Homework #3. Dataset Collection and Understanding.

Student ID - 17. Sofiia Zakharuk, CS-2.

Team #17.

Option 1 — Reddit Dataset.

a. Dataset description and source:

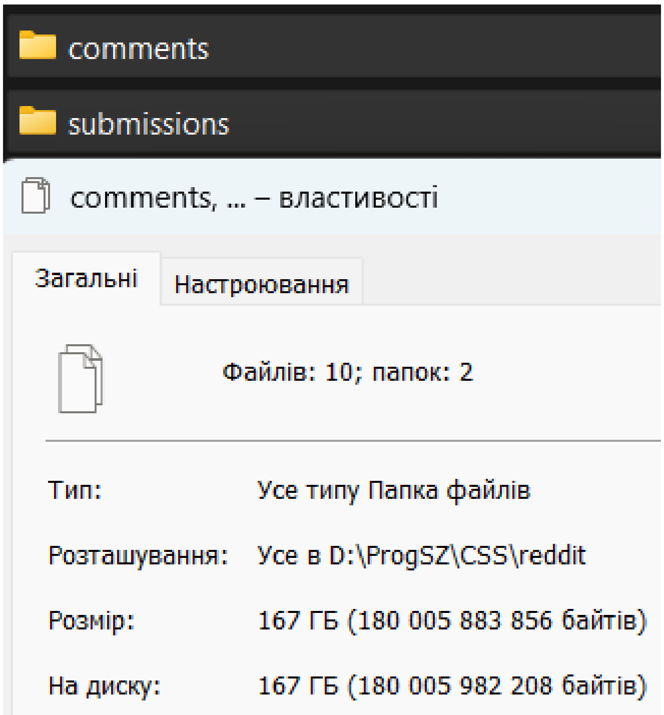
We have chosen 5 months to demonstrate changes in the prevalence and adoption of AI in artistic professions – 09.2018, 09.2021, 05.2023, 10.2023, 09.2024 – To ensure maximum comparability of data and minimize the influence of seasonal factors on the results, September was chosen as a single reference month for 2018, 2021, 2024 studied years. This month is characterized by stable online activity and marks the beginning of the academic and business season, making it representative for the analysis of professional and creative discourses. May and October were chosen to better show the difference after the outbreak of AI mass popularity in 2023.

name	AI in creative industries
subreddits	Art, illustration, DigitalArt, painting, graphic_design, photography, writing, copywriting, blender, gamedev

Source:

```
@article{,
title= {Reddit comments/submissions 2005-06 to 2025-06},
journal= {},
author= {stuck_in_the_matrix, Watchful1, RaiderBDev},
year= {2025},
url= {https://academictorrents.com/details/30dee5f0406da7a353aff6a8caa2d54fd01f2ca1},
abstract= {Reddit comments and submissions from 2005-06 to 2025-06 collected by pushshi
ft and u/RaiderBDev.

These are zstandard compressed ndjson files. Example python scripts for parsing the dat
a can be found here https://github.com/Watchful1/PushshiftDumps

The more recent dumps are collected by u/RaiderBDev},
keywords= {'reddit'},
terms= {},
license= {},
superseded= {}
}
```

Downloading:



comments

submissions

comments, ... – властивості

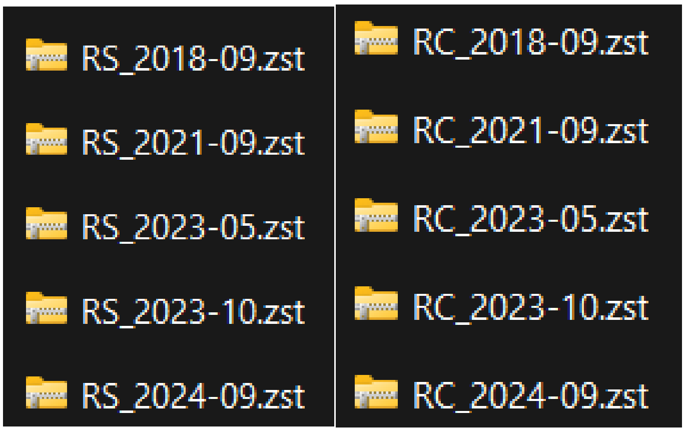
Файлів: 10; папок: 2

Тип: Усе типу Папка файлів

Розташування: Усе в D:\ProgSZ\CSS\reddit

Розмір: 167 ГБ (180 005 883 856 байтів)

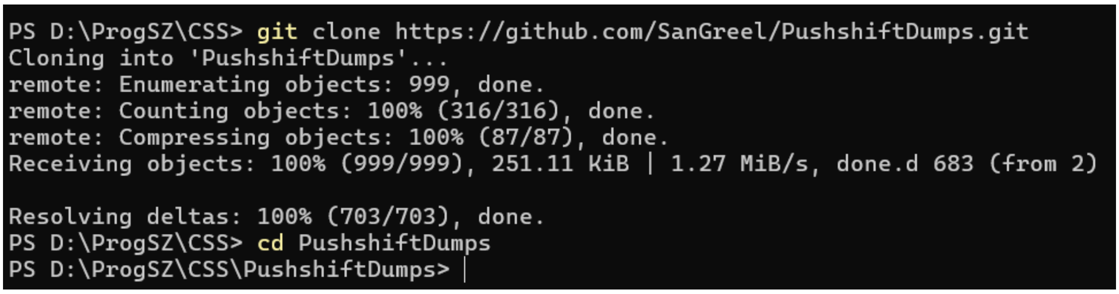
На диску: 167 ГБ (180 005 982 208 байтів)



RS_2018-09.zst
RS_2021-09.zst
RS_2023-05.zst
RS_2023-10.zst
RS_2024-09.zst

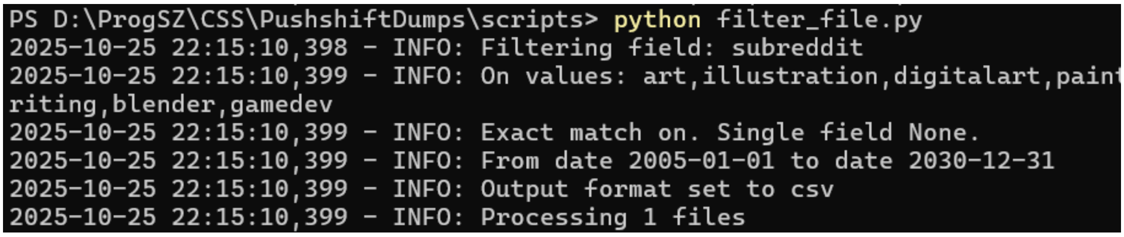
RC_2018-09.zst
RC_2021-09.zst
RC_2023-05.zst
RC_2023-10.zst
RC_2024-09.zst

Cloning PushshiftDumps repository:



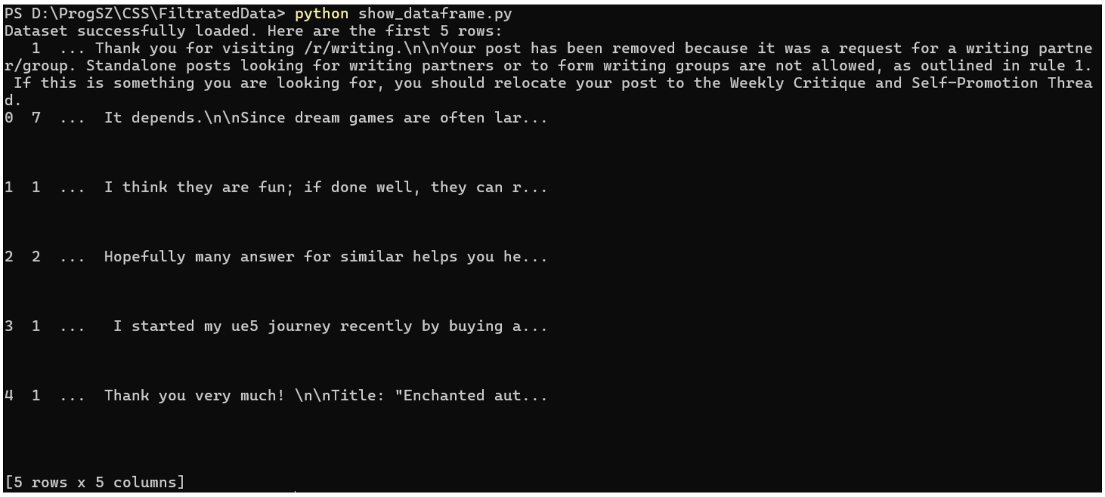
```
PS D:\ProgSZ\CSS> git clone https://github.com/SanGreel/PushshiftDumps.git
Cloning into 'PushshiftDumps'...
remote: Enumerating objects: 999, done.
remote: Counting objects: 100% (316/316), done.
remote: Compressing objects: 100% (87/87), done.
Receiving objects: 100% (999/999), 251.11 KiB | 1.27 MiB/s, done.d 683 (from 2)

Resolving deltas: 100% (703/703), done.
PS D:\ProgSZ\CSS> cd PushshiftDumps
PS D:\ProgSZ\CSS\PushshiftDumps> |
```

Filtering by subreddits:



```
PS D:\ProgSZ\CSS\PushshiftDumps\scripts> python filter_file.py
2025-10-25 22:15:10,398 - INFO: Filtering field: subreddit
2025-10-25 22:15:10,399 - INFO: On values: art,illustration,digitalart,paint
riting,blender,gamedev
2025-10-25 22:15:10,399 - INFO: Exact match on. Single field None.
2025-10-25 22:15:10,399 - INFO: From date 2005-01-01 to date 2030-12-31
2025-10-25 22:15:10,399 - INFO: Output format set to csv
2025-10-25 22:15:10,399 - INFO: Processing 1 files
```

b. Screenshot of loaded DataFrame:

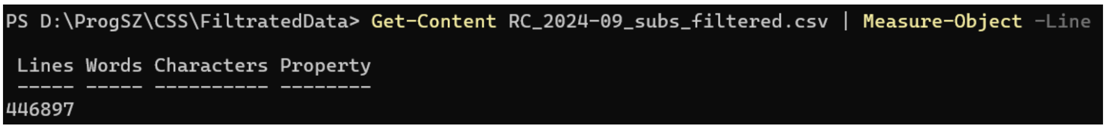


```
PS D:\ProgSZ\CSS\FiltratedData> python show_dataframe.py
Dataset successfully loaded. Here are the first 5 rows:
   1  ... Thank you for visiting /r/writing.\n\nYour post has been removed because it was a request for a writing partne
r/group. Standalone posts looking for writing partners or to form writing groups are not allowed, as outlined in rule 1.
 If this is something you are looking for, you should relocate your post to the Weekly Critique and Self-Promotion Threa
d.
0  7  ...  It depends.\n\nSince dream games are often lar...


1  1  ...  I think they are fun; if done well, they can r...


2  2  ...  Hopefully many answer for similar helps you he...


3  1  ...   I started my ue5 journey recently by buying a...


4  1  ...  Thank you very much! \n\nTitle: "Enchanted aut...


[5 rows x 5 columns]
```

c. Link to our dataset:
https://drive.google.com/drive/folders/1tLfxoc8LADqZDDmII0ZvutbVpkorcvh1?usp=sharing

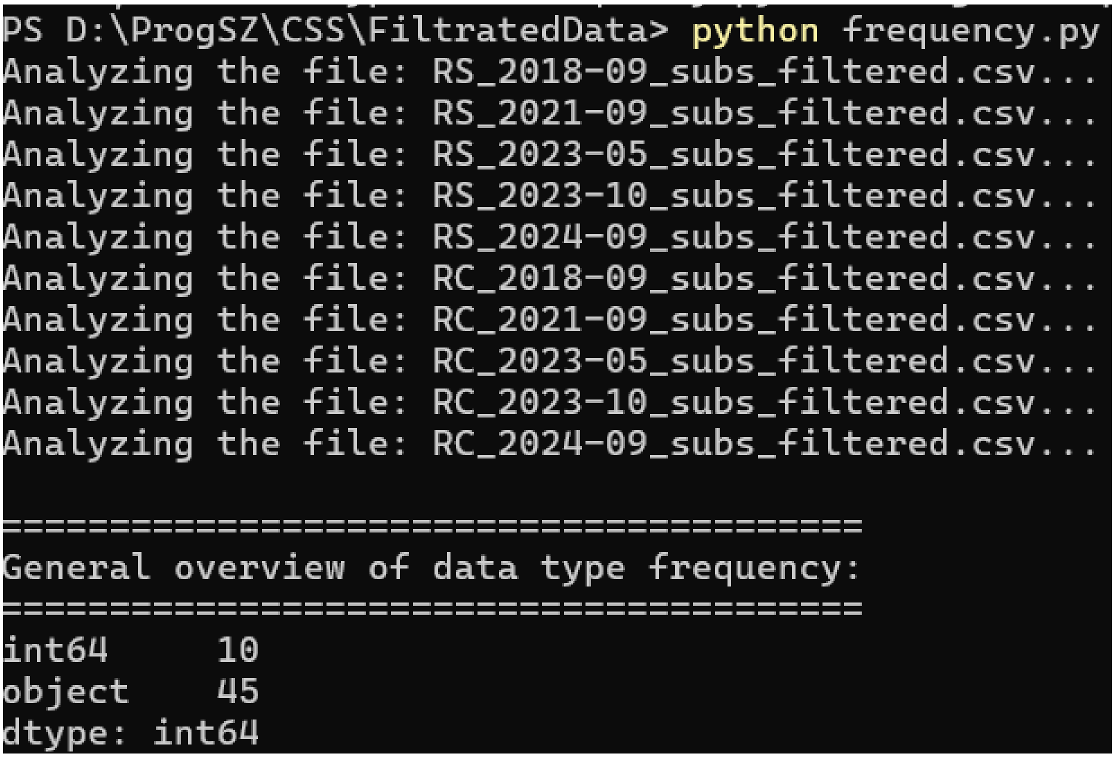
d. Table with data size (MB) and row count:

Name	Size	Rows
RC_2018-09_subs_filtered	61.9MB	323974
RC_2021-09_subs_filtered	81MB	433082
RC_2023-05_subs_filtered	83MB	425859
RC_2023-10_subs_filtered	94.7MB	455675
RC_2024-09_subs_filtered	93.5MB	446897
RS_2018-09_subs_filtered	8.33MB	44190
RS_2021-09_subs_filtered	11.3MB	63463
RS_2023-05_subs_filtered	11.3MB	62725
RS_2023-10_subs_filtered	13.5MB	74745
RS_2024-09_subs_filtered	13.5MB	71932



```
PS D:\ProgSZ\CSS\FiltratedData> Get-Content RC_2024-09_subs_filtered.csv | Measure-Object -Line

 Lines Words Characters Property
 ----- ----- ---------- --------
446897
```

e. Frequency overview of data types:



```
PS D:\ProgSZ\CSS\FiltratedData> python frequency.py
Analyzing the file: RS_2018-09_subs_filtered.csv...
Analyzing the file: RS_2021-09_subs_filtered.csv...
Analyzing the file: RS_2023-05_subs_filtered.csv...
Analyzing the file: RS_2023-10_subs_filtered.csv...
Analyzing the file: RS_2024-09_subs_filtered.csv...
Analyzing the file: RC_2018-09_subs_filtered.csv...
Analyzing the file: RC_2021-09_subs_filtered.csv...
Analyzing the file: RC_2023-05_subs_filtered.csv...
Analyzing the file: RC_2023-10_subs_filtered.csv...
Analyzing the file: RC_2024-09_subs_filtered.csv...

========================================
General overview of data type frequency:
========================================
int64     10
object    45
dtype: int64
```

f. Brief interpretation of the dataset context:

The 10 filtered CSV files represent a specialized and filtered data corpus derived from the public Reddit archives.

The data was sourced exclusively from large, topic-specific subreddits (such as r/Art, r/illustration, r/graphic_design, r/writing and others). This allows us to assume that the authors of the comments and submissions are primarily artists, designers, writers, and other creative professionals or enthusiasts. This focus ensures we are capturing the authentic voice and immediate reaction of the communities most directly impacted by this technological shift.

The primary value of our dataset lies in its representation of five distinct time periods, each capturing a different stage of the public discourse on generative AI.

September 2018: During this period, AI's role in creative fields was largely experimental and niche. We expect these files to contain a small volume of data, reflecting theoretical discussions about procedural generation, GANs, and conceptual ideas rather than practical, widespread tools.

September 2021: With the release of GPT-3, text generation technology became more prominent. In these files we anticipate seeing growing awareness and the first discussions about practical applications (like AI writer), alongside initial concerns regarding academic integrity and plagiarism. The data volume is being greater than in 2018.

May & October 2023: By this time, tools like Midjourney, Stable Diffusion, and ChatGPT had become global phenomena. These files are expected to reflect an explosive surge in urgent and often polarized discussions. The data should contain conversations about specific tools, widespread community debate on job displacement, copyright infringement, AI ethics, and the very definition of art. This will likely be the largest and richest portion of our dataset.

September 2024: This period represents the "new normal." We anticipate the discourse will have shifted from shock and debate towards integration and pragmatism. The data will likely reflect discussions about practical workflows, new integrated tools (like Adobe's Generative Fill), and the establishment of community rules regarding AI-generated content.

Also, our dataset is separated into two file types, mirroring the communication structure of Reddit: submissions(posts) and comments. Analyzing posts will show what topics were most important to the community. And analyzing comments will reveal the community's attitude toward the initiated topics, the level of engagement, and the overall sentiment.

g. Link to git repo: https://github.com/Sonafi3/Computational_Social_Science_2025