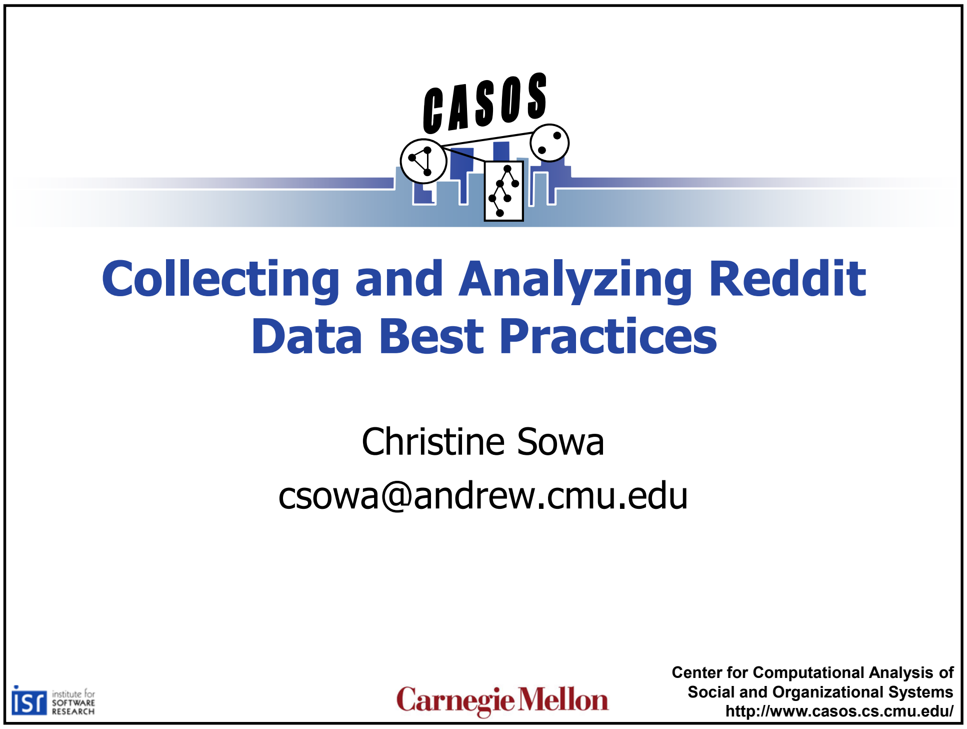
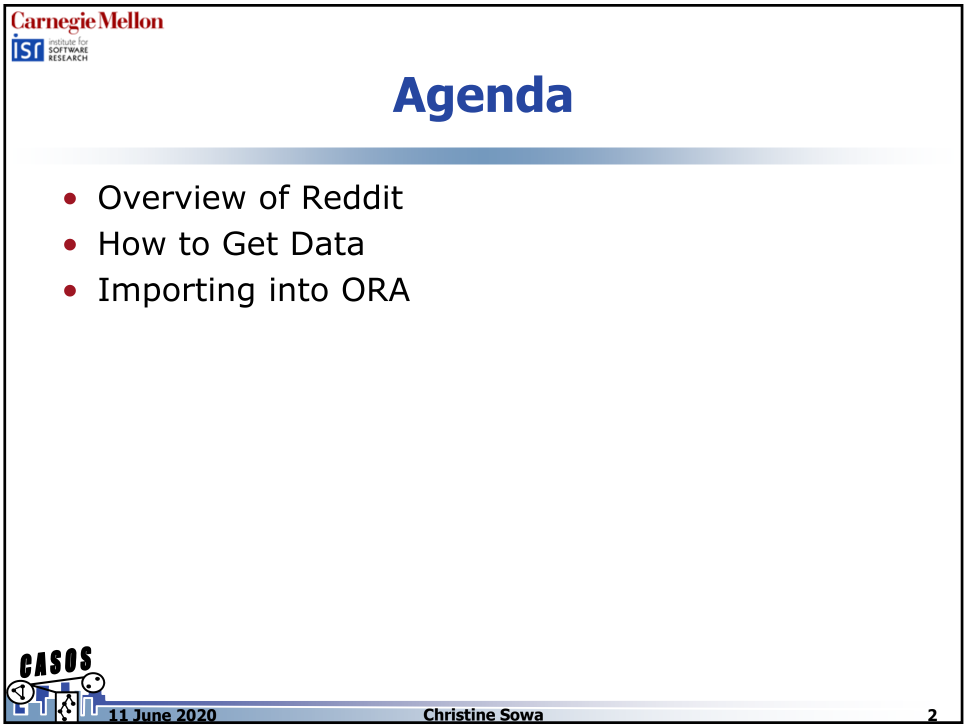


Collecting and Analyzing Reddit Data Best Practices

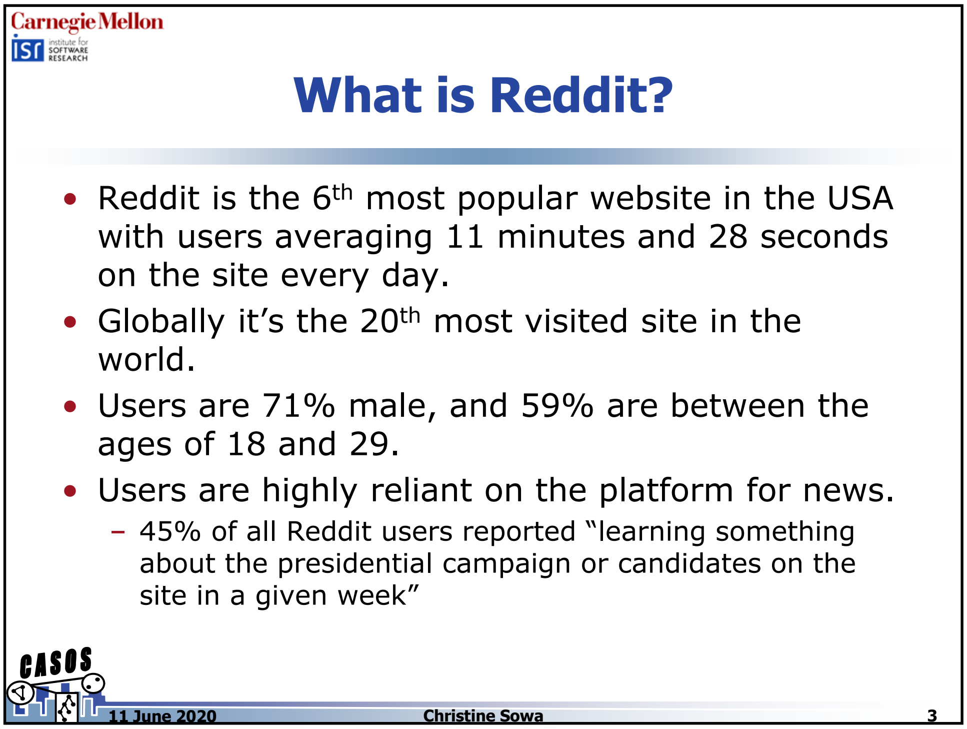
Christine Sowa
csowa@andrew.cmu.edu

Center for Computational Analysis of
Social and Organizational Systems
http://www.casos.cs.cmu.edu/



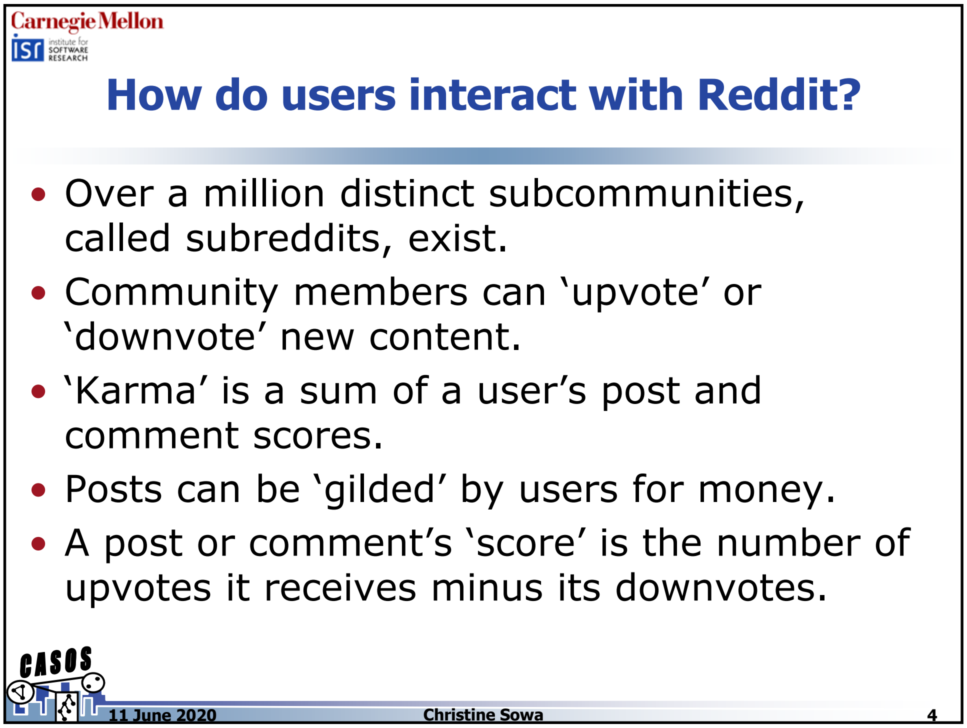
Agenda

- Overview of Reddit
- How to Get Data
- Importing into ORA



What is Reddit?

- Reddit is the 6th most popular website in the USA with users averaging 11 minutes and 28 seconds on the site every day.
- Globally it’s the 20th most visited site in the world.
- Users are 71% male, and 59% are between the ages of 18 and 29.
- Users are highly reliant on the platform for news.
  - 45% of all Reddit users reported “learning something about the presidential campaign or candidates on the site in a given week”



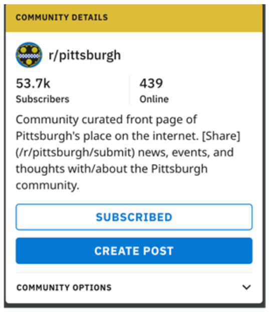
How do users interact with Reddit?

- Over a million distinct subcommunities, called subreddits, exist.
- Community members can ‘upvote’ or ‘downvote’ new content.
- ‘Karma’ is a sum of a user’s post and comment scores.
- Posts can be ‘gilded’ by users for money.
- A post or comment’s ‘score’ is the number of upvotes it receives minus its downvotes.

What makes Reddit unique?

- Moderation
  - Each subreddit has moderators that enforce community standards for posts

Example Interactions



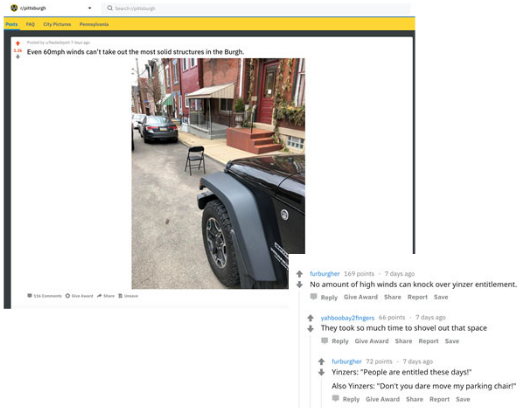
COMMUNITY DETAILS

r/pittsburgh

53.7k Subscribers

439 Online

Community curated front page of Pittsburgh's place on the internet. [Share](/r/pittsburgh/submit) news, events, and thoughts with/about the Pittsburgh community.



Posted by u/[illegible] 7 days ago

5.2k

Even 60mph winds can't take out the most solid structures in the Burgh.

furburgher 169 points · 7 days ago
No amount of high winds can knock over yinzer entitlement.

yahboobay2fingers 66 points · 7 days ago
They took so much time to shovel out that space

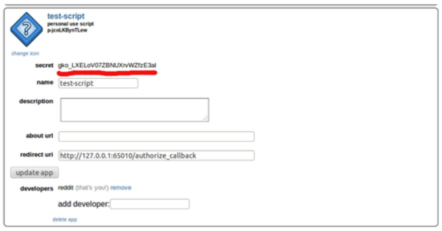
furburgher 72 points · 7 days ago
Yinzers: "People are entitled these days!"

Also Yinzers: "Don't you dare move my parking chair!"

Carnegie Mellon

The Reddit API

- First must read the terms and register to use the API
- API data format comes out as a JSON
  - One JSON per post or comment
- Can use wrappers (like praw or PushShift for Python).



test-script
personal use script
p-jcoLKBynTLew

secret gko_LXELoV07ZBNUXrvWZfzE3al

name test-script

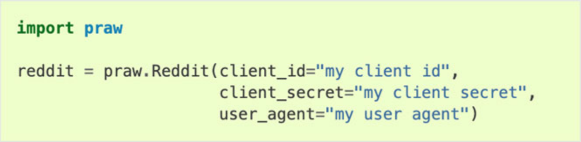
description

about url

redirect uri http://127.0.0.1:65010/authorize_callback

developers reddit (that's you!) remove

add developer:



```
import praw

reddit = praw.Reddit(client_id="my client id",
                     client_secret="my client secret",
                     user_agent="my user agent")
```

Carnegie Mellon

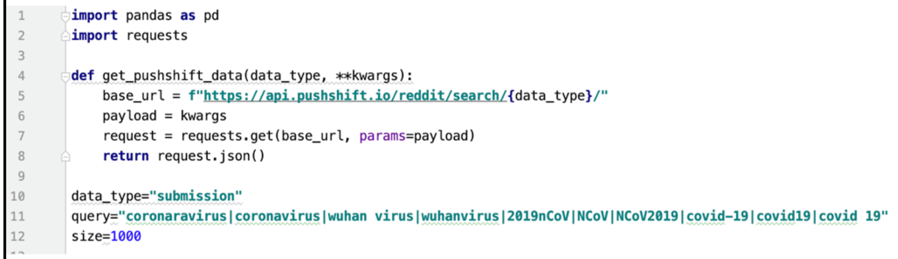
Type of Data to Pull

- Get all of the posts (Submissions) from a given subreddit from the past 30 days
  - Get post title, score, id, url, number of comments, author, score
- Get all posts from a given Redditor
- Obtain all comments to a set of posts
  - Get comment author, time, score, text

Reddit Networks

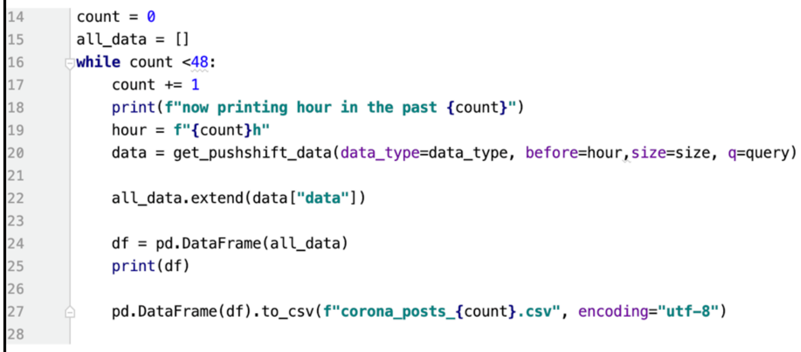
- User x Subreddit
- User x Post
- User x User
- …

Walking through API using PushShift



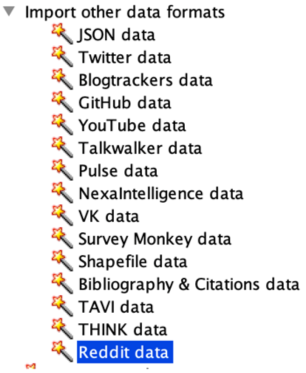
```python
import pandas as pd
import requests

def get_pushshift_data(data_type, **kwargs):
    base_url = f"https://api.pushshift.io/reddit/search/{data_type}/"
    payload = kwargs
    request = requests.get(base_url, params=payload)
    return request.json()

data_type="submission"
query="coronaravirus|coronavirus|wuhan virus|wuhanvirus|2019nCoV|NCoV|NCoV2019|covid-19|covid19|covid 19"
size=1000
```

Pulling Data with Pushshift



```
count = 0
all_data = []
while count <48:
    count += 1
    print(f"now printing hour in the past {count}")
    hour = f"{count}h"
    data = get_pushshift_data(data_type=data_type, before=hour,size=size, q=query)

    all_data.extend(data["data"])

    df = pd.DataFrame(all_data)
    print(df)

    pd.DataFrame(df).to_csv(f"corona_posts_{count}.csv", encoding="utf-8")

```

Uploading Data into Ora



▼ Import other data formats
- JSON data
- Twitter data
- Blogtrackers data
- GitHub data
- YouTube data
- Talkwalker data
- Pulse data
- NexaIntelligence data
- VK data
- Survey Monkey data
- Shapefile data
- Bibliography & Citations data
- TAVI data
- THINK data
- Reddit data