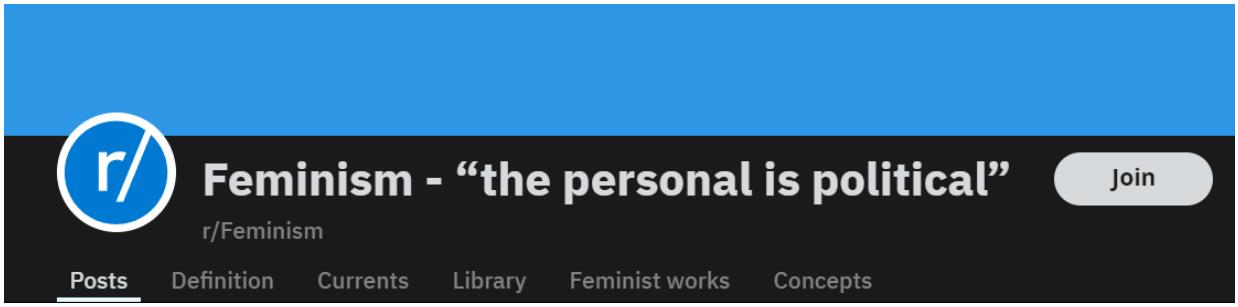
Collecting data from r/Feminism

A report on scraping Reddit for research in the Digital Humanities



Github: https://github.com/Jeli-sh/MA-Digital-Humanities-group-project-Collecting-Data

Group Project Report

Xi Yang(s5269121), Sabina Moulder (s3148637), Jeli Haagsma (s4095030), Shanshan Liu(s5336651), Jing Li (s5374553)

24-01-2023

Table of contents

Introduction and Background

Feminism, a term which is constantly misinterpreted by many due to individuals' understanding through past experiences and their backgrounds. It is becoming an increasingly significant social concern due to global economic development and cultural changes. As society develops, the definition of 'feminism' changes. It started as a way to fight for women's rights and identity in society (Raina, 2020) such as abortion, equal pay and discrimination (Harrison, & Boyd, 2018), and now it changed to a kind of social movement to abrogate sexism and aims to attain full gender equality within laws and practices (Council of Europe Portal, n.d.).

Reddit is the world's largest, user-created social news community. On Reddit, users have essentially no restrictions and they can discuss everything that interests them on what is called a subreddit. Not only that, but Reddit boasts a significant number of daily visitors. For example, in 2021, Reddit had over 52 million daily users, a cumulative total of over 366 million posts, over 2.3 billion comments and a total of 46 billion likes (Staff, 2021). Reddit also boasts an extremely high interaction rate(Staff, 2021). Users on Reddit are more eager to learn, talk, share, and collect information compared to other social media platforms where users browse content and information in solitude. In general, a higher interaction rate corresponds to a higher conversion rate. That users spend a lot of time on the platform is also increased by frequent interactions, helping to tie users together.

We decided to explore the subreddit r/Feminism because we believe we will collect much interesting data that could represent various valuable topics and discussions. Since the number of daily users and posts on Reddit is large enough to give us a good sample for our data collection. At the same time, users are relatively free to speak anonymously on any subreddit, they are less afraid as they are behind screens, which avoids the risk of being blocked for some unknown reasons. For this reason, we could gain some interesting insights into the ways users, in their own online community, propose their topics for discussion.

As a group of female students, also students of Digital Humanities, we want to explore the ways how 'ordinary' social-media users express their views and propose their topics about feminism. Also, we want to apply what we have learned to a specific project, in order to experiment with various data collection tools and support other digital humanities students in doing the same. Therefore, our research question is: *Which topics are mostly discussed by Reddit*

users on subreddits dedicated to feminist topics like r/Feminism? This research question helps us to think about the various ways in which data collection within the Digital Humanities can be done. Our goal is not to provide a clear conclusion to this question. On the contrary, it helps us to engage with digital tools to collect our own data and make sense of the information within a dataset. In addition, this then can guide many other researchers or students within the field. Hence, we want to provide insights into the ways the collection of data can be useful to make sense of the ever-existing questions of the humanities and provide new forms of information and answers.

Our research question is based on the popular book *Data Feminism*, written by Catherine D'Ignazio and Lauren F.Klein (2020). We were inspired and wanted to include a feminist perspective in data research. In their work, D'ignazio and Klein reflect on the ever-increasing importance of data science. By doing this, they ask some crucial questions: *Data Science by whom? Data Science for whom?* and *Data Science with whose interests in mind?* According to what they brought forward in this book, they emphasize feminist perspectives in data science, which have been very much neglected. According to D'ignazio and Klein, the narratives around big data and data science are overwhelmingly white, male and techno-heroic.

In this project about the collection of data from r/Feminism, we want to show our modest ways in which we have thought about these questions and issues. In order to do so, we tried to centralize the collection of data of an online political emancipatory movement, which could be overlooked in other data collection processes within the Digital Humanities. Moreover, the collection of data from this online community brings the possibility to gain an understanding of the proposed topics and discourses of the users on r/Feminism. Furthermore, it helps us to bring more feminist perspectives to the fields of data science and Digital Humanities. According to Posner, most datasets we see in these fields deal with structures of power, like gender and race with a crudeness that would not be accepted in the traditional humanities (2016). One could say that this brief report, on the collection of data to explore discussed feminist topics on Reddit, is a modest contribution to the work of the Critical Digital Humanities. This is an intersectional field that emphasizes anti-racist, postcolonial, queer and feminist perspectives in their use and understanding of digital technologies (Critical Digital Humanities International Conference, 2022).

FAIR principles and our Data Management Plan (DMP)

In every research, one must oblige by data collection law, the FAIR principles and conduct a data management plan (DMP). To begin, we must follow the General Data Protection Regulations (GDPR) as we are collecting data within the European Union (European Commission, n.d.a). Additionally, we constructed a DMP in order to set our goals and ensure that our objectives are consistent as a group. We did this by following the FAIR principles which consist of Findability, Accessibility, Interoperability and Reusability.

For findability we made sure that our dataset is easy to find via indicators such as Uniform Resource Locator (URL) as well as being machine-readable (GO-FAIR, n.d.). In relation to accessibility, our data is open to everyone as we do not include personal data that is identifiable to an individual (European Commission, n.d.b; GO-FAIR, n.d.). Additionally, for interoperability, we checked that we all could gain access to the dataset as well as save the data as a Comma Separated Values (CSV) file for universal access (GO-FAIR, n.d.). Finally, for reusability, we ensure that our data is open for other researchers for future research by describing what and how we collected the data (GO-FAIR, n.d.). Furthermore, in order to complete the DMP we must take the GDPR into consideration. Although in this project we collected data from Reddit which is considered as personal data, we decided not to collect information that may lead to identifying the users in our dataset in order to oblige by the GDPR.

Tutorial: Creation and exploration of the r/Feminism dataset

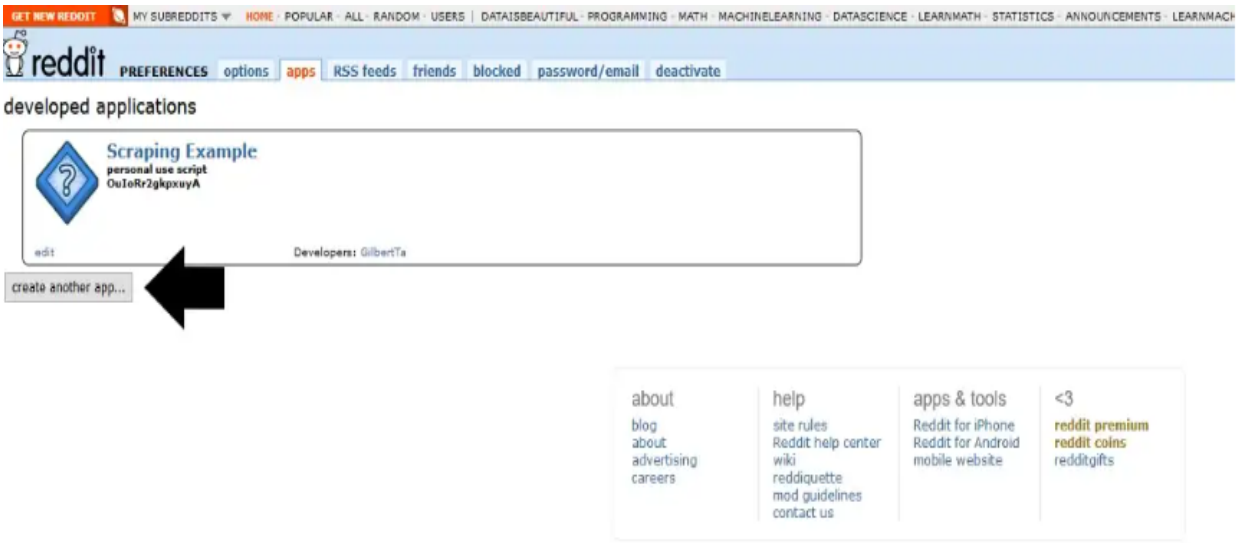
On Reddit, various topics are discussed according to the theme of a particular subreddit. Especially on r/Feminism. With the subtitle “the personal is political,” we already get an understanding of the, sometimes heavily loaded, topics which are proposed by users on this subreddit. In order to get a clear overview of the topics discussed within this sociopolitical digital space, to see general trends as well as thoroughly debated controversial themes, we recognize the importance of gathering data from this particular subreddit itself.

How can this be done in the most user-friendly way possible? As we ourselves are only just beginning to engage with the methods of the Digital Humanities, this is a question that we regard as particularly important. Moreover, the discipline of Digital Humanities itself is growing in importance and so are the researchers who want to understand how to engage with various methods of data collection. For this reason, this tutorial on how we gathered data for our own brief topic analysis, is also created with the intention to serve as a guide for other researchers. This step-by-step tutorial on how we collected our own data is easily reusable for other purposes. Hence, in the next section of this report, we encourage other researchers (or anyone that is interested in Digital Humanities or Reddit research) to create their own Reddit scraper with our active learning exercises.

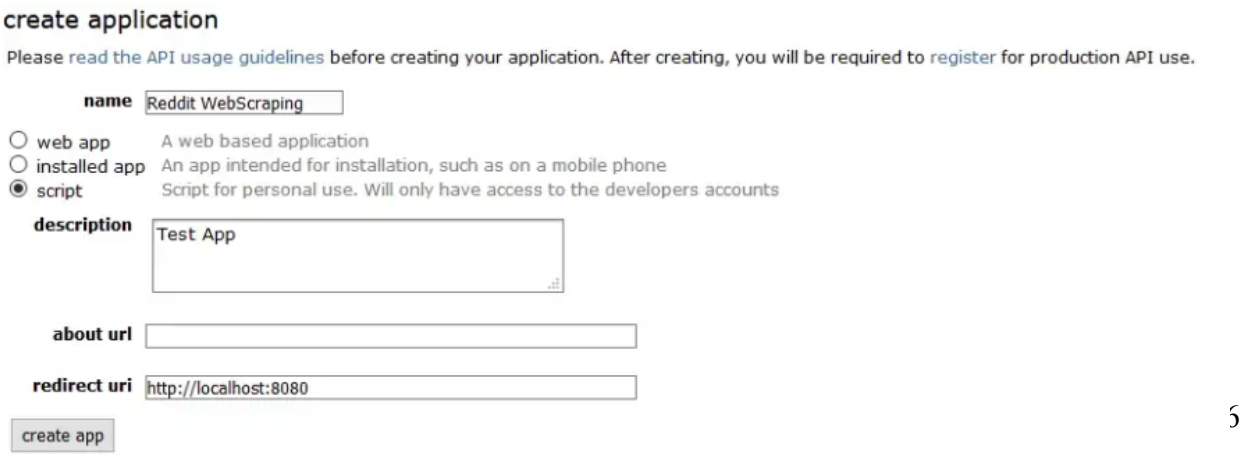
To explore the topics discussed on r/Feminism, we decided to use the Python Reddit API Wrapper (PRAW). As its name suggests, it is a Python wrapper for the Reddit API. PRAW is a very user-friendly tool that allows you to scrape all kinds of data from all subreddits or one of your choosing. In addition, it is also possible to create a bot on Reddit with PRAW. This multi-diverse Python library is in our experience the most suited for research purposes like a topic analysis of a specific subreddit. There are other Reddit API wrappers. These connect to the Pushshift API. Although PSAW and PMAW are also great for scraping a subreddit, we found that PRAW is more stable and reliable for research that is similar to ours. PRAW allows its users to decide which values they want to gather from a subreddit of choice.

Installing PRAW and making a web App:

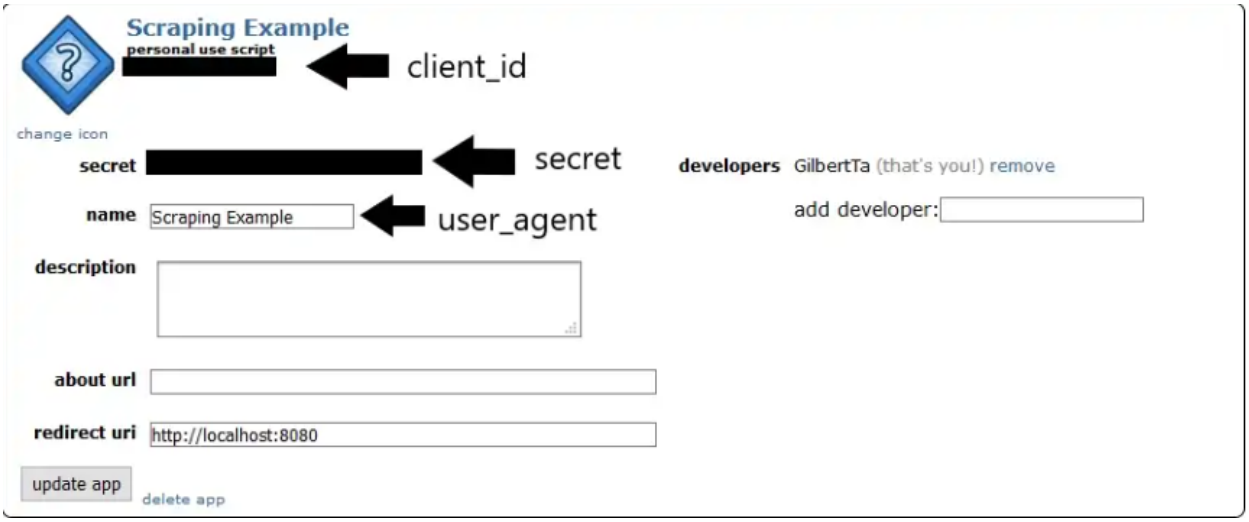
In order to get started, one needs to install PRAW to their device. This can be done with this line of code: `!pip install praw` After this you have to import PRAW and Pandas. To eventually gather data from r/Feminism, an app on the website of Reddit has to be created. This can be done via this link: https://www.reddit.com/prefs/apps. You have to choose 'Create App.' Here a name for the app can be filled in, which is the user agent. You have to provide a description of the App and a redirect URI. As described in the PRAW documentation, for the URI you should choose http://localhost:8080. As these images show, the creation of the Reddit app should look like this (Tanner, 2019):



Now it is time to fill in the information, which you later on have to add to the PRAW commands:



You pick the option: ‘script’. The name of the App is the ‘user-agent’. Other than this image shows, we actually recommend using a user-agent without the words ‘scraping’ or ‘bot.’ Although not always the case, it can be difficult to get authorization. In this next image you see what serves as the ‘user_agent,’ ‘client_secret’ and ‘client_id.’



You can fill this in the following lines of code:

```
reddit_read_only = praw.Reddit(client_id="",        #your client id
                               client_secret="",    #your client secret
                               user_agent="")       # your user agent
subreddit = reddit_read_only.subreddit("Feminism") #The name of the subreddit, in our case: (r/)Feminism.

#With these lines of code you can check if PRAW is connected to the subreddit of your choice.

# Display the name of the Subreddit
print("Display Name:", subreddit.display_name)

# Display the title of the Subreddit
print("Title:", subreddit.title)

# Display the description of the Subreddit
print("Description:", subreddit.description)
```

After filling in the client_id, client_secret and user_agent, you can choose the subreddit you want to scrape. If you want to scrape all subreddits you can simply write down “all.” In our case, this was “Feminism.” After this, we ran a few simple commands to see if we were actually connected to the subreddit.

Creating a Pandas DataFrame:

After we got authorization and were connected to r/Feminism, it was time to decide our values and save them in a Pandas Dataframe:

```
posts = []

for post in subreddit.hot(limit=2000):
    posts.append([post.title, post.score, post.id, post.subreddit, post.url, post.num_comments, post.selftext, post.created])
feminism_df = pd.DataFrame(posts,columns=['title', 'score', 'id', 'subreddit', 'url', 'num_comments', 'body', 'created'])
print(feminism_df)
```

We created an empty list in which we could store our chosen values and eventually save it in a Dataframe. For our project, we decided to collect all 'hot posts'. The reason for doing this is that r/Feminism is not a very large subreddit so in this way we could already gather the most prominent posts from the past year. Note that we decided on the limit of the collection, which we set at 2000. However, we only gathered around 800 posts. For larger subreddits, the limit could be more applicable as you would probably collect way more posts. It is also possible to gather the 'top posts.' Instead of 'subreddit.hot()' you then write 'subreddit.top().' As an argument, you can give either 'year,' 'month,' or 'week' in order to gather the top posts you want. One limitation of using PRAW is that you cannot specify a timeframe from which you want to collect posts.

Moving on, we decided on our values. In our case we wanted to collect the title of the thread (title), the number of upvotes (score), a unique id (id), the name of the subreddit (subreddit), the url of the post or other content in the post (url), the number of comments (num_comments), the text within the thread (selftext) and when it was created (created). After this, you specify the name of the columns in the Dataframe. If you are satisfied with the Dataframe as it is, you can save it as a CSV file.

```
feminism_df.to_csv("feminism reddit dataset.csv", index=True)
```

Inspecting and cleaning the dataset with Pandas:

For our project, we wanted to inspect and change a couple of things in our dataset. First of all, as we want to explore some trends of the topics over time, it was important that we added clear dates to our dataset. Our 'created' column only showed a code.

```
import datetime as dt
feminism_df['date'] = pd.to_datetime(feminism_df['created'], utc=True, unit='s')
```

```
feminism_df['date']
```

In addition, the ‘created’ column was now no longer needed. We also noticed that there was an unwanted column ‘Unnamed: 0’, which is the same as the index of the Dataframe. Therefore we could both drop these columns. As you can see in the pictures below, the date column also contains the time in one column. It is wise to split this into separate columns. Especially when you want to make some visualizations. We did this later on in excel but you can also split the columns in python and create a new column.

```
feminism_df = feminism_df.drop(columns=['created'])
feminism_df = feminism_df.drop(columns=['Unnamed: 0'])
```

To inspect the dataset further we wanted to take a closer look at the value types and possible null-values:

```
feminism_df.dtypes
```

```
Unnamed: 0        int64
title            object
score             int64
id               object
subreddit        object
url              object
num_comments      int64
body             object
created         float64
date             object
dtype: object
```

Now lets see how many observations the dataset has:

```
feminism_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 796 entries, 0 to 795
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Unnamed: 0    796 non-null    int64
 1   title         796 non-null    object
 2   score         796 non-null    int64
 3   id            796 non-null    object
 4   subreddit     796 non-null    object
 5   url           796 non-null    object
 6   num_comments  796 non-null    int64
 7   body          282 non-null    object
 8   created       796 non-null    float64
 9   date          796 non-null    object
dtypes: float64(1), int64(3), object(6)
memory usage: 62.3+ KB
```

For the ‘body’ column we figured out that there were some null-values. This is not a surprise. Reddit users could make posts with only a title. In most cases, these posts include a link to a video or image.

We wanted to inspect which posts actually had missing values. As the url’s show they indeed include links to videos or images:

```
feminism_df[feminism_df.isnull().any(axis=1)]
```

	Unnamed: 0	title	score	id	subreddit	url	num_comments	body	created	date
1	1	On New Year's Eve, Iranian women express their...	189	10134y8	Feminism	https://v.redd.it/r6d7rq8b4k9a1	3	NaN	1.672633e+09	2023-01-02 04:23:21+00:00
2	2	She led two historic victories for abortion ri...	204	101106p	Feminism	https://www.theguardian.com/world/2023/jan/01/...	4	NaN	1.672627e+09	2023-01-02 02:37:39+00:00
3	3	Amal(12) | Pregnant child bride: A story of ma...	84	1014ivy	Feminism	https://v.redd.it/w4k473wkyg9a1	6	NaN	1.672638e+09	2023-01-02 05:35:48+00:00
5	5	Women are more critical of female toplessness ...	146	100vlhd	Feminism	https://www.psypost.org/2022/10/women-are-more...	39	NaN	1.672613e+09	2023-01-01 22:35:35+00:00
6	6	This has bothered me for a long time	2609	100c37f	Feminism	https://i.redd.it/2xwf53zf3d9a1.png	88	NaN	1.672548e+09	2023-01-01 04:46:51+00:00

Due to the length of this project, we only focused on the title of the threads. However, it would be interesting, for various reasons, to also take a closer look at the actual texts in the thread. Therefore, we made a filter of our DataFrame with posts that do not have images or videos.

```
body_df = feminism_df[feminism_df['body'].notna()]
body_df
```

	Unnamed: 0	title	score	id	subreddit	url	num_comments	body	created	date
0	0	This is a comprehensive list of resources for ...	2547	phrcrn	Feminism	https://www.reddit.com/r/Feminism/comments/phr...	236	**Update** I guess I've been mass reported for...	1.630761e+09	2021-09-04 13:15:02+00:00
4	4	My brother watches Hamza and it's scaring me	76	1010yt6	Feminism	https://www.reddit.com/r/Feminism/comments/101...	46	My little brother (14M) listens to a lot of "r...	1.672627e+09	2023-01-02 02:35:44+00:00
7	7	Being a teen girl sucks	236	100ofn7	Feminism	https://www.reddit.com/r/Feminism/comments/100...	12	As I am in my last year of highschool and am g...	1.672594e+09	2023-01-01 17:28:45+00:00

The last part of this process of inspecting our dataset, was taking a closer look at our values:

```
feminism_df.describe(include='all')
```

	Unnamed: 0	title	score	id	subreddit	url	num_comments	body	created	date
count	796.000000	796	796.000000	796	796	796	796.000000	282	7.960000e+02	796
unique	NaN	792	NaN	796	1	792	NaN	282	NaN	795
top	NaN	Yes we can.	NaN	phrcrn	Feminism	https://theconversation.com/women-in-antarctic...	NaN	**Update** I guess I've been mass reported for...	NaN	2022-12-09 03:43:49+00:00
freq	NaN	3	NaN	1	796	2	NaN	1	NaN	2
mean	397.500000	NaN	258.898241	NaN	NaN	NaN	23.077889	NaN	1.668580e+09	NaN
std	229.929699	NaN	442.125368	NaN	NaN	NaN	40.938321	NaN	2.703460e+06	NaN
min	0.000000	NaN	0.000000	NaN	NaN	NaN	0.000000	NaN	1.630761e+09	NaN
25%	198.750000	NaN	15.000000	NaN	NaN	NaN	2.000000	NaN	1.666444e+09	NaN
50%	397.500000	NaN	84.000000	NaN	NaN	NaN	6.000000	NaN	1.668573e+09	NaN
75%	596.250000	NaN	328.000000	NaN	NaN	NaN	23.000000	NaN	1.670594e+09	NaN
max	795.000000	NaN	3388.000000	NaN	NaN	NaN	278.000000	NaN	1.672650e+09	NaN

This simple line of code gives us the opportunity to already gather a lot of information about our dataset. For instance, we can see the average score (upvotes) which is around 259 per post. We can see the average amount of comments per post is 23. The maximum number of upvotes is 3388. The maximum number of comments is 278. In comparison to other subreddits, these numbers are not that high. This could indicate that feminist topics are maybe not popular or discussed in great amounts on this particular subreddit or even on Reddit in general.
We also see that the 'title' column has four values that are not unique. This means that there could be a few reposts.

```
feminism_df[feminism_df.duplicated(subset=['title'])]
```

	title	score	id	subreddit	url	num_comments	body	date
118	Yes we can.	813	zqn682	Feminism	https://i.redd.it/021o2vz4r17a1.jpg	8	NaN	2022-12-20 12:29:06+00:00
121	Women Heavily Underrepresented in Political De...	3	zrfcid	Feminism	https://web-mind.io/artificial-intelligence/wo...	0	NaN	2022-12-21 09:15:24+00:00
210	Yes we can.	895	zfibca	Feminism	https://i.redd.it/2csazo9cck4a1.jpg	19	NaN	2022-12-07 23:47:35+00:00
774	Women in Antarctica face assault and harassmen...	9	xwlrzs	Feminism	https://theconversation.com/women-in-antarctic...	0	NaN	2022-10-05 20:43:51+00:00

In our case, we see that this is not actually the case. The "*yes we can*" posts contain different URLs and the other posts have different titles. We do not have to remove these rows. However, it could be that there are unnoticed reposts. A researcher has to decide for themselves whether the rows have to be removed or not. To see the actual programming of our web scraper as well as the inspection and cleaning of our dataset, please go to our Github page: https://github.com/Jeli-sh/MA-Digital-Humanities-group-project-Collecting-Data/blob/main/Collecting%20data%20Group%20Project%20-%20building%20the%20reddit%20scraper.ipynb.

Exploration of the r/Feminism dataset with Flourish

In order to explore our dataset and, more importantly, create a snapshot of the topics discussed on r/Feminism, we used Flourish. Flourish is an online tool which easily turns your data into stunning charts, maps and interactive stories. It is effective in engaging with your audience when visualizing a dataset. Per usual, there are many tools available to make data visualizations. Seaborn, Mathplotlib and Plotly to name a few. For an intensive data analysis, we recommend looking into other visualization tools. Flourish is a great tool to make aesthetically pleasing visualizations, which can easily be shared online. Yet, it does have its limitations with regard to an extensive functionality for data alteration, analysis and visualization. For our project, which is

mainly focused on the data collection process, a tool like Flourish is actually very sufficient. We only wish to explore our dataset and the topics proposed on r/Feminism. Moreover it helps to make our visualizations interactive and easily accessible in our report as well as on our Github, which allows us to demonstrate some brief possibilities of a dataset created by a Reddit scraper with PRAW.

For this reason, we made a scatterplot on Flourish. To see the data visualization and the interactive elements, please go to: https://public.flourish.studio/visualisation/12415427/. In the scatterplot, you can see the date on the x axis and the number of comments on the y axis. The amount of upvotes is visible with color. The darker the dot, the more upvotes a post got.

This scatterplot can already give us some interesting insights into the discussed topics on r/Feminism. For instance, the dot which got the highest amount of comments with the title: “Boyfriend thinks I’m bias[ed] for not wanting to watch an Andrew Tate video” did not get the highest amount of upvotes. How could this be? Andrew Tate is a topic which is heavily debated in the last couple of months. This explains the highest amount of comments. The fact that the post did not get as many upvotes as other posts, could be because the topic is very controversial. However considering this is a subreddit dedicated to feminist narratives only, it would be more probable that users downvote posts with topics they do not like in general instead of actually agreeing with the argument made in the post.

One of the dots in the graph represents a post that respectively got the highest amount of upvotes and comments. The post has the title: “This is a comprehensive list of resources for those in need of an abortion.” The thread was posted on the 4th of September in 2021. It is no surprise that a subreddit like r/Feminism would discuss these heavily debated topics. What is even more interesting is that this thread was posted way before the overturning of Roe v. Wade. A political event which caused a lot of debate on Reddit and elsewhere, about women's rights and abortions. This could give an indication that the topic of abortion takes a central place at all times on the r/Feminism subreddit.

These are only a few interpretations one could make about the discussed topics on r/Feminism according to this graph. One thing that is particularly interesting about these topics is the division between heavily charged political topics and light-hearted everyday subjects. One of the threads which is posted on the 4th of November in 2022 with a large number of upvotes (2754) has the title:

“This is Nika Shakarami. She went missing while being chased by security forces during the protests in Iran. Her body was handed back to her family with a broken nose and skull. The security forces then invaded her funeral and stole her body so they could bury her themselves and arrested her aunt.”

This horrible story of the protests in Iran shows the seriousness of the topics discussed on this subreddit. However, another thread which was posted around the same month, with also a decent amount of upvotes has the title: “ladies: PLEASE STOP BRINGING YOUR BOYFRIEND AND HUSBANDS INTO VICTORIA'S SECRET.” Furthermore, when looking at the graph you can see that there exist quite a few posts on the subreddit asking for movie or book recommendations with female leads. These are again followed up with posts about violence against women or political discussions about women's position in the labor market.

Although not even close to a full in-depth data analysis, by looking at this graph we already get some understanding of the discussed topics on the subreddit. By hovering over all the dots, we see a variety of topics: sociopolitical or activist charged posts and discussions, but also everyday problems that women face. By looking at the number of upvotes and comments we get some ideas about topics which are popular, overseen or controversial. Following up by taking a closer look at these actual posts one could already make some preliminary conclusions or build a hypothesis around these findings.

Recommendations for further data analysis of the r/Feminism dataset

We do want to provide some recommendations for further, more in-depth, data analysis of our dataset. Although we only discussed a general overview of the topics on r/feminism according to our graph, there are several ways to do an extensive topic modeling which can be applied to the r/Feminism dataset or the one you created yourself (see the next section).

By using data science and Natural Language Processing approaches in Python it is possible to engage in critical analysis of online discourse communities such as r/Feminism. A popular approach to explore an online community is topic modeling. This is a type of statistical modeling that allows you to discover abstract topics which occur in a collection of documents.

As Tom van Nuenen explains in his guide to analyzing Reddit communities with Python, topic modeling is used frequently as a text-mining tool to find hidden semantic structures in textual data (2022). As van Neunen further explains, it is an unsupervised machine learning technique that allows us to scan a set of documents, detect word and phrase patterns within them, and automatically cluster word groups and similar expressions that best characterize a set of documents. There are many ways of doing topic modeling and one of the most used is Latent Dirichlet Allocation (LDA). In Python you can use the `gensim` package to create topic models. You can explore these models by using `PyLDAvis` and eventually evaluate and visualize the coherence of these models. For further information on this method please look at van Nuenen's guide: https://tomvannuenen.medium.com/analyzing-reddit-communities-with-python-part-5-topic-modeling-a5b0d119add.

One thing we have not discussed in this tutorial are the usernames of Reddit users. We have not added them to our dataset because we did not want to deal with several privacy regulations which you can read in our Data Management Plan (DMP). Nonetheless, it still could be very interesting to look at the ways users of Reddit interact with each other on the r/Feminism subreddit. For instance, Daniele Linkevicus de Andrade and Demival Vasques Filho did an extensive network analysis of users on several subreddits dedicated to the topics of history (2021). They noticed one fascinating thing about this Reddit community: the plurality of proposals. As they argue, all kinds of Reddit users could engage in and propose topics. What is interesting, is that they only analyzed interactions between users who engaged in conversations. Although the r/Feminism dataset does not contain any information about comments, these can be added as values as well. You can also manually change usernames and give them their own ID. Still, what is promising about doing a network analysis, is that you do not even have to represent the actual usernames in the network analysis, as you can represent them as nodes. Here, you could gain deeper insights into the discussions of specific topics on r/Feminism. In addition, you could see which users within the community actually have authority in proposing these topics.

Active learning exercises: Creating your own Reddit scraper

Contemporary feminism is playing out in increasingly digitized ways as women from all over the world use the Internet and social media to discuss, disseminate and debate key feminist ideas in spaces dedicated to women's perspectives. (Mowle, A, 2021). Since social media has become a significant platform for people to debate gender-related issues, we decided to scrape data from r/Feminism in order to explore the topics proposed by the users. However, as Reddit serves as a big social media corpus, researchers could analyze the platform in various ways. A post in Reddit must be made in an assigned "subreddit" for any and every interest. Within each of these subreddits exists a unique community with a distinct subculture. (Anderson, K. E., 2015).

There are various gender-oriented subreddits present besides r/Feminism, such as r/MensRight and r/AskWomen. Different gendered groups present distinct points of view, which are also useful in exploring gender-related issues. Researchers who want to dig deeper into gender issues through a social media analysis can create their own Reddit scraper by following our tutorial. Moreover, our tutorial as well as these active learning exercises give researchers an opportunity to scrape any subreddit they like. Delving deeper into gender related issues or any other topic in the humanities, can be done by using our example assignments which serve as a template for any researcher.

Our tutorial may be valuable to those who wish to present a data mining analysis of gender-related social movements using Reddit. Since there are more than a million different subreddits, our decision to use the data from the larger subreddit r/Feminism is significant, but it also disregards the opinions of the smaller member groups. So learners can select subreddits of interest and follow the guidelines we provide for scraping data on Reddit.

Subreddits are identified by the naming pattern "/r/subreddit," which is used in the URL for direct access, making it simple to approach. By following our tutorial, researchers can simply replace the name of the subreddit in this step to get the desired data from other subreddits.

```
reddit_read_only = praw.Reddit(client_id="",       #your client id
                               client_secret="",   #your client secret
                               user_agent="")      # your user agent
subreddit = reddit_read_only.subreddit("Feminism") #The name of the subreddit, in our case: (r/)Feminism.
```

As for us, we restricted our search to the hot posts in the r/Feminism subreddit. But learners are free to select their own data scale. They can choose the time restriction for the top posts or they can go one step further and set the limit themselves, not only for the hotposts; it depends on how much data they want to scrape.

```
posts = []

for post in subreddit.hot(limit=2000):
    posts.append([post.title, post.score, post.id, post.subreddit, post.url, post.num_comments, post.selftext
feminism_df = pd.DataFrame(posts,columns=['title', 'score', 'id', 'subreddit', 'url', 'num_comments', 'body'
print(feminism_df)
```

For more specific coding, we've made our notebook publicly available on GitHub. Learners may easily follow the tutorial and example assignments which can be found in the file: *Workbook-Making your own scraper.ipynb*. In this step by step guide the learner is encouraged to gather the required subreddit data of their choosing by using our notebook.

Conclusion

In this report, we presented how we used digital tools to gather our own data from Reddit and interpret the data in our dataset. Moreover, we provided a tutorial and exercises for other researchers to do the same.

The large amount of data from Reddit is relatively accessible, but collecting them all takes time. Therefore we explored the ways of building our own scraper by using PRAW. In order to narrow down the data that we could gather, we focused on the hot posts shared on the subreddit r/Feminism. By doing so, we tried to illustrate how this method of data collection was efficient in gathering the data, whereas our research question could guide us in this process. In addition, our research question helped us to explore the dataset we created and show others the possibilities of a dataset created with PRAW.

As mentioned in the introduction, the purpose of this research is not to provide definitive answers to the research question. Rather our goal was to focus on how we could collect our data, what we can learn from the dataset with regard to the proposed topics and for others to do the same. In addition, we used Flourish to dig deeper into our data story in order to understand the data we collected. We created a scatterplot using Flourish, which can already provide us with some intriguing insights about the topics discussed on r/Feminism, despite not even coming close to a comprehensive in-depth data analysis. We obtained some insights into popular topics by observing the amount of upvotes and the amount of comments, which focus on a variety of political, legal and social equality issues affecting women, such as abortion and other societal problems, as well as common challenges that women experience on a daily basis. Moreover, we provided insights into some methods which could be used to do further, more in-depth data analysis.

Social media is becoming an increasingly common source of data. For various reasons, Reddit is an attractive data source. Its popularity means that there is a large amount of data available for research. Furthermore, most subreddits are open to the public, and Reddit provides an anonymous environment in which Redditors can express their ‘real’ ideas (Amaya, Bach, Keusch & Kreuter, 2021). As more researchers use Reddit, we believe that data-driven findings will be used on an ever increasing scale within the Humanities.

References

Amaya, A., Bach, R., Keusch, F., & Kreuter, F. (2021). New data sources in social science research: Things to know before working with Reddit data. *Social science computer review*, *39*(5), 943-960.

Anderson, K. E. (2015). Ask me anything: what is Reddit?. Library Hi Tech News.

Andrade, D. L. de, & Filho, D. V. (2021, November 12). Diving into Reddit: authority networks in history forums. *Digital Humanities Lab*. https://dhlab.hypotheses.org/2297

Council of Europe Portal. (n.d.) Feminism and women's rights movements—Gender matters—Publi. Coe. Int. (n.d.). Gender Matters. Retrieved 15 December 2022. From: https://www.coe.int/en/web/gender-matters/feminism-and-women-s-rights-movements

European Commission. (n.d.a) Data Protection. Retrieved 23 January 2023. From: https://commission.europa.eu/law/law-topic/data-protection_en.

European Commission. (n.d.b) Data Protection. Retrieved 23 January 2023. From: https://commission.europa.eu/law/law-topic/data-protection/reform/what-personal-data_en

GO-FAIR. (n.d.). FAIR Principles. Retrieved 23 January 2023. From: https://www.go-fair.org/fair-principles/

Harrison, K., & Boyd, T. (2003). *Understanding political ideas and movements*. Manchester University Press.

Khan, A., & Golab, L. (2020). Reddit Mining to Understand Gendered Movements. In EDBT/ICDT Workshops.

Nuenen, T. van. (2022, March 30). *Analyzing Reddit communities with Python — Part 5: topic modeling*. Medium. https://tomvannuenen.medium.com/analyzing-reddit-communities-with-python-part-5-topic-modeling-a5b0d119add.

Posner, M. (2016). What's Next: The Radical, Unrealized Potential of Digital Humanities. In M.K. Gold & L.F. Klein (Eds.), *Debates in the Digital Humanities* 2016 (32-41) University of Minnesota Press.

Projects | *EADH - The European Association for Digital Humanities*. (n.d.). Retrieved October 25, 2022. From: https://eadh.org/projects.

Raina, J. A. (2020). Feminism: An Overview. Retrieved 15 December 2022. From: https://www.researchgate.net/publication/339939198_Feminism_An_Overview#:~:text=Abstract,and%20social%20equality%20of%20sexes.

Staff (2022). *Reddit recap 2021*, *Upvoted*. Accessed: January 18, 2023. Available at: https://www.redditinc.com/blog/reddit-recap-2021.

Tanner, G. (2021, December 7). *Scraping Reddit data - Towards Data Science*. Medium. https://towardsdatascience.com/scraping-reddit-data-1c0af3040768.