Pushshift Reddit Dataset

Introduction to Pushshift Reddit Dataset

What is the Pushshift Reddit Dataset?

- A large-scale social media dataset that includes all Reddit submissions and comments from 2005 to the present, updated in real-time.
- Enables social media researchers to reduce the time spent on data collection, cleaning, and storage.

Why Reddit?

Reddit is a significant platform for social media studies due to its millions of subreddits, billions of comments, and relatively open data access compared to platforms like Facebook and Twitter.

Goal of Pushshift

Make Reddit data more accessible and easy to analyze for researchers through APIs and monthly data dumps.

What is the Pushshift Reddit Dataset?

- A large-scale social media dataset that includes all Reddit submissions and comments from 2005 to the present, updated in real-time.

- Enables social media researchers to reduce the time spent on data collection, cleaning, and storage.

Why Reddit?

Reddit is a significant platform for social media studies due to its millions of subreddits, billions of comments, and relatively open data access compared to platforms like Facebook and Twitter.

Goal of Pushshift

Make Reddit data more accessible and easy to analyze for researchers through APIs and monthly data dumps.

Dataset Composition

Scale of the Dataset

- 651 million submissions and 5.6 billion comments from over 2.8 million subreddits between June 2005 and April 2019.
- The dataset includes metadata (e.g., post IDs, timestamps, authors) and content (e.g., comment text, submission titles).

Data Format

- Submissions and comments are stored in newline-delimited JSON (ndjson) format.
- Includes information such as score, author, subreddit, and number of comments.

Submission Json Response

Comment Json Response

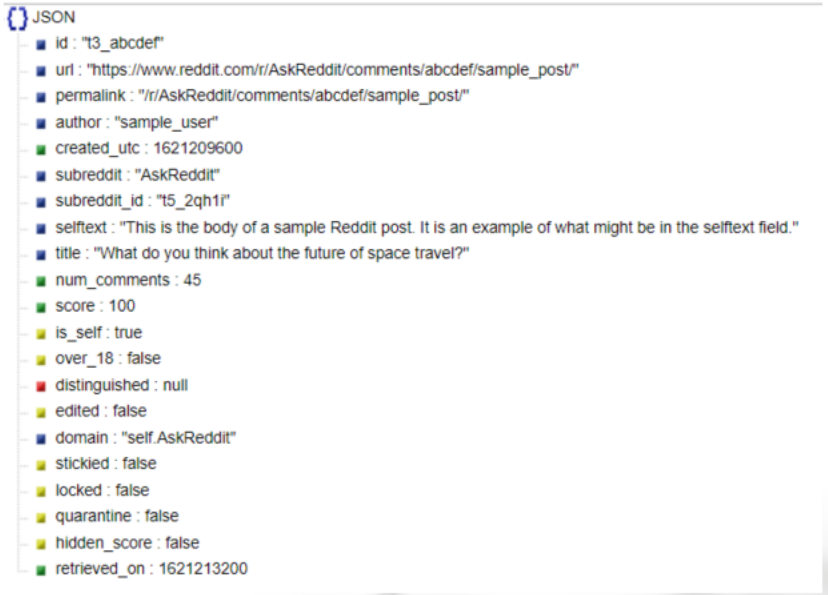
Scale of the Dataset

- 651 million submissions and 5.6 billion comments from over 2.8 million subreddits between June 2005 and April 2019.
- The dataset includes metadata (e.g., post IDs, timestamps, authors) and content (e.g., comment text, submission titles).

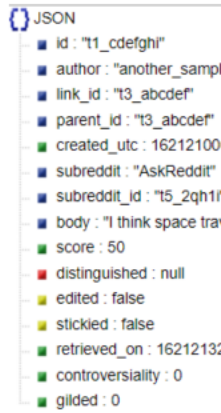
Data Format

- Submissions and comments are stored in newline-delimited JSON (ndjson) format.
- Includes information such as score, author, subreddit, and number of comments.

Submission Json Response



C

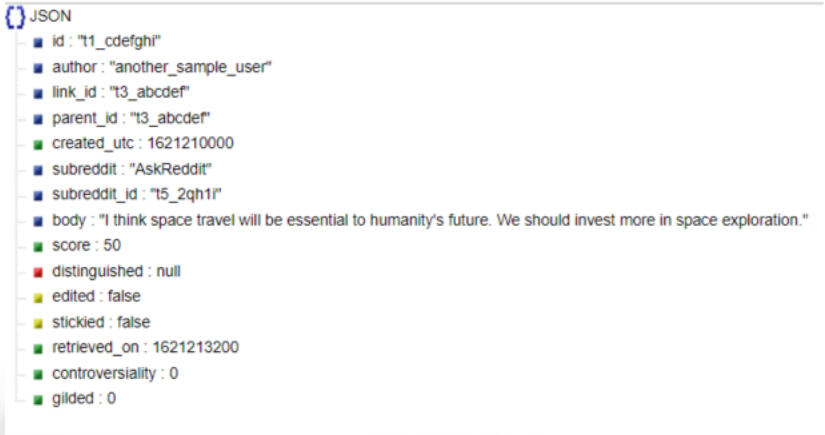


son

"

at might be in the selftext field."

Comment Json Response



Pushshift's Role

Navigating the Post-API Landscape

Challenges in the Post-API Age

- The restriction of data access by major platforms like Facebook and Twitter is known as the "post-API age.
- Due to privacy concerns, researchers face increasing difficulties in gathering social media data, particularly for studying issues like online misinformation.
- With the restriction of APIs, collecting data in a timely and efficiently has become increasingly difficult.

Pushshift's Role

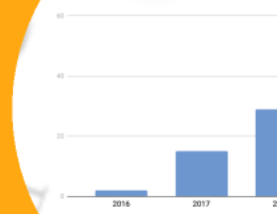
- In response to these challenges, Pushshift has emerged as a valuable tool for social media researchers.
- Unlike platforms with limited data access, Pushshift provides an extensive dataset of Reddit posts and comments, making both real-time and historical data available.
- This allows researchers to access large volumes of information that would otherwise be restricted.

Enhanced API Functionality

- Pushshift's API goes beyond what the official Reddit API offers by allowing full-text searches and supporting larger data queries.
- With its ability to handle five times more data per query and aggregate data efficiently, Pushshift simplifies the process of data collection and storage.
- It helps researchers to focus on actual analysis rather than technical tasks.

Data Format

Papers Published using PushShi



Over 100 peer-reviewed p
published using Pushshift

Challenges in the Post-API Age

- The restriction of data access by major platforms like Facebook and Twitter is known as the "post-API age.
- Due to privacy concerns, researchers face increasing difficulties in gathering social media data, particularly for studying issues like online misinformation.
- With the restriction of APIs, collecting data in a timely and efficiently has become increasingly difficult.

Pushshift's Role

- In response to these challenges, Pushshift has emerged as a valuable tool for social media researchers.
- Unlike platforms with limited data access, Pushshift provides an extensive dataset of Reddit posts and comments, making both real-time and historical data available.
- This allows researchers to access large volumes of information that would otherwise be restricted.

Enhanced API Functionality

- Pushshift's API goes beyond what the official Reddit API offers by allowing full-text searches and supporting larger data queries.
- With its ability to handle five times more data per query and aggregate data efficiently, Pushshift simplifies the process of data collection and storage.
- It helps researchers to focus on actual analysis rather than technical tasks.

Pushshift's Data collection process

Pushshift utilizes several backend software components for data collection, storage, cataloging, indexing, and dissemination

Ingest Engine

- Responsible for collecting raw data from Reddit, supporting multiple data sources such as APIs and web scraping.
- Uses a Redis queue for intermediate data storage before processing.
- Stores data in a custom format like newline-delimited JSON (ndjson)

PostgreSQL Database

Advanced querying of collected data and metadata, providing efficient and structured data retrieval

ElasticSearch

- Indexes and aggregates data, supporting full-text search and dynamic mapping to handle Reddit's evolving API.
- Ensures scalability with a cluster-based setup and offers Unicode, including emoji search.

Pushshift's Reddit data collection platform

Pushshift utilizes several backend software components for data collection, storage, cataloging, indexing, and dissemination

Ingest Engine

- Responsible for collecting raw data from Reddit, supporting multiple data sources such as APIs and web scraping.
- Uses a Redis queue for intermediate data storage before processing.
- Stores data in a custom format like newline-delimited JSON (ndjson)

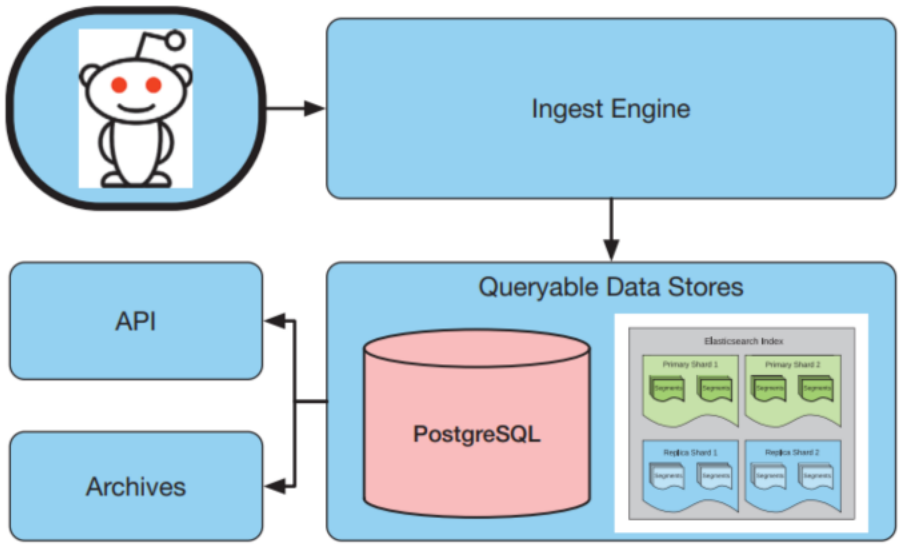
PostgreSQL Database

Advanced querying of collected data and metadata, providing efficient and structured data retrieval

ElasticSearch

- Indexes and aggregates data, supporting full-text search and dynamic mapping to handle Reddit's evolving API.
- Ensures scalability with a cluster-based setup and offers Unicode, including emoji search.

Pushshift's Reddit data collection platform



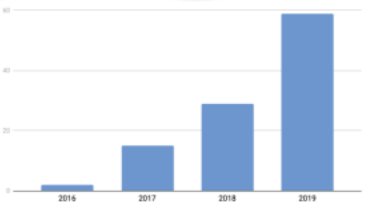
ape

Enhanced API Functionality

- Pushshift's API goes beyond what the official Reddit API offers by allowing full-text searches and supporting larger data queries.
- With its ability to handle five times more data per query and aggregate data efficiently, Pushshift simplifies the process of data collection and storage.
- It helps researchers to focus on actual analysis rather than technical tasks.

PushShift Reddit Dataset Impact

Papers Published using PushShift Data 2016-2019

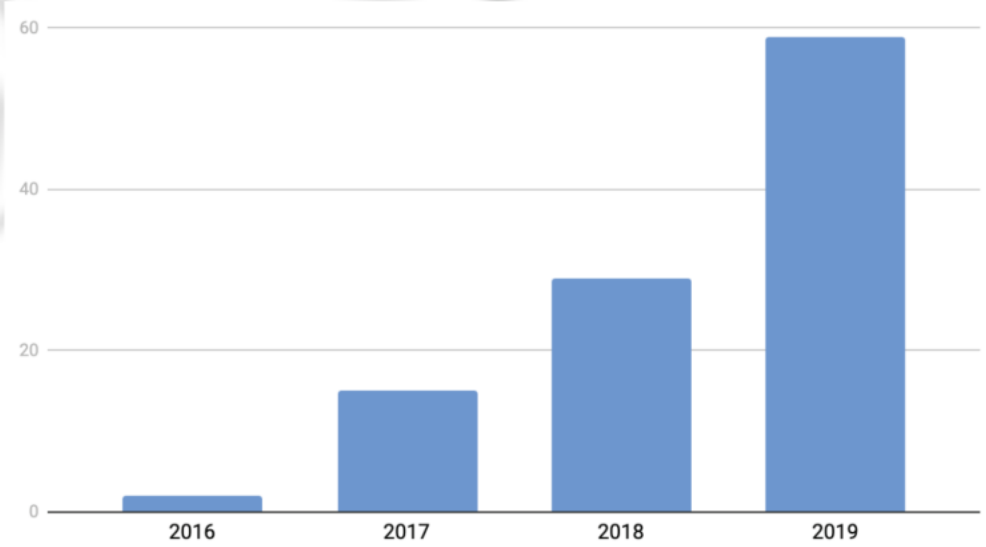


Over 100 peer-reviewed papers have been published using Pushshift data

Academic Influence

- Online Community Governance
- Extremism Research
- Disinformation Research
- Web Science Research
- Big Data and Health Informatics
- Robust Intelligence and NLP

Papers Published using PushShift Data 2016-2019



Over 100 peer-reviewed papers have been published using Pushshift data

2019

019

been

Academic Influence

- Online Community Governance
- Extremism Research
- Disinformation Research
- Web Science Research
- Big Data and Health Informatics
- Robust Intelligence and NLP

Related Work

Data Collection Services v/s PushShift

Data Collection
- Existing data collection services are available that cater to researchers' needs.
- These services move away from the traditional 'storage buckets' model.
- Focus on large-scale real-time social media data collection.

PushShift
- Pushshift is one of many real-time social media data collection services.
- It is designed to meet the needs of researchers and build upon existing models.
- Several other services have influenced Pushshift's goals and design.

Features Comparison

Feature	PushShift (now)	Media Cloud	GDELT	Stats Exchange	Wikimedia
Public API	Fully Supported	Fully Supported	Fully Supported	Fully Supported	Fully Supported
Data archive/dump	Fully Supported	Fully Supported	Fully Supported	Fully Supported	Fully Supported
Regularly updated	Fully Supported	Fully Supported	Fully Supported	Fully Supported	Fully Supported
Interactive computing	Fully Supported	Fully Supported	Fully Supported	Partially Supported	Fully Supported
Tutorials & demos	Fully Supported	Fully Supported	Partially Supported	Fully Supported	Fully Supported
Online community	Fully Supported	Not Supported	Partially Supported	Not Supported	Fully Supported
Outreach	Not Supported	Fully Supported	Partially Supported	Fully Supported	Fully Supported

Key Alternate Services

- **Media Cloud**: Tracks news stories with a semi-public API.
- **GDELT**: Global news monitoring with event tracking and analysis.
- **Stats Exchange**: Social Q&A data with SQL querying through Data Explorer.
- **Wikimedia**: Offers data archives, APIs, and Jupyter Notebooks for analysis.

Data Collection Services v/s PushShift

Data Collection

- Existing data collection services are available that cater to researchers' needs.
- These services move away from the traditional "storage buckets" model.
- Focus on large-scale real-time social media data collection.

PushShift

- Pushshift is one of many real-time social media data collection services.
- It is designed to meet the needs of researchers and build upon existing models.
- Several other services have influenced Pushshift's goals and design.

Key Altern

esearchers and build upon existing

ed Pushshift's goals and design.

Supported	Supported	Supp

Key Alternate Services

- **Media Cloud**: Tracks news stories with a semi-public API.
- **GDELT**: Global news monitoring with event tracking and analysis.
- **Stats Exchange**: Social Q&A data with SQL querying through Data Explorer.
- **Wikimedia**: Offers data archives, APIs, and Jupyter Notebooks for analysis.

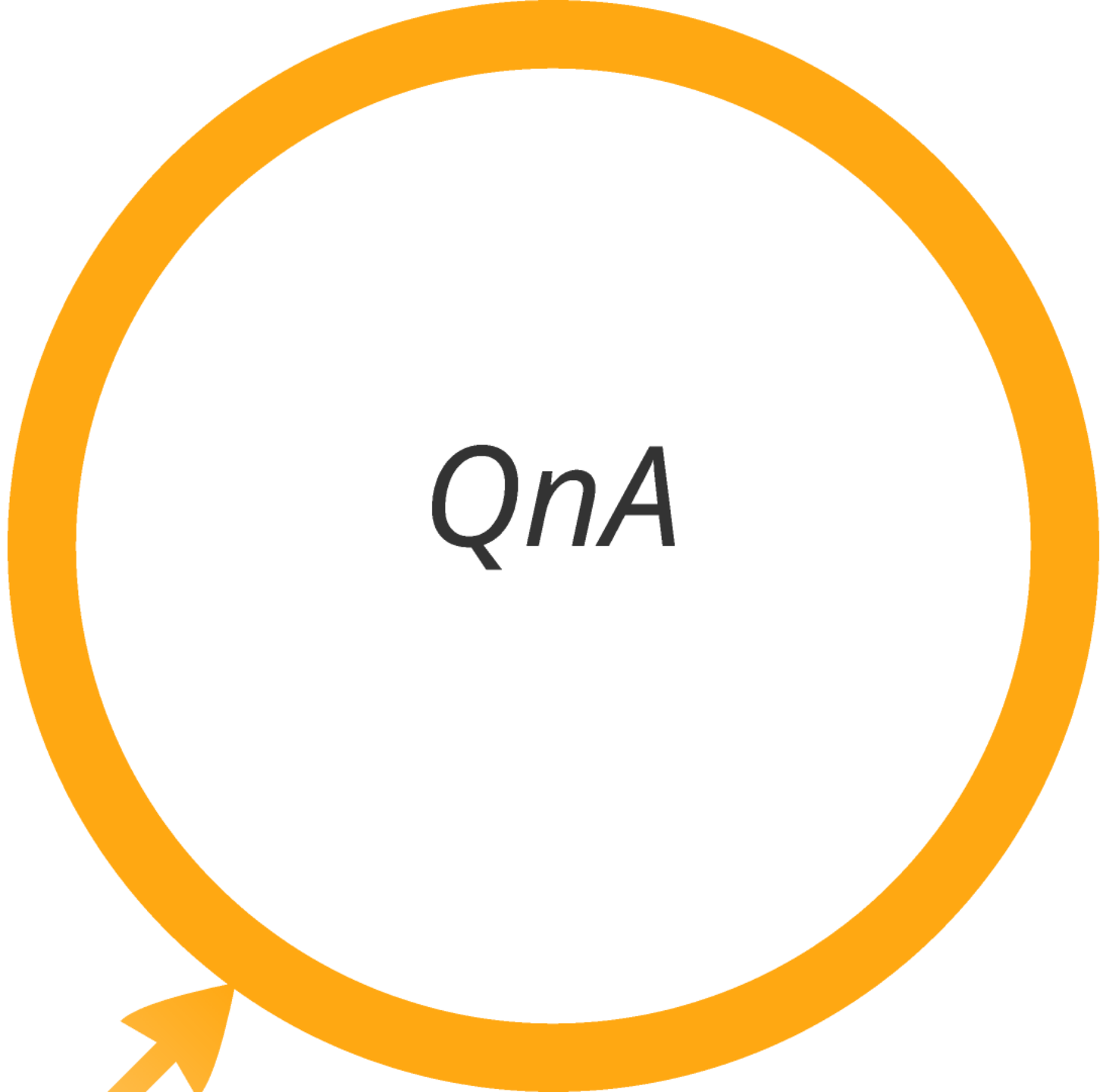
Features Comparison

Feature	PushShift (now)	Media Cloud	GDELT	Stats Exchange	Wikimedia
Public API	Fully Supported	Fully Supported	Fully Supported	Fully Supported	Fully Supported
Data archive/dump	Fully Supported	Fully Supported	Fully Supported	Fully Supported	Fully Supported
Regularly updated	Fully Supported	Fully Supported	Fully Supported	Fully Supported	Fully Supported
Interactive computing	Fully Supported	Fully Supported	Fully Supported	Partially Supported	Fully Supported
Tutorials & demos	Fully Supported	Fully Supported	Partially Supported	Fully Supported	Fully Supported
Online community	Fully Supported	Not Supported	Partially Supported	Not Supported	Fully Supported
Outreach	Not Supported	Fully Supported	Partially Supported	Fully Supported	Fully Supported

rnate Services

Conclusion

- Pushshift Reddit Dataset offers extensive Reddit data (651M submissions, 5.6B comments) from 2005 to now.
- Accessible via APIs and monthly dumps, it simplifies data collection and cleaning for researchers.
- Tools like APIs and Slackbot enable easy data interaction and visualization, boosting research efficiency.
- Used in over 100 papers, it's a crucial resource across fields like social science and disinformation research.
- Future Impact: Pushshift will continue to drive insights into online behavior and community dynamics.


QnA

Pushshift Reddit Dataset