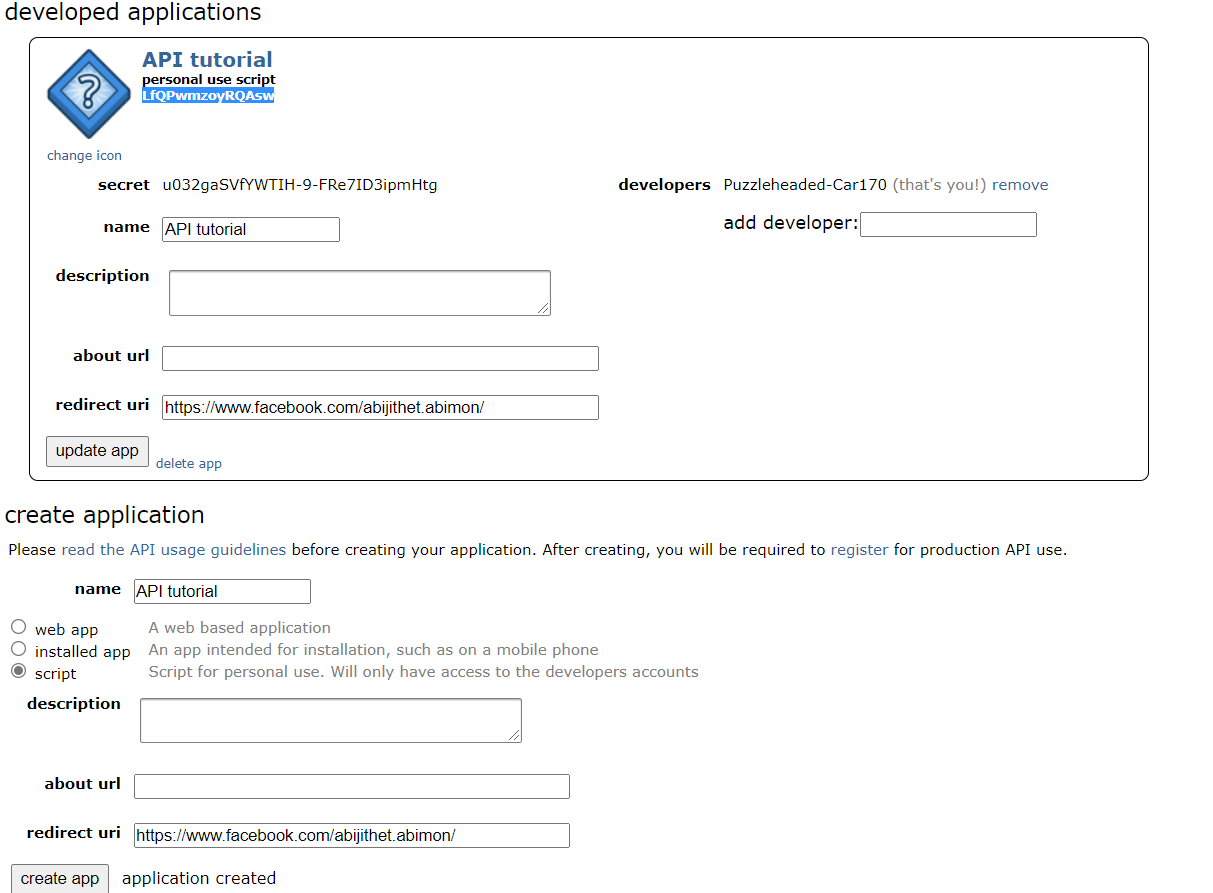


Figure: 1

Secret ID and secret keys are in this figure 1 is the blue shade text and secret text respectively.

With help of these credentials and name of the script app is used to create the API for implementation.

**Implementation in Luigi**

First of all, Luigi is a **Python (3.6, 3.7 tested) package** that helps you build complex pipelines of batch jobs. It handles dependency resolution, workflow management, visualization, handling failures, command line integration, and much more.

Here the solution is implemented as single task in luigi with “local scheduler”.

The solution run twice in day and generate an output excel.

**Pre-requirements and installation steps**

#Python 3.7

#Luigi

* For installation of luigi, run the command “pip install luigi”

#Praw

* For installation of luigi, run the command “pip install praw”

#Schedule

* For installation of luigi, run the command “pip install schedule”

#Openpyxl

* For installation of luigi, run the command “pip install openpyxl”

#Pandas

* For installation of luigi, run the command “pip install pandas”

**Running the solution in locally**

Clone the git hub repo <https://github.com/abijithraaz/Rank-the-topics-in-reddit-with-the-help-of-luigi> into the local machine.

Get the secret ID and secret code from the reddit app window and modify the arguments in the “subreddit\_twice\_day.py”

Run the code **python subreddit\_twice\_day.py** in command terminal.

An output excel with subreddit ranks and posts will be generated inthe **output\_log** directory.

**Docker containerization**

### **Installation Docker**

#### Install Docker on Ubuntu

* Update software repositories

sudo apt-get update

* Uninstall Old versions of docker

sudo apt-get remove docker-engine docker.io

* Install docker

sudo apt install docker.io

#### Install Docker on Windows

* Download Docker desktop for windows

<https://hub.docker.com/editions/community/docker-ce-desktop-windows/>

* Double-click docker desktop installer.exe and run installation.

**Dockerfile creation**

Create an empty document file with filename “Dockerfile” in the docker location

Add the following lines in Dockerfile for creating the environment to run the code.

* Import python and pip

FROM python:3.7-slim-buster

* Install all needed packages to run the inference code

RUN pip3 install --upgrade pip

RUN pip3 install luigi

RUN pip3 install praw

RUN pip3 install schedule

RUN pip3 install openpyxl

RUN pip3 install pandas

* Set working directory as workspace

WORKDIR /workspace

* Save Dockerfile and open terminal from where the Dockerfile is located.
* All the above steps are already done in the dockerfile inside the docker directory ithe repo.
* Run in terminal inside from docker and run the commands
  1. Build base image

*docker build –t <*base image name*>*

*eg: docker build –t subreddit\_ranking* .

* 1. Verify whether the image is created

*docker images*

* Do all the above steps once in a system for creating the environment.

Create an empty document file with filename “Dockerfile” for running the code in already created environment.

This document is already created in the repo the below steps just for the clarrification.

Add the following lines in Dockerfile for run the code.

* Import already created environment docker image

FROM subreddit\_ranking

* Set working directory as workspace

WORKDIR /workspace

* Copy code and depended folders into container

COPY subreddit\_ranking\_with\_luigi.py ./

COPY subreddit\_twice\_day.py ./

* Run script

CMD ["python", "subreddit\_twice\_day.py"]

* Save Dockerfile and open terminal from where the Dockerfile is located.
* Run in terminal from the work directory
  1. Build base image

*docker build –t <*base image name*>*

*eg: docker build –t* subreddit\_ranking .

* 1. Verify whether the image is created

*docker images*

* Before run the code create an output folder in the work directory and give the full path for the volume mounting.
* Run docker container

Docker run –v /host/fullpath:/container/path *<*base image name*>*

Eg: docker run -v D:/Allianz/output\_log:/workspace/output\_log subreddit\_ranking