Things to know when working with reddit data

Show & Tell – Social Media-Daten in der Forschungspraxis II

Dr. Carolina Haensch
Thank you to Ashley Amaya et al., authors of the article „New Data Sources in Social Science Research: Things to Know Before Working With Reddit Data” on which I base part of my talk
1. Reddit overview
Reddit as platform

- Popular social media site
  - Worldwide Nr. 20
  - Germany Nr. 37
  - US Nr. 9 (all numbers from 2021)
- Self-proclaimed “front page of the Internet”
- Founded in 2005
Reddit basics

• Registered users can submit content
  • Content includes links, text posts, images, videos, GIFs
  • Upvoting and downvoting possible, influences position on pages

• Content is divided into boards/subpages called subreddits or communities
  • Example: r/dataisbeautiful or r/dataisugly

• Reddit administrators and subreddit-specific members moderate the subreddits
can anyone enlighten me of that the UK leaving means and or benefits? (in avg joe terms please)
I see all the fuss but i don't know what is all about?

Thread root

First level Comment

Third level Comment

Reddit as a data source for research

• Amount/Size: Due to its popularity, there is A LOT of data available

• Availability: Most subreddits (exact statistics are unavailable) are public
  • Public subreddits can be viewed by all
  • Posts may be commented and voted on by all registered users (free and anonymous)
  • The data may be downloaded for free by any registered user

• Identification of rare subpopulations: 130,000 active subreddits
  • Researchers of rare populations can identify and study eligible participants
Ein Subreddit für alle, die Franken lieben & im schönen Frankenland leben. Hier könnt ihr euch über Kultur, schöne Orte, gutes Essen und sonstige Interessen und Neuigkeiten in Franken austauschen! Teilt gerne eure Tipps.

Erstellt am 15. Dez. 2020

227 Mitglieder, 4 Online, Oberste 50%
Content created on Reddit in 2021

- Posts: 429,952,687
- Comments: 2,717,216,237
- PMs: 713,505,544
- Chats: 1,912,574,822

Total: 5,773,249,290
Examples for data analyses with Reddit data

- Gender roles (Ammari, Schoenebeck, & Romero, 2018)
  - Variation across parenting subreddits (mommit and daddit). Similarities in topics across the boards, such as sleep training, as well as differences, such as fathers talking about custody cases and Halloween

- News consumption (Wasike, 2011)
  - Human interest frame was the most common generic frame followed by technology. Science and technology was the most common issue-specific frame.

- Sexual identity (En, En, & Griffiths, 2013)
  - Evaluate through subreddit self-descriptions which aspects of identity are seen – and thereby made – to be markers of personal and collective sexual identities

- Mental health (Choudhury & De, 2014)
  - Self-disclosure in mental illness communities and characterization of mental health social support
Examples

• Right-wing accounts (Gaudette et al., 2021)
  • Group building dynamics in subreddits through calls for violence against minorities and political opponents of right-wingers

• Corona-pandemic and news consumption (Chipidza et al., 2022)
  • News sources in liberal and conservative subreddits
2. A closer look from the social science perspective
2.1 The reddit population
Reddit population

- Reddit is anonymous + anyone can view most subreddits
- Posting, voting, or commenting on Reddit requires registration
- **Demographics of the 330 million active users are unknown**
  - Users are more likely to be male and younger
  - Pew:
    - 6% of all U.S. adults used Reddit
    - Men twice as likely as women to use
  - Reddit:
    - 79% of worldwide users are 18–34 years of age
First half of 2022, the United States accounted for 47.13 percent of traffic.

The United Kingdom was ranked second, accounting for 7.48 percent of web visits to the social media platform.

Canada, Australia and Germany are the next biggest countries.
Reddit may be an appropriate source of data if a researcher is attempting to target a rare population but less ideal for general population research.

The definition of Redditors (e.g., Reddit posters, viewers/consumers, commenters) influences your findings.
2.2 Getting reddit data
Getting reddit data

- Two ways to access Reddit content
  
  - Via the Reddit API
    - PRAW for python
    - RedditExtractoR for R
  
  - Pushshift.io (maintained by redditor Jason Baumgartner, Stuck_In_The_Matrix)
    - Download via http://files.pushshift.io/reddit/
    - PSAW for python
    - pushshiftR for R
Getting reddit data via the Reddit API (python)

- API rules: https://github.com/reddit-archive/reddit/wiki/API#rules

- Any registered user may use the Reddit API, but they must fill out a special request and be authenticated. (This will take less than an hour.)

- The guide of the PRAW (Python Reddit API Wrapper) package provides documentation to start the process in Python

- Allows data collection from subreddits, threads, users

- Depending on the search being conducted and unit of analysis, there are limits to the number of units the system will return
Submission

A class for submissions to Reddit.

**Typical Attributes**

1. **Note**

   This table describes attributes that typically belong to objects of this class. PRAW dynamically provides the attributes that Reddit returns via the API. Since those attributes are subject to change on Reddit's end, PRAW makes no effort to document any new/removed/changed attributes, other than to instruct you on how to discover what is available. As a result, this table of attributes may not be complete. See Determine Available Attributes of an Object for detailed information.

   If you would like to add an attribute to this table, feel free to open a pull request.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>author</td>
<td>Provides an instance of <code>Redditor</code></td>
</tr>
<tr>
<td>clicked</td>
<td>Whether or not the submission has been clicked by the client.</td>
</tr>
<tr>
<td>comments</td>
<td>Provides an instance of <code>CommentForest</code></td>
</tr>
<tr>
<td>created_utc</td>
<td>Time the submission was created, represented in Unix Time.</td>
</tr>
<tr>
<td>distinguished</td>
<td>Whether or not the submission is distinguished.</td>
</tr>
<tr>
<td>edited</td>
<td>Whether or not the submission has been edited.</td>
</tr>
<tr>
<td>Field</td>
<td>Description</td>
</tr>
<tr>
<td>---------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>clicked</td>
<td>Whether or not the submission has been clicked by the client.</td>
</tr>
<tr>
<td>comments</td>
<td>Provides an instance of CommentForest.</td>
</tr>
<tr>
<td>created_utc</td>
<td>Time the submission was created, represented in Unix Time.</td>
</tr>
<tr>
<td>distinguished</td>
<td>Whether or not the submission is distinguished.</td>
</tr>
<tr>
<td>edited</td>
<td>Whether or not the submission has been edited.</td>
</tr>
<tr>
<td>id</td>
<td>ID of the submission.</td>
</tr>
<tr>
<td>is_original_content</td>
<td>Whether or not the submission has been set as original content.</td>
</tr>
<tr>
<td>is_self</td>
<td>Whether or not the submission is a selfpost (text-only).</td>
</tr>
<tr>
<td>link_flair_template_id</td>
<td>The link flair's ID.</td>
</tr>
<tr>
<td>link_flair_text</td>
<td>The link flair's text content, or None if not flaired.</td>
</tr>
<tr>
<td>locked</td>
<td>Whether or not the submission has been locked.</td>
</tr>
<tr>
<td>name</td>
<td>Fullname of the submission.</td>
</tr>
<tr>
<td>num_comments</td>
<td>The number of comments on the submission.</td>
</tr>
<tr>
<td>over_18</td>
<td>Whether or not the submission has been marked as NSFW.</td>
</tr>
<tr>
<td>permalink</td>
<td>A permalink for the submission.</td>
</tr>
<tr>
<td>poll_data</td>
<td>A PollData object representing the data of this submission, if it is a poll submission.</td>
</tr>
<tr>
<td>saved</td>
<td>Whether or not the submission is saved.</td>
</tr>
<tr>
<td>score</td>
<td>The number of upvotes for the submission.</td>
</tr>
<tr>
<td>selftext</td>
<td>The submissions’ selftext - an empty string if a link post.</td>
</tr>
<tr>
<td>spoiler</td>
<td>Whether or not the submission has been marked as a spoiler.</td>
</tr>
</tbody>
</table>
Getting reddit data via Reddit API

- Overview over options in R from APIs for Social Scientists (Bauer et al)
  - [https://bookdown.org/paul/apis_for_social_scientists/reddit-api.html](https://bookdown.org/paul/apis_for_social_scientists/reddit-api.html)

- Simple API calls via subreddit webaddresses *without authentication*
  - [https://www.reddit.com/r/cats/.json](https://www.reddit.com/r/cats/.json)
  - Possible to extracts threads, upvotes and so on

- RedditExtractoR: Reddit Data Extraction Toolkit
  - Limited functionalities

```r
find_subreddits
find_thread_urls
get_thread_content
get_user_content
```
Getting reddit data via pushshift download

- Contains all Reddit content
  - Data sets are stored as large compressed json files organized by month and by type of content (e.g., post or comment)
  - Data is copied into Pushshift at the time it is posted to reddit

- The data set structure and its contents changes over time as changes were implemented in Reddit

- Contains variables and content that may have since been removed or changed (not available on Reddit API)

- Current scores not available via pushshift
Getting reddit data via pushshift download

Directory Contents

Please consider making a donation (https://pushshift.io/donations) if you download a lot of data. This helps offset the costs of my time collecting data and providing bandwidth to make these files available to the public. Thank you!

If you have any questions about the data formats of the files or any other questions, please feel free to contact me at jason@pushshift.io.

<table>
<thead>
<tr>
<th>Filename</th>
<th>Type</th>
<th>Size (bytes)</th>
<th>Date Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>69M_reddit_accounts.csv.gz</td>
<td>69M_REDDIT_ACCOUNTS.CSV.GZ File</td>
<td>1,051,903,601</td>
<td>Sep 8 2021 6:10 AM</td>
</tr>
<tr>
<td>RA_2018-09.gz</td>
<td>RA_2018-09.GZ File</td>
<td>1,104,136,829</td>
<td>Sep 8 2021 6:11 AM</td>
</tr>
<tr>
<td>RS_2019-09-01.gz</td>
<td>RS_2019-09-01.GZ File</td>
<td>256,537,945</td>
<td>Sep 8 2021 6:12 AM</td>
</tr>
<tr>
<td>authors</td>
<td>&lt;Directory&gt;</td>
<td></td>
<td>Jun 23 2022 2:09 AM</td>
</tr>
<tr>
<td>authors.dat.zst</td>
<td>AUTHORS.DAT.ZST File</td>
<td>1,444,191,053</td>
<td>Sep 8 2021 6:14 AM</td>
</tr>
<tr>
<td>comments</td>
<td>&lt;Directory&gt;</td>
<td></td>
<td>Oct 11 2022 11:47 AM</td>
</tr>
<tr>
<td>daily</td>
<td>&lt;Directory&gt;</td>
<td></td>
<td>Sep 8 2021 5:54 AM</td>
</tr>
<tr>
<td>joel.data.ndjson.zst</td>
<td>JOEL_DATA.NDJSON.ZST File</td>
<td>105,415,750</td>
<td>Oct 19 2021 11:30 AM</td>
</tr>
</tbody>
</table>
Getting reddit data via pushshift API

- Pushshift also offer an API and there are API wrappers for R and python
  - Python package psaw
    - https://psaw.readthedocs.io/en
  - R Package pushshiftR
    - https://github.com/nathancunn/pushshiftR

- The Python package offers more options and is better documented.
# Comparison of pushshift and Reddit API

**Table 2. Differences Between Two Reddit Data Sources.**

<table>
<thead>
<tr>
<th>Downloadable Data Set (from <code>http://files.pushshift.io/reddit/</code>)</th>
<th>Reddit API</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Contains all Reddit content, making it larger than one may want or be able to process</td>
<td>• Caps ability to pull data to 60 items per minute. An item would be a single post or a comment along with the metadata associated with that submission</td>
</tr>
<tr>
<td>• Maybe some missingness (due to download error), but the rate of missing data has declined over time</td>
<td>• More variables at more levels of analysis (e.g., thread, user, subreddit) are available through the API than the downloadable data set</td>
</tr>
<tr>
<td>• The data set structure and its contents changes over time as changes were implemented in Reddit</td>
<td>• Will be more consistent if downloaded all at once</td>
</tr>
<tr>
<td>• Contains variables and content that may have since been removed or changed (not available on Reddit API)</td>
<td>• Depending on the search being conducted and unit of analysis, there are limits to the number of units (i.e., data rows) the system will return (either 100, 1,000, or unlimited)</td>
</tr>
</tbody>
</table>

*Note. API = Application Programming Interface.*
Combination of pushshift and Reddit API

- Search for content through Pushshift, save IDs of the content
- Forward IDs to the Reddit API
- Get updated (meta-) data through the Reddit API
Summary

- Reddit is large and will likely have sufficient data for several types of research objectives and questions.

- The sociodemographic characteristics of redditors are not comparable to the general population but insights into rare subpopulations might be possible.

- Pushshift and the Reddit API have different advantages and disadvantages for collecting Reddit data, but can also be combined.
Thank you for your attention!