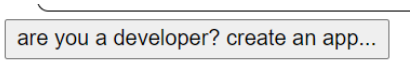
GR5067 – HW#4

Prerequisite (should be done):

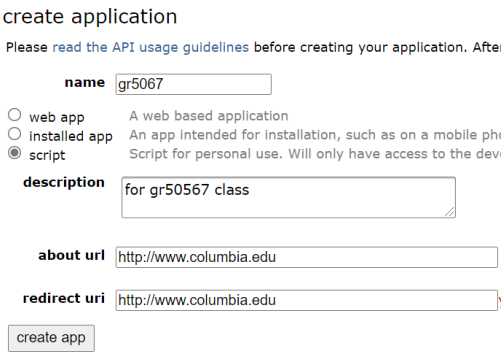
1. Install praw by performing a "pip install praw" from a conda shell or from linux shell. Praw is an API that interfaces with the Reddit API. The API enables a user to access both historical and real-time subreddit channel messages.

2. Sign up for a Reddit account if you don't have one already

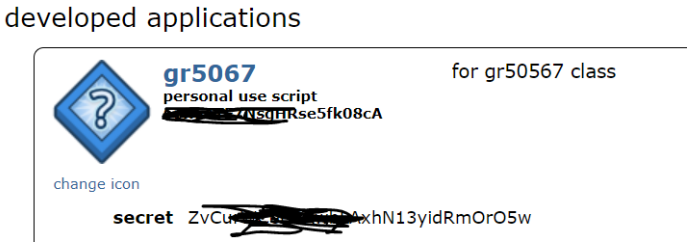
3. Create a Reddit App

   - Click onto 

   - Fill out the "Create Application" exactly as you see below:



   - Once you click the "create app" button, please scroll up and copy your "personal use script" and "secret":



   - Now go to the Python script called my_reddt.py and paste your "client_id", "client_secret", "username", "password" into the proper fields, below:

```
reddit = praw.Reddit(
    client_id="YOURS",
    client_secret="YOURS",
    user_agent="testscript by u/fakebot3",
    username="YOURS",
    password="YOURS",
    check_for_async=False
)
```

Question (100 Points)

The following files exist in CourseWorks files -> hw -> hw4

my_reddit.py – top level code

vectorizer.pk – vectorizer stage

pca.pk – dimension reduction stage

my_model.pk – model that was already trained on corpuses related to 'republican policies' and 'democrat policies', each being one of two class labels

Just running my_reddit.py will result in real-time messages from the subreddit channel "politics" to stream in your Spyder Console.

➔ You need to use the functions available to you from lecture code *utils.py* and refactor anything in *my_reddit.py* and *utils.py* and classify the text located in the *tmp_df* dictionary key called "body". You need to leverage functions from *utils.py* and replicate the following pre-processing steps (some functions will need to be slightly modified to accommodate the flow below), transformations and classifying the text in the exact order below:

clean_text → rem_sw → my_stem → vectorize (vectorizer.pk) → pca (pca.pk) → classify (my_model.pk)

Your code should simply output the class label prediction and the likelihood score in real-time.