StackExchange

Pruthvi Billa, Afia Simeen, Zizheng Zhang, Thanmai Reddy

Dataset Overview

Consists of user-contributed content on the Stack Exchange network.

AI, Gaming, History, Movies, Music, and Software engineering are the six chosen sites for our analysis.

Each site archive includes Posts, Users, Votes, Badges, Comments, PostHistory, and PostLinks.

Dataset Size: 3.05GB

Tools

aws

Amazon S3

Amazon EC2

hadoop HDFS

PySpark

CO

PyTorch

File Conversions

- Used the '**stackexchange-xml-converter**' CLI tool to convert from XML to CSV.

```
(base) pruthvishyambilla@Pruthvis-MBP-2 stackexchange-xml-converter % find /Users/pruthvish
ackExchange/ai.stackexchange.com -type f -name "*.xml" -exec ./stackexchange-xml-converter
-source-path {} -store-to-dir=/Users/pruthvishyambilla/Desktop/StackExchange_csv/ \;
2023/12/06 16:59:03 Total 1 file(s) to convert
2023/12/06 16:59:03 [Comments] Converting is started
2023/12/06 16:59:03 [Comments] File is converted. 27,350 of 27,350 row(s) has been processe
2023/12/06 16:59:03 Total 1 file(s) to convert
2023/12/06 16:59:03 [Users] Converting is started
2023/12/06 16:59:04 [Users] File is converted. 66,629 of 66,629 row(s) has been processed s
2023/12/06 16:59:04 Total 1 file(s) to convert
2023/12/06 16:59:04 [Votes] Converting is started
2023/12/06 16:59:04 [Votes] File is converted. 88,548 of 88,548 row(s) has been processed s
2023/12/06 16:59:04 Total 1 file(s) to convert
2023/12/06 16:59:04 [Tags] Converting is started
2023/12/06 16:59:04 [Tags] File is converted. 983 of 983 row(s) has been processed successf
2023/12/06 16:59:04 Total 1 file(s) to convert
2023/12/06 16:59:04 [PostHistory] Converting is started
2023/12/06 16:59:05 [PostHistory] File is converted. 100,535 of 100,535 row(s) has been pro
2023/12/06 16:59:05 Total 1 file(s) to convert
2023/12/06 16:59:05 [PostLinks] Converting is started
2023/12/06 16:59:05 [PostLinks] File is converted. 2,356 of 2,356 row(s) has been processed
2023/12/06 16:59:05 Total 1 file(s) to convert
2023/12/06 16:59:05 [Posts] Converting is started
2023/12/06 16:59:06 [Posts] File is converted. 25,296 of 25,296 row(s) has been processed s
2023/12/06 16:59:06 Total 1 file(s) to convert
2023/12/06 16:59:06 [Badges] Converting is started
2023/12/06 16:59:06 [Badges] File is converted. 58,820 of 58,820 row(s) has been processed
(base) pruthvishyambilla@Pruthvis-MBP-2 stackexchange-xml-converter %
```

Data Transfer

Objects (5) Info

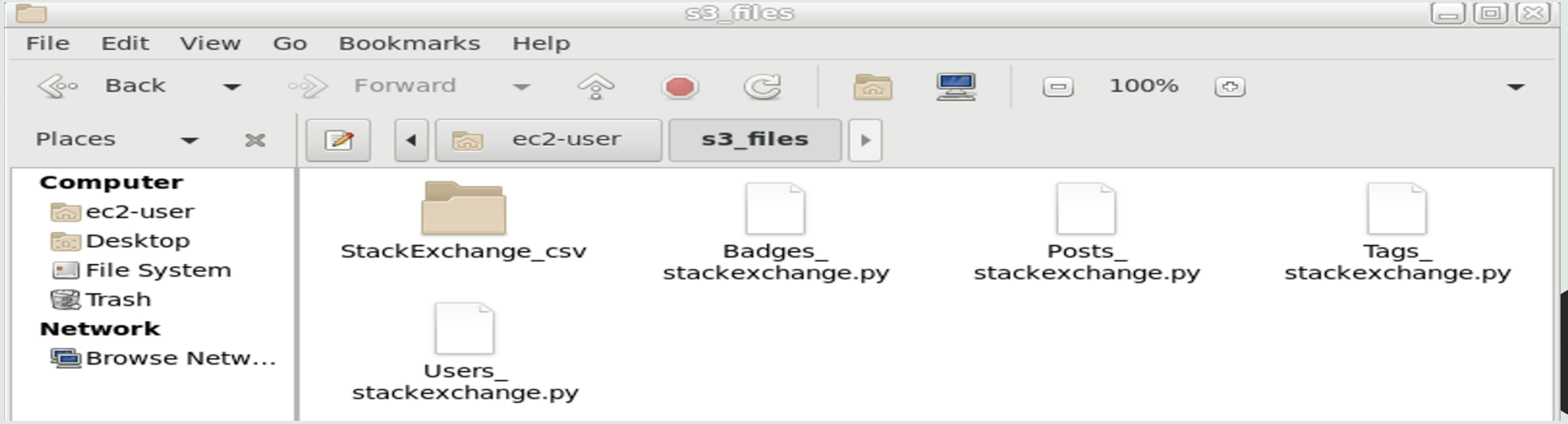
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

Copy S3 URI | Copy URL | Download | Open | Delete | Actions | Create folder | Upload

Find objects by prefix

Name	Type	Last modified	Size	Storage class
Badges_stackexchange.py	py	December 5, 2023, 01:03:03 (UTC-05:00)	5.4 KB	Standard
Posts_stackexchange.py	py	December 5, 2023, 01:03:04 (UTC-05:00)	1.2 KB	Standard
StackExchange_csv/	Folder	-	-	-
Tags_stackexchange.py	py	December 6, 2023, 17:57:18 (UTC-05:00)	2.1 KB	Standard
Users_stackexchange.py	py	December 5, 2023, 01:30:39 (UTC-05:00)	2.9 KB	Standard

```
[ec2-user@ip-172-31-61-143 ~]$ aws s3 sync s3://budt758-billa-330pm ./s3_files
```



Data Storage


```
[[root@quickstart s3_files]# ls
 StackExchange
[[root@quickstart s3_files]# hdfs dfs -put StackExchange/ /project/
 [root@quickstart s3_files]#
```

HUE | Query Editors | Data Browsers | Workflows | Search | Security

File Browser

Search for file name | Actions | Move to trash

Home / project / **StackExchange** | History | Trash

Name	Size	User	Group	Permissions	Date
↑		root	supergroup	drwxr-xr-x	December 04, 2023 01:23 PM
.		root	supergroup	drwxr-xr-x	December 04, 2023 01:23 PM
.DS_Store	38.0 KB	root	supergroup	-rw-r--r--	December 04, 2023 01:23 PM
ai.meta.stackexchange.com		root	supergroup	drwxr-xr-x	December 04, 2023 01:23 PM
ai.stackexchange.com		root	supergroup	drwxr-xr-x	December 04, 2023 01:23 PM
gaming.meta.stackexchange.com		root	supergroup	drwxr-xr-x	December 04, 2023 01:23 PM
gaming.stackexchange.com		root	supergroup	drwxr-xr-x	December 04, 2023 01:23 PM
history.meta.stackexchange.com		root	supergroup	drwxr-xr-x	December 04, 2023 01:23 PM
history.stackexchange.com		root	supergroup	drwxr-xr-x	December 04, 2023 01:23 PM
movies.meta.stackexchange.com		root	supergroup	drwxr-xr-x	December 04, 2023 01:23 PM
movies.stackexchange.com		root	supergroup	drwxr-xr-x	December 04, 2023 01:23 PM
music.meta.stackexchange.com		root	supergroup	drwxr-xr-x	December 04, 2023 01:23 PM
music.stackexchange.com		root	supergroup	drwxr-xr-x	December 04, 2023 01:23 PM
softwareengineering.meta.stackexchange.com		root	supergroup	drwxr-xr-x	December 04, 2023 01:23 PM
softwareengineering.stackexchange.com		root	supergroup	drwxr-xr-x	December 04, 2023 01:24 PM

Descriptive Analytics

Platform Growth: How does badge distribution bolster Stack Exchange's user participation and expertise recognition?

Cross-Community Engagement: How can cross-topic engagement data improve community-building across Stack Exchange?

Content & Support Strategy: How does understanding user post volume on Stack Exchange topics enhance the platform's content and support?

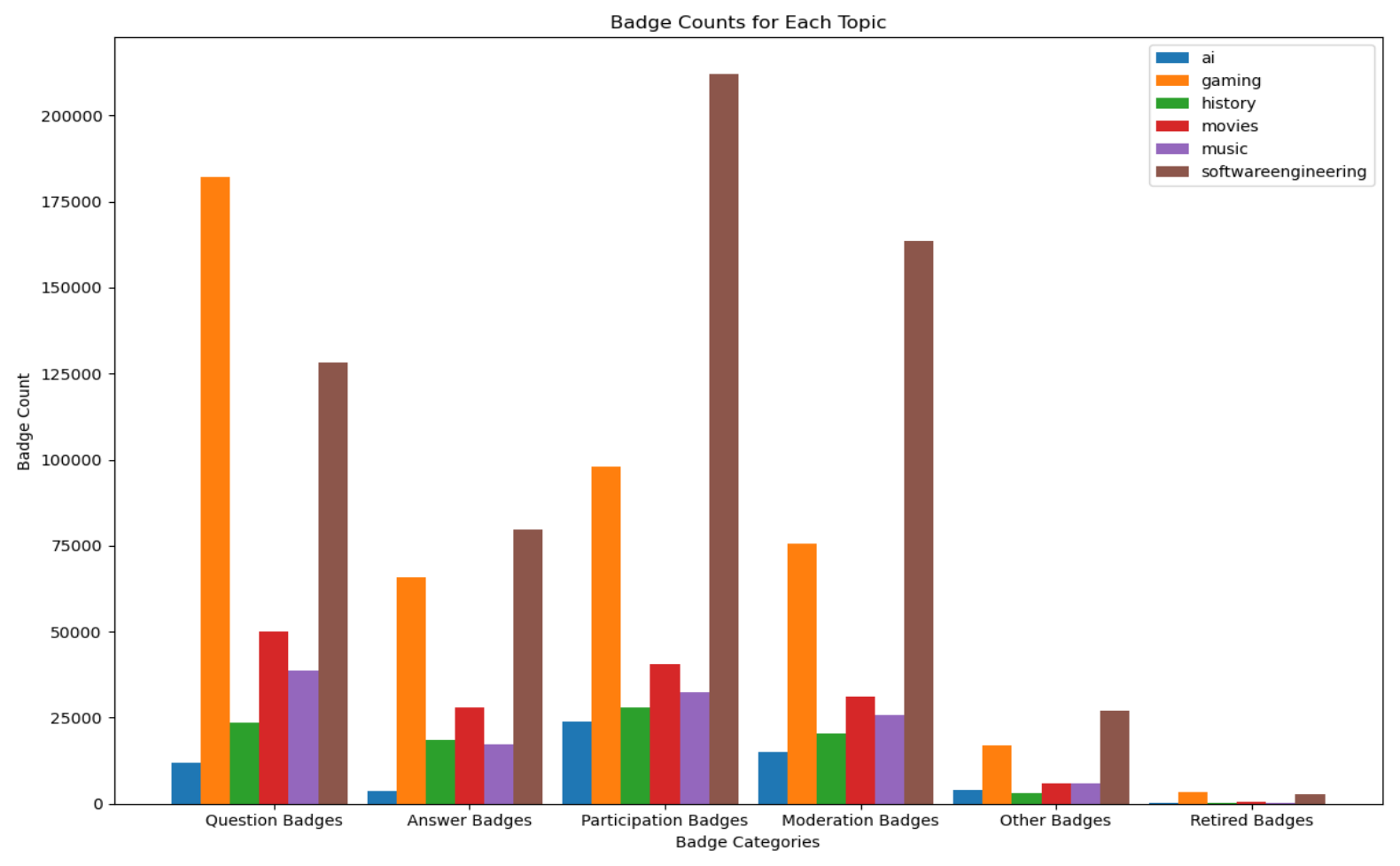
Tag Analysis for UX: How can tag frequency analysis refine user experience on Stack Exchange?

Business Case 1 – Platform Growth

Username root Text Search for text Succeeded Running Failed Killed

Logs	ID	Name	Application Type	Status	User	Maps	Reduces	Queue	Priority	Duration	Submitted
	1701749287690_0018	PigLatin:common_badges_etl.pig	MAPREDUCE	SUCCEEDED	root	100%	100%	root.root	N/A	32s	12/04/23 21:00:03
	1701749287690_0016	PigLatin:badges_etl.pig	MAPREDUCE	SUCCEEDED	root	100%	100%	root.root	N/A	1m:33s	12/04/23 20:39:07

- Provide incentives such as enhanced visibility and recognition within the community or even physical rewards for the best answers.
- Implement collaborative events, peer sessions, and dedicated platforms to replicate high participation in software engineering across other topics.

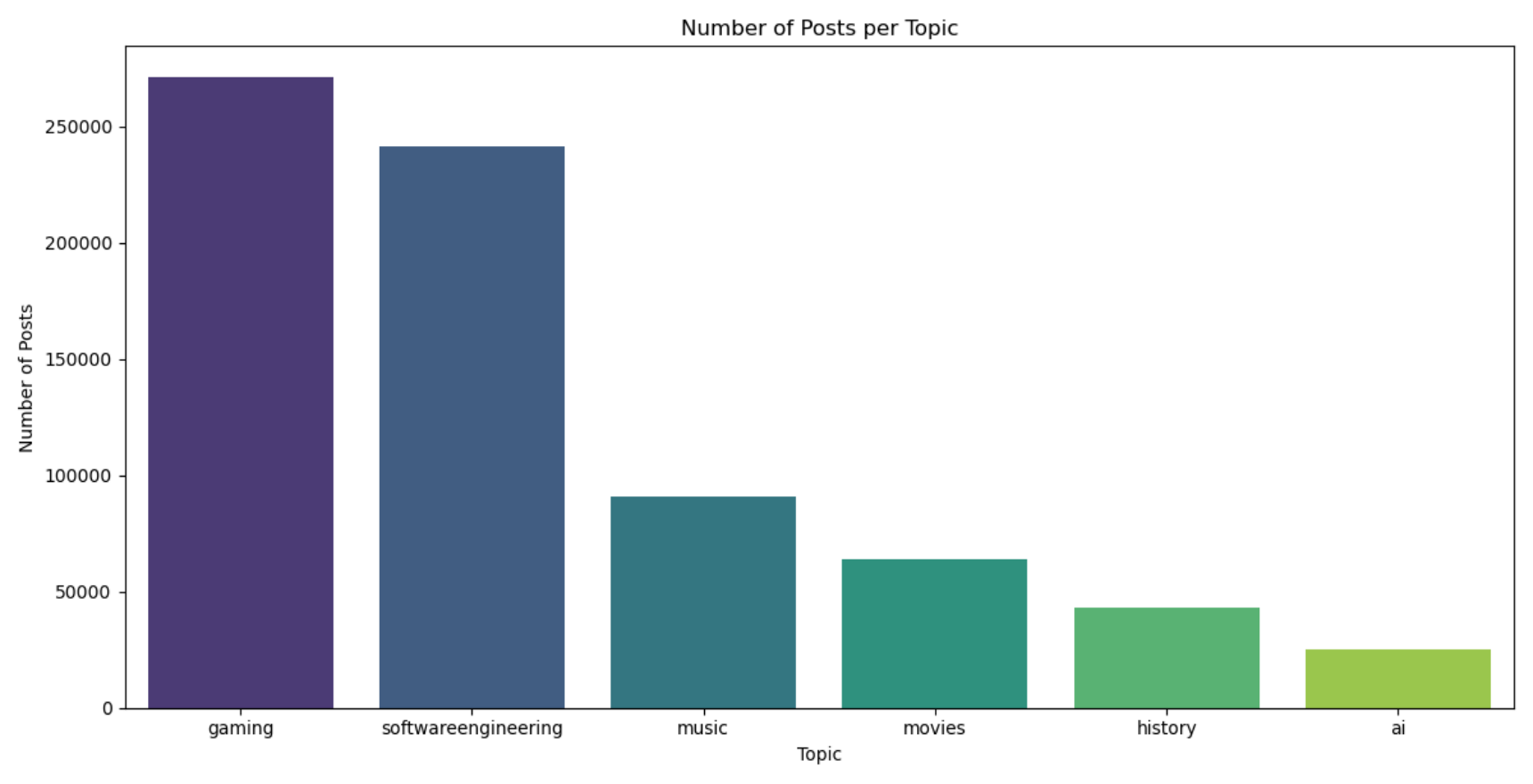


Business Case 2 – Cross Community Engagement

```
[[ec2-user@ip-172-31-61-143 ~]$ python3 Users_stackexchange.py
                        ai  gaming history  movies   music softwareengineering
ai                    100%   8.81%   6.15%   7.10%   6.87%              26.69%
gaming               2.83%    100%   4.68%   7.59%   5.21%              20.02%
history              9.34%  22.11%    100%  21.24%  14.96%              32.06%
movies               6.40%  21.26%  12.60%    100%  12.05%              30.27%
music                6.74%  15.91%   9.66%  13.12%    100%              28.73%
softwareengineering  4.82%  11.25%   3.81%   6.07%   5.29%               100%
Total number of users in ai: 66622
Total number of users in gaming: 207116
Total number of users in history: 43831
Total number of users in movies: 73902
Total number of users in music: 67872
Total number of users in softwareengineering: 368617
There are 1071 number of users who are active in all the 6 topics.
There are a total of 822605 users in all the 6 topics.
```

- Launch initiatives like cross-topic events, challenges, or discussions that appeal to users with interests in multiple areas.
- Provide collaborative spaces to connect users with overlapping interests.

Business Case 3 – Content & Support Strategy



- Strategically allocate resources such as time, staff, and money to enhance content and support in areas with the highest engagement, like gaming and software engineering, ensuring that user needs are effectively met.

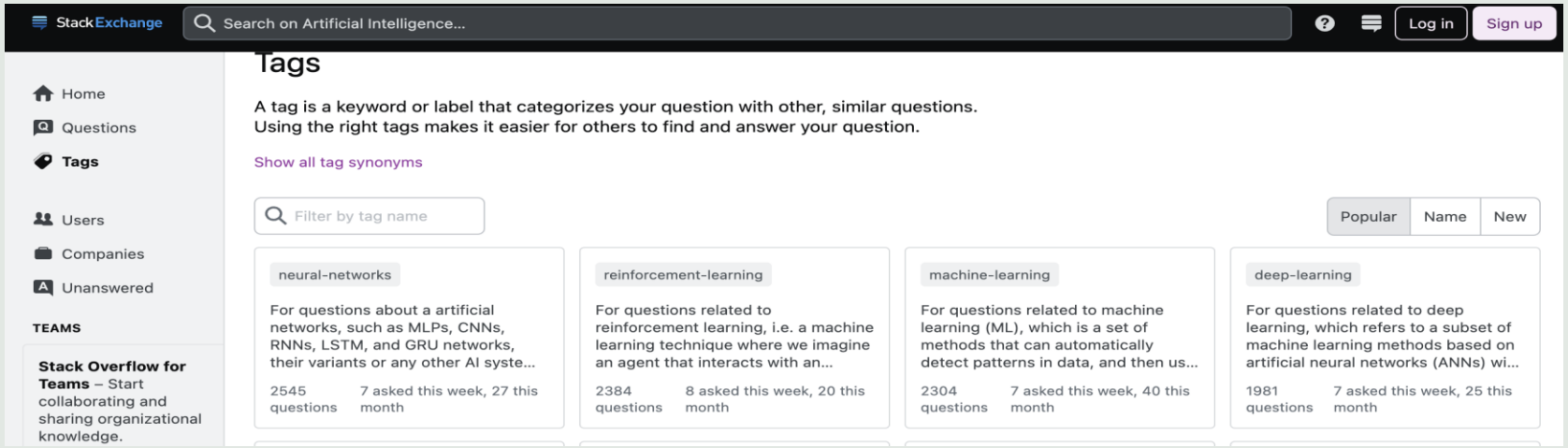
Business Case 4 – Tag Analysis for UX

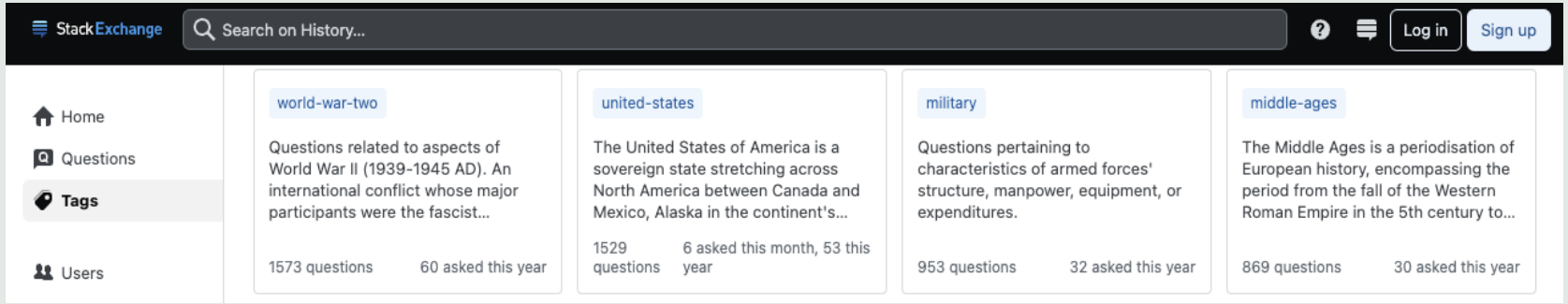
```
Top 5 most frequent words for: movies
plot-explanation: 1.0
character: 0.28635536688902363
analysis: 0.16810187992722864
marvel-cinematic-universe: 0.16203759854457248
dialogue: 0.11400848999393572


Top 5 most frequent words for: ai
neural-networks: 1.0
reinforcement-learning: 0.9353932584269663
machine-learning: 0.8888443017656501
deep-learning: 0.7724719101123596
convolutional-neural-networks: 0.4550561797752809


Top 5 most frequent words for: music
theory: 1.0
guitar: 0.905587668593449
piano: 0.8420038535645472
notation: 0.6229011835948252
chords: 0.5733553537021745


Top 5 most frequent words for: gaming
minecraft-java-edition: 1.0
minecraft-commands: 0.3793723316605498
the-elder-scrolls-v-skyrim: 0.3281402142161636
steam: 0.20373005767358252
diablo-iii: 0.1987117069882406


Top 5 most frequent words for: softwareengineering
design: 1.0
c#: 0.9582271033535987
java: 0.9580309864679349
design-patterns: 0.8603647774073347
architecture: 0.6760149048833105


Top 5 most frequent words for: history
world-war-two: 1.0
united-states: 0.974293059125964
military: 0.609254498714653
middle-ages: 0.5546272493573264
ancient-history: 0.4723650385604113
```

```
topics_list = ['ai', 'gaming', 'history', 'movies', 'music', 'softwareengineering']
```

```
# Parallelize the task using Spark and collect results locally
spark.sparkContext.parallelize(topics_list).foreach(generate_wordcloud)
```

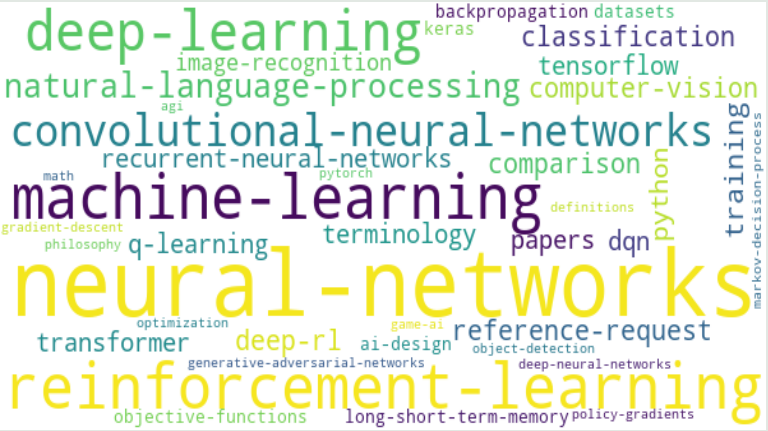
PySpark





Word Clouds

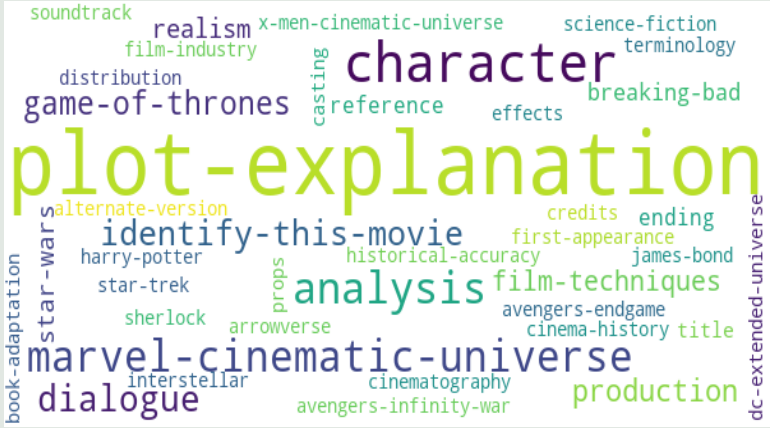
AI



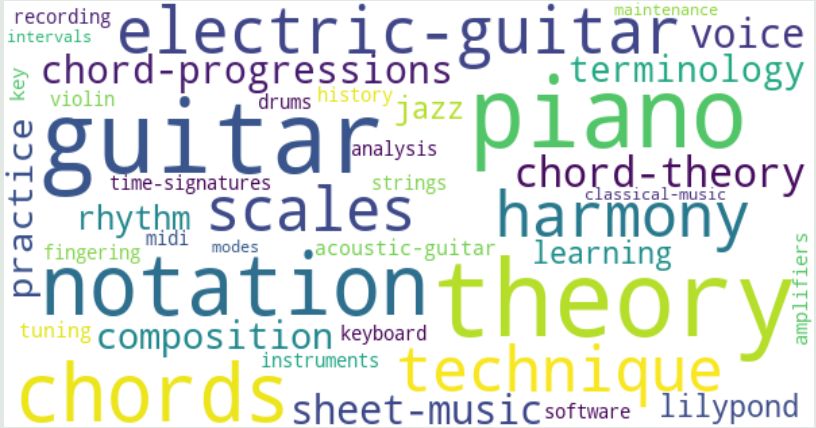
Gaming

History

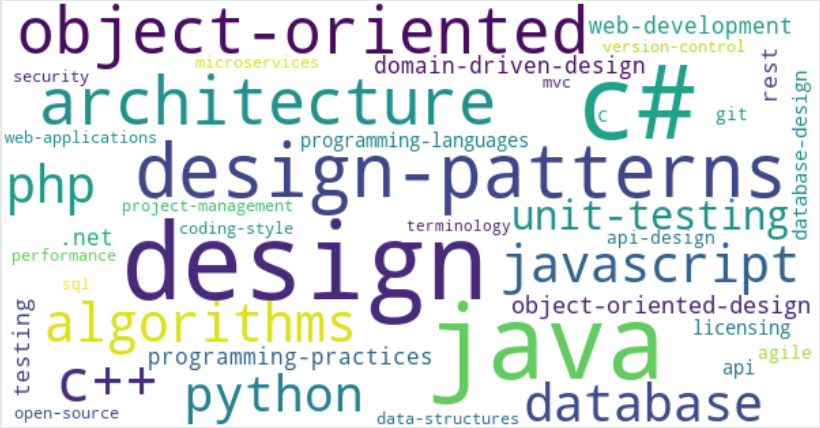
Movies



Music



Software Engineering



- Optimize search and recommendations to surface discussions on popular tags for better information discovery.
- Visually emphasizing frequent tags with larger fonts or distinct colors, and featuring them in a 'trending' section, streamlines navigation and facilitates a smoother, more intuitive user experience.

Predictive Analytics - Modelling

Topic Classification: Given a user's post, predict the topic to which this post belong.
Problem type: Text classification

Reasoning:

- All posts under a given topic are relevant.
- The relevance could be captured through a statistical approach, i.e., it is statistically learnable.
- Anomalies, such as toxic as well as irrelevant posts do not share that relevance.
- Anomalies could be identified through the use of a machine learning model.

Business Value:

- Improve the contents quality on StackExchange.
- Automate part of the censorship workflow and therefore improve productivity.
- Utilize the state of the art AI technology.

Predictive Analytics - Implement

Model Selection: BERT Model for Text Classification

- A large language model that captures contextual information bidirectionally
- Adept at multiclass classification tasks.
- Can be fine-tuned to adapt to the nuances of the target classification problem

Training Details:

- Hardware: Google Colab T4 GPU (16GB VRAM)
- Software: PyTorch, HuggingFace, and other common Python libraries
- Dataset: User posts from StackExchange
- Loss function: cross entropy
- Epoch: 2
- Batch size: 16 (owing to limited VRAM)

Predictive Analytics - Performance

Decreasing Loss Trends:

- Decreasing training loss (0.0487 to 0.0235) indicates effective learning from the training data.
- Decreasing validation loss (0.0446 to 0.0397) reflects the model's generalization capability.

Epoch	Training Loss	Validation Loss
1	0.048700	0.044623
2	0.023500	0.039681

Optimal Model Performance:

- Converging training and validation losses at low values suggests the model is reaching an optimal state.
- The diminishing gap between the two losses also implies good generalization to new, unseen data.

```
{'eval_loss': 0.03968135267496109,
 'eval_runtime': 229.8726,
 'eval_samples_per_second': 63.614,
 'eval_steps_per_second': 3.976,
 'epoch': 2.0}
```

Evaluation Accuracy:

- **96.9%** on sample evaluation dataset.

Predictive Analytics - Examples

A random post from a user (positive case):

I have created a short melody that uses these notes:What mode contains these notes?I have tried Dorian, Aeolian, Lydian, Phrygian and Mixolydian by starting at the scale in each mode that contains all naturals and working up via 5ths. But I can't find a scale that incorporates these notese.f. for Dorian, I started on D, then A, then E, etc. etc.D E F G A B CA B C D E F# GE F# G A B C# DB C# D E F# G# AF# G# A B C# D# EC# D# E F# C# A# B

Model prediction: music
Actual label: music

Predictive Analytics - Examples

A random post from a user (positive case):

I am disappointed by the lack of imagination displayed by my fellow programmers here.It seems to me the client did some research. He may have read somewhere that quality code typically contains about 25% of comments. Obviously he cares/worries about maintenance further down the road. Now, how does he make that concrete in a requirements document that is to be tied to a contract? That is not easy. It may even be impossible…(truncated)

Model prediction: software engineering

Actual label: software engineering

Predictive Analytics - Examples

A random post from a user (negative case):

According to this unofficial wiki:Jeanne was captured by English troops, accused of witchery, and burned at the stake on May 31 1431, at the age of 19. However, the Templar Order had orchestrated her execution in order to steal her Sword.An ancestor of Warren Vidic was present at the trial and execution.Since the Templars apparently orchestrated her execution, it is very likely they already knew about the Sword beforehand.

Model prediction: gaming

Actual label: history

Predictive Analytics - Examples

A random post from a user (negative case):

Why did algorithm S do better at beating humans than algorithm H? Because S was a better model of human behaviour. The obvious difference is that S (Shannon) had a short memory and H (Hagelbarger) had one that was longer. We can hypothesise that humans play this game with more short term than long term consistency. Obviously to check our hypothesis will require more experiments.Why did algorithm S beat algorithm H? … (truncated)

Model prediction: software engineering

Actual label: AI

Predictive Analytics - Deliverables

Tuned model available at HuggingFace Hub

- **Repo Id: Chaconne/BDAI**
- **Repo link:** https://huggingface.co/Chaconne/BDAI
- **Demo notebook:** https://colab.research.google.com/drive/1iGJXVLkDsLqhZrPYkltGbMovpT1xPNht#scrollTo=9sdV93cyQHhK

Predictive Analytics - Limitation

- Original schedule was to train the model on A100 GPU with 40GB VRAM.
- Owing to unavailability to high-performance computation resource, only a small portion of the original dataset (58,881 rows) were used in training to accommodate the mere 16 GB GPU memory.
- Based on the loss information, the model has not yet been tuned to its optimal state.
- Time consuming data manipulation and training process make it very inefficient to debug.
- Model is prune to mistakes when trying to predict closely related topics.

Conclusion

Successfully fine tuned a large language model that classifies Stack Exchange posts, improving the contents quality and reducing human labor cost.

Conducted a comprehensive user behavior analysis to strategically allocate resources.

Revealed correlations between user activity and topic preferences, providing a basis for focused cross community engagement.

These insights offer Stack Exchange data-driven strategies to optimize user experience and drive platform growth.

Thank You