Homework #3. Dataset Collection and Understanding.

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

Team #8.

Option 1 — Reddit Dataset.

a. Dataset description and source:

We have chosen 6 months to demonstrate changes in the prevalence and adoption of AI in educational processes – 09.2018, 09.2021, 05.2023, 10.2023, 09.2024, 06.2025 – September is a standard as a month of the beginning of the school year and the corresponding revival of conversations about learning process. May and October were chosen to better show the difference after the outbreak of AI mass popularity in 2023. And June as the last month possible to analyze in 2025.

name	AI in education
subreddits	education, highereducation, University, school, Learning, college

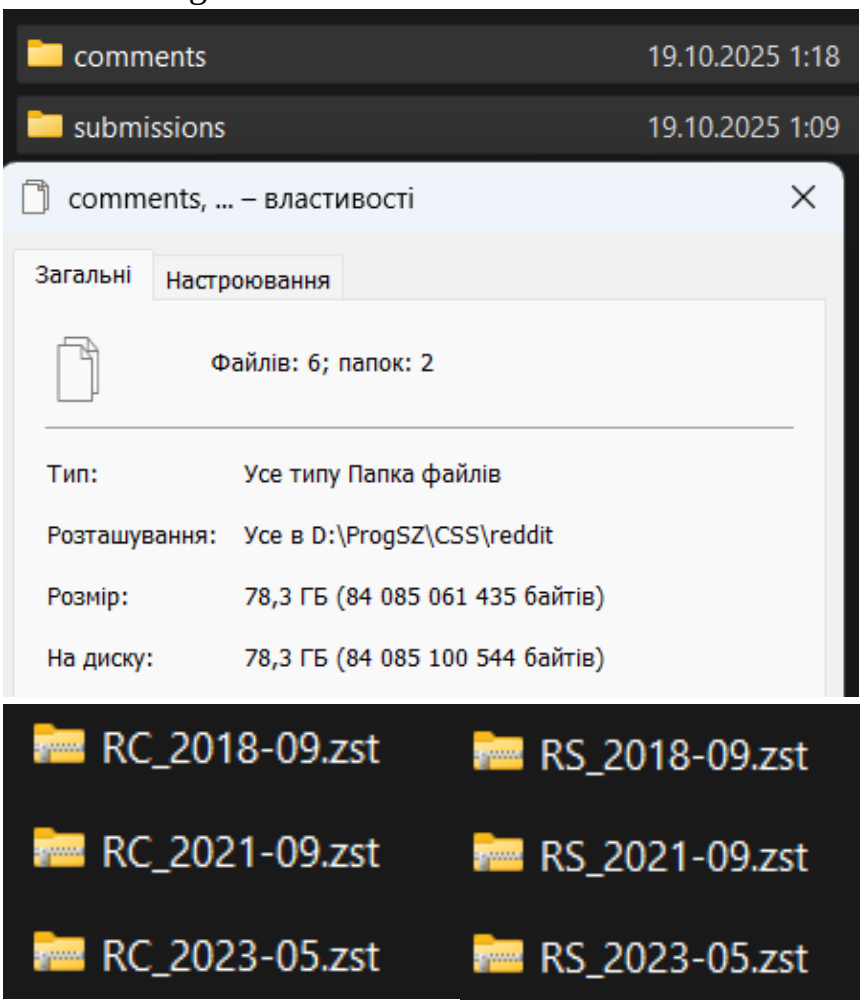
Source:

```
@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superseded= {}
}
```

Within the team, we split the months between each other. I downloaded and filtered 09.2018, 09.2021 and 05.2023.

Month	Topics
09.2018	"artificial intelligence", "machine learning", "deep learning", "neural network", "natural language processing", "nlp", "computer-based learning", "intelligent systems", "robot teacher", "edtech", " ai ", "ai ", " ai", " ml ", " ml", "ml "
09.2021	"artificial intelligence", "machine learning", "gpt-3", "openai", "generative model", "text generation", "ai writer", "essay writer", "online proctoring", "remote learning", "edtech", "academic integrity", "gpt", " ai ", "ai ", " ai", " ml ", " ml", "ml "
05.2023	"chatgpt", "chat gpt", "gpt-4", "gpt 4", "openai", "bard", "google bard", "gpt", "bing chat", "claude", "llm", "large language model", "artificial intelligence", "machine learning", "gpt-3", "generative model", "text generation", "ai writer", "essay writer", "generative ai", "generative artificial intelligence", "ai tool", "ai in education", "teaching with ai", "ai for homework", "ai tutor", "academic integrity", "honor code", "ai detection", "detecting ai", "ai policy", " ai ", "ai ", " ai", " ml ", " ml", "ml "

Downloading:



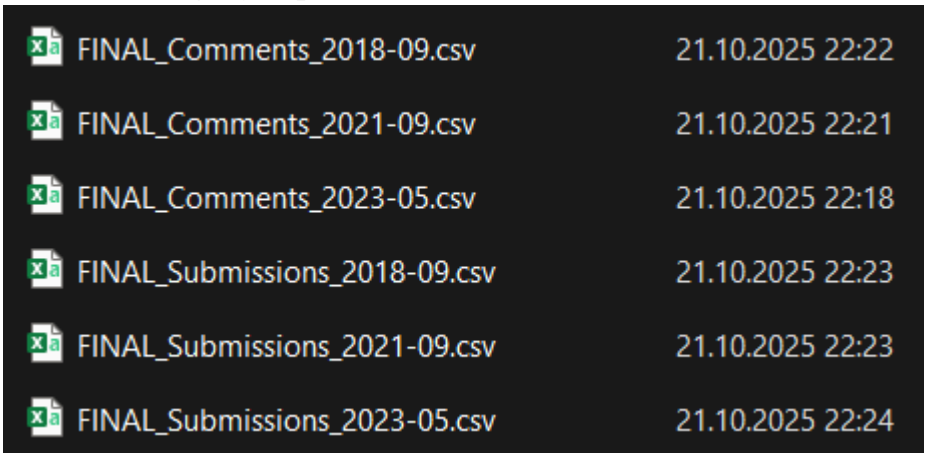
Cloning PushshiftDumps repository:

```
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
2025-10-20 15:57:57,617 - INFO: Filtering field: subreddit
2025-10-20 15:57:57,618 - INFO: On values: education,highereducation,univer
2025-10-20 15:57:57,618 - INFO: Exact match on. Single field None.
2025-10-20 15:57:57,618 - INFO: From date 2005-01-01 to date 2030-12-31
2025-10-20 15:57:57,618 - INFO: Output format set to zst
2025-10-20 15:57:57,618 - INFO: Processing 1 files
```

After filtering by topics:



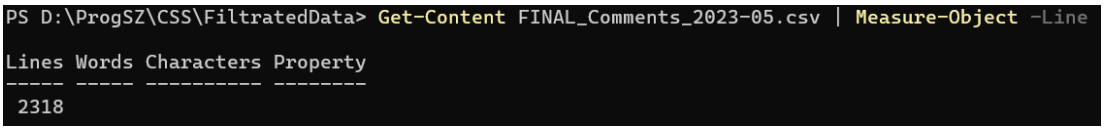
b. Screenshot of loaded DataFrame:

```
PS D:\ProgSZ\CSS\FiltratedData> python show_dataframe.py
Dataset successfully loaded. Here are the first 5 rows:
   2  ... Ap credits can only be taken until age 21 before college enrollment so this wouldn't apply here. \n\nThe CC r
oute is prob the best way to go because it may result in smaller class sizes and more one-on-one supported compared to a
 large university.
0  -6  ...                     Hmmm..... ChatGPT anyone? /s

1  14  ...  I saw “violations” too, and I was expecting to...

2   1  ...                      ChatGPT + Netus AI bypasser

3  13  ...  Try explaining this thoughtfully and calmly to...

4   3  ...  AI isn't that advance, you see the polished wo...

[5 rows x 5 columns]
```

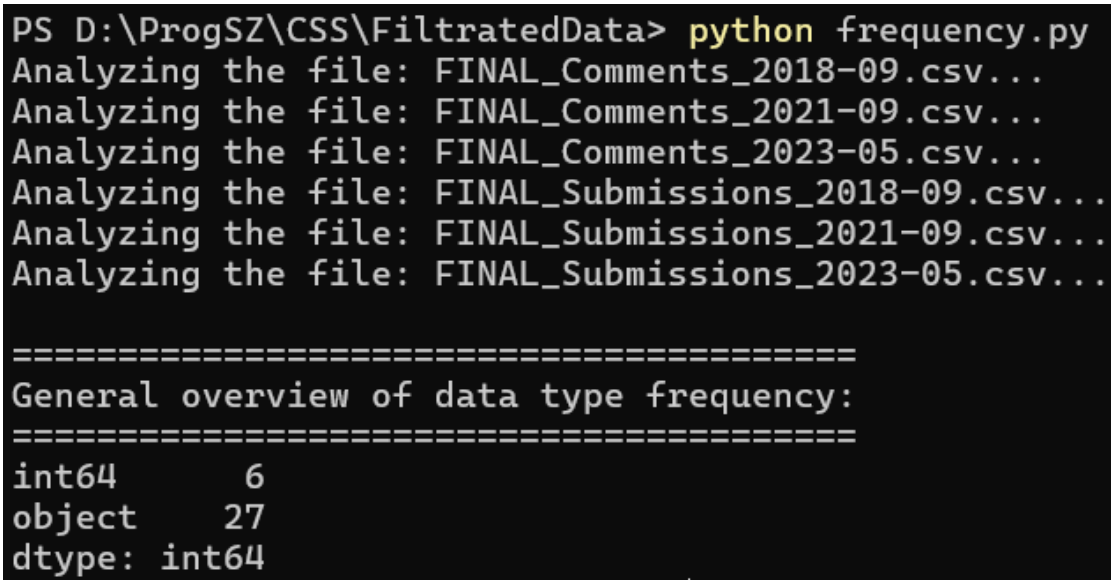
c. Link to our dataset: https://drive.google.com/drive/folders/1tLfxoc8LADqZDDmII0ZvutbVpkorcvh1?usp=sharing

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

Name	Size	Rows
FINAL_Comments_2018-09	358KB	1262
FINAL_Comments_2021-09	422KB	1478
FINAL_Comments_2023-05	651KB	2318
FINAL_Submissions_2018-09	94.4KB	343
FINAL_Submissions_2021-09	158KB	567
FINAL_Submissions_2023-05	265KB	899



e. Frequency overview of data types:



f. Brief interpretation of the dataset context:

The six filtered CSV files represent a specialized and highly filtered data corpus derived from the public Reddit archives. This is not raw data, but rather a targeted sample created to investigate the evolution of discussions surrounding Artificial Intelligence (AI) within educational contexts.

The data was sourced exclusively from topic-specific subreddits (r/education, r/highereducation, r/University, r/school, r/ Learning and r/ college). This allows us to assume that the authors of the comments and submissions are primarily students, educators and professionals within the field of education.

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

September 2018: During this period, discussions about AI in education were largely theoretical, academic, and niche. Our filtering focused on terms like machine learning and intelligent systems. We expected these files to contain a

relatively small amount of data, reflecting hypothetical discussions about the future, specific educational technologies (edtech), and conceptual ideas.

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 2023: By this time, ChatGPT had become a global phenomenon. These files reflect an explosion of practical and urgent discussions. The data contains conversations about specific tools (ChatGPT, Bard), the widespread use of AI by students for assignments, faculty concerns, institutional policy debates, and tools for AI detection. This is the largest and richest portion of my part of our dataset.

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