Social Data Aquisition with APIs

Layla Bouzoubaa

7/11/23

Hello

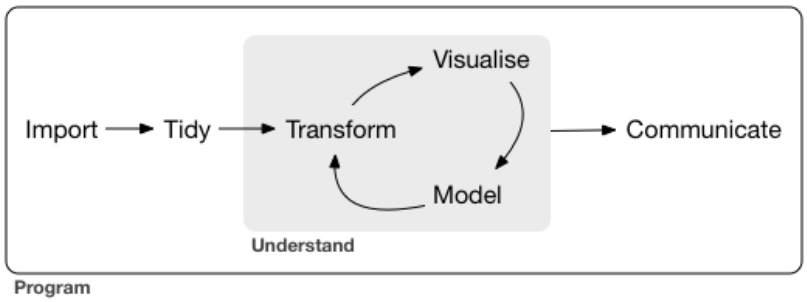
Doctoral student, IS

NLP, social computing, drugs

Passionate about

Outline

- Introduce API
- Setup Reddit API
- Reddit API with examples
- Play time



Typical Data Science Pipeline

Reddit

- The "front page of the internet"
- Reddit has an API that allows developers to access and extract data from the platform.

Reddit is a popular social news aggregation and discussion website where users can post content and interact with others. In this tutorial, we will learn how to use Python to get data from Reddit using its API.

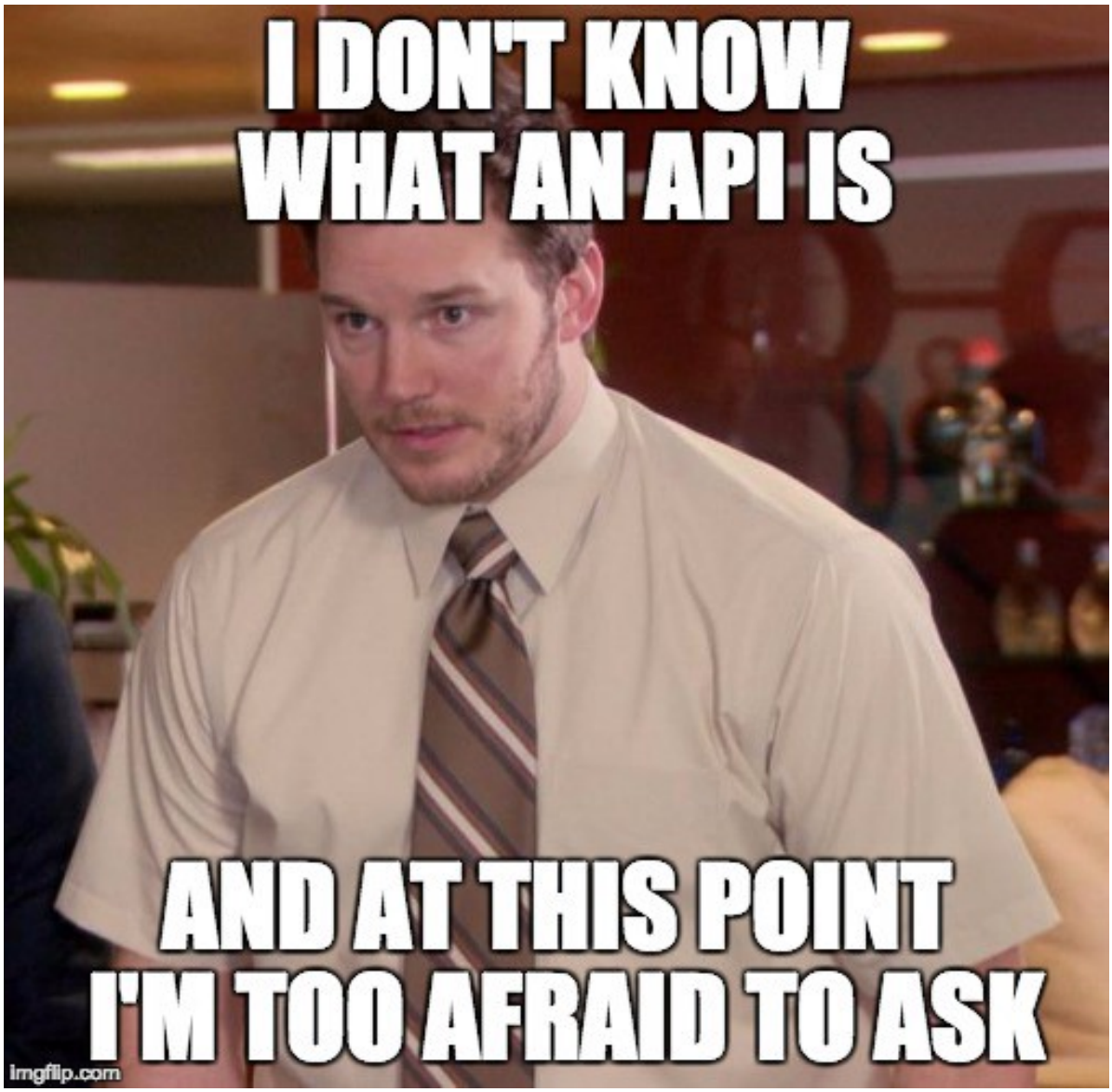
What is an API?

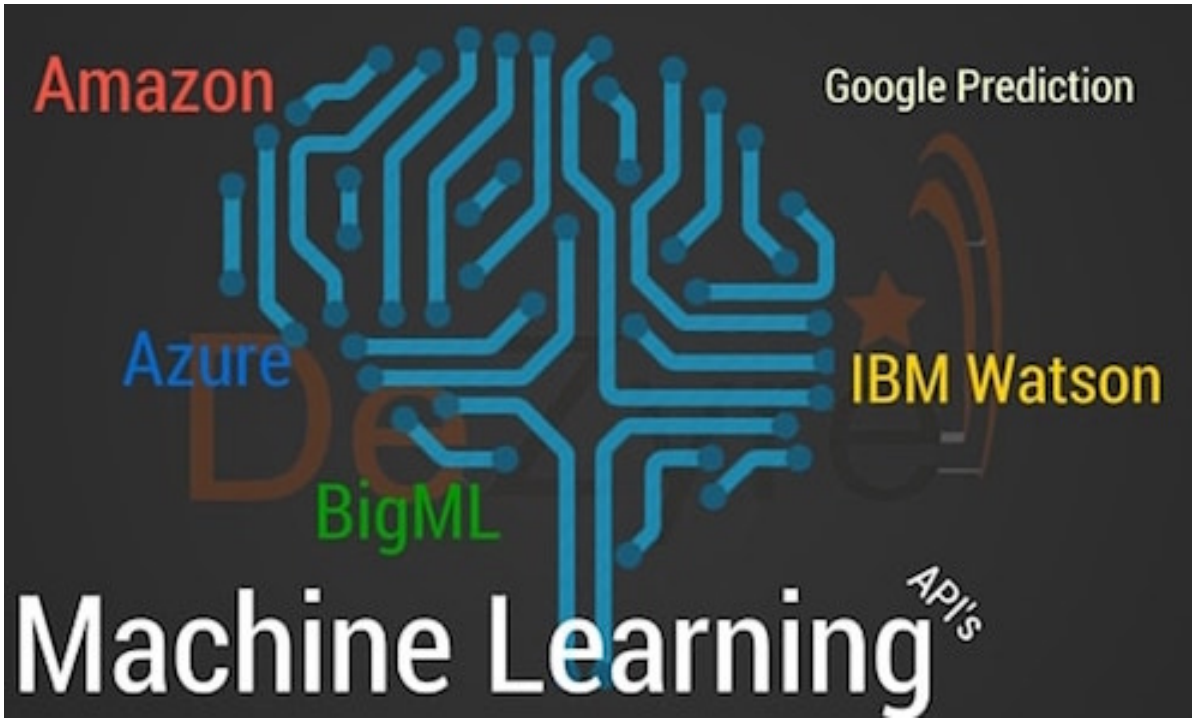
Application **P**rogram **I**nterface

- Allows a user to programatically pull data from a source given that source provides one
  - *ex. The NYT, Twitter, Facebook, Google, US Census*

How Does it Work?

- Like a Graphical User Interface (GUI) allows you to interact with your code, an API lets your code interact with other code
- Set of commands, functions, objects, and protocols


I DON'T KNOW
WHAT AN API IS
AND AT THIS POINT
I'M TOO AFRAID TO ASK
imgflip.com



- used to create software or interact with an external system (databases)
- convenient, ready-made code to simplify the life of a programmer

Not Only for Getting Data

- Build APIs for your ML models so others can use them
- Write (post) to a database
  - and update (put) and delete (delete)
- Functions
  - more control over what is exposed to end user

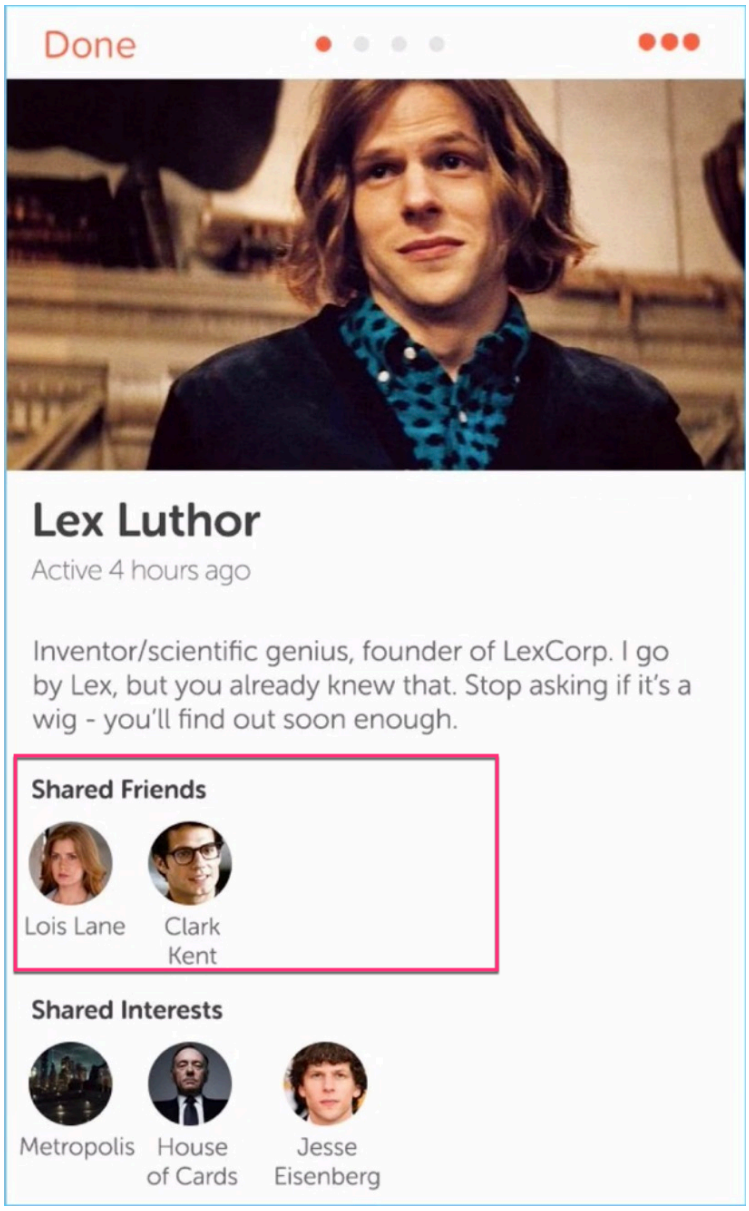
The API

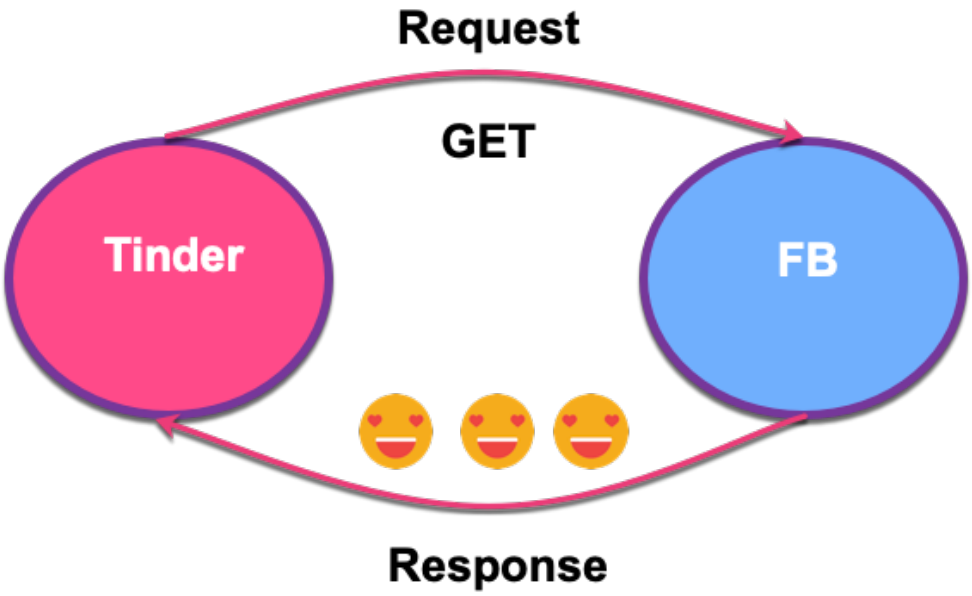
- **G**ET: used to retrieve data from the server
- **P**OST: send data to the server to create a new resource
- **P**UT: update an existing resource on the server
- **D**ELETE: delete an existing resource on the server

USER

- **G**ET: used to retrieve data from the server. When a client sends a GET request, the server responds with the requested data (if it exists)
- **P**OST: send data to the server to create a new resource. For example, a client might send a POST request to create a new user account on Reddit
- **P**UT: update an existing resource on the server. For example, a client might send a PUT request to update a user's profile information
- **D**ELETE: delete an existing resource on the server. For example, a client might send a DELETE request to delete a user account

There is this App...

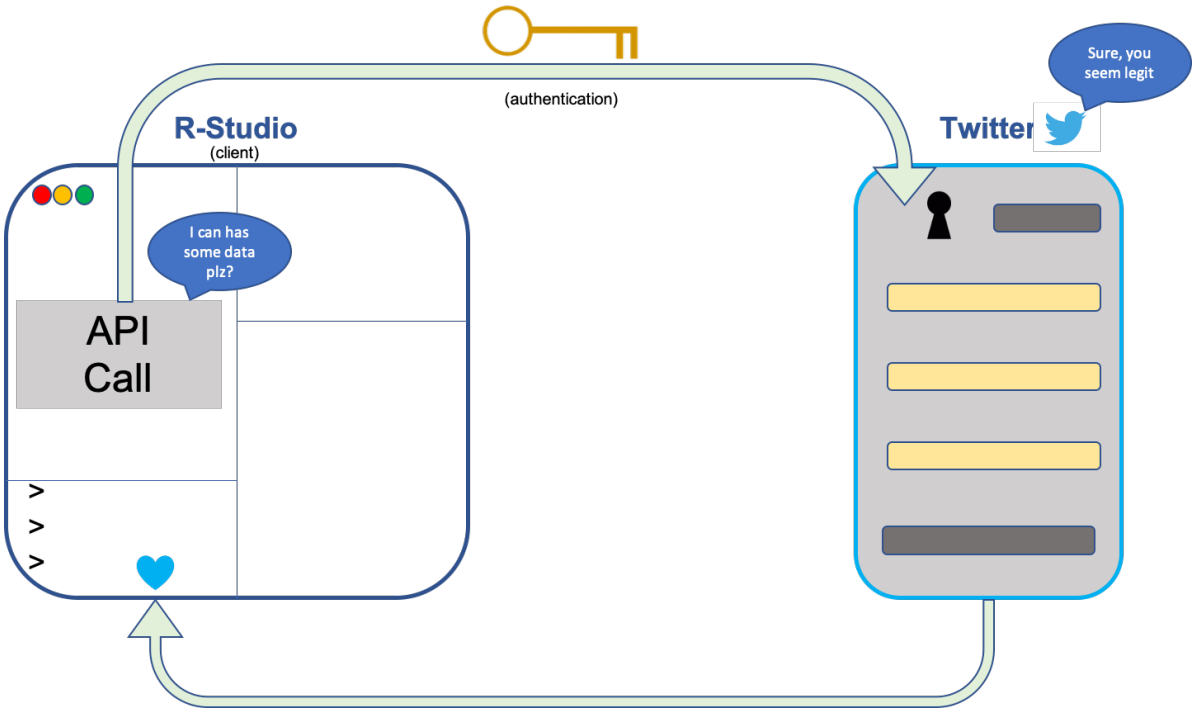




It's that easy?

- 1. You must first agree to the terms of use and obtain an API Key.
  - 1. This authenticates you to the server that holds the data
  - 2. Sometimes it's simple and sometimes it's a process...be prepared
- 2. Store this key in a local `.py` file
  - 1. DO NOT SHARE/PUBLISH YOUR API KEY!!
- 3. Run your API call
  - 2. DO NOT SHARE/PUBLISH YOUR API KEY!!

Keeper of the Keys



Understanding Reddit API

- Reddit API is a RESTful API that uses HTTP requests to communicate with the server.
- API endpoints are used to perform specific tasks such as getting information about a subreddit, submitting posts or comments, and searching for content.
- To access the Reddit API, we need to authenticate with our API key.

The Problem

Which Friends characters are discussed more?

The One with the Steps

Source: reddit.com/r/friends_tv_show

1. Obtain Reddit API Key
2. Install required libraries

    1. PRAW
    2. requests
3. Authenticate with Reddit
4. Do data science

```
pip install praw
pip install requests
```

- PRAW is a Python Reddit API Wrapper that provides a more user-friendly interface for accessing the API.
- Requests is a Python library that allows us to send HTTP requests.
- To authenticate with the Reddit API, we need to create a Reddit instance and pass our API key and secret as arguments.
- After creating the instance, we can use it to perform various tasks such as getting information about subreddits, retrieving posts and comments, and submitting content.

The One with the Notebook

See the notebook here to follow along

Resources

- What is an API?
- Reddit API Documentation