MINECRAFT DATA ENGINEERING

SPONSORED BY: MICROSOFT MINECRAFT

Our Team:

We are The Data Mine's 2020-2021's Corporate Partnership cohort for Microsoft Minecraft.

Here's a little bit about us:
- We all come from very different backgrounds.
- We are united by our love for Minecraft and data science.
- With our mentors' help, we've analyzed data from across multiple social media platforms!
- Doing this was no easy task, so we needed to divide into three groups to achieve our goals.
- From artists to data scientists, this project used every one of our skills.

Our Problem:

When it comes to creating an excellent user experience, gathering feedback and the general sentiment from the masses is essential.

Here's an overview of what our goals were and why:
- With so many different social media websites and outlets, gathering data was challenging.
- To solve this problem, we narrowed down what websites to look at.
- We developed scrapers to gather data from social media platforms.
- Our aim was to gather feedback and understand general user sentiment.
-This data is invaluable for our stakeholders. (Game developers, marketers, etc.)

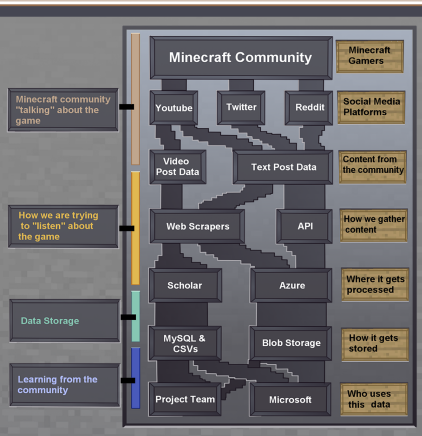
Our Names:

TAs: Kelly Addison, Laura Humphrey
Students: Joseph Bushagour, Kevin Choe, Supriya Dixit, Sumeeth Guda, Eric Han, Ramitha Kotarkonda, Michael Kruse, Kiersten Lofton, Sean McGuckin, Anton Nagy, Jennifer O'Connell, Dhanya Prem Sankar, Sid Rao, Vaishakh Vinod Kumar

About Minecraft:

Minecraft has many different versions, which poses a problem for us when determining what version people are talking about. However, each version has some essential differences that help us determine what people are talking about.
- Java is a popular version of Minecraft that can be modded
- Bedrock has the Marketplace and supports cross-play
- Pocket edition is the mobile game equivalent of Minecraft.

Also note:
- Minecraft is a sandbox genre video game where players have control over a 3D-generated world.
- It is one of the best-selling video games of all time, with 200 million copies sold and 126 million active users.
- Spin-offs include Minecraft Dungeons, Story Mode, and Earth.
- Its massive following has enabled a solid social media presence among its users.



APIs and Web Scrapers:

For each social media platform, we can collect data with an API or a web scraper.

What is an API?
- An API, Application Programming Interface, is software that helps us make programs that can collect data and interact with its respective social media platform.
- We used these to gather Twitter data.

Some challenges associated with APIs are:
- They may not be customizable.
-They may provide limited access to data.
- They can be challenging to learn.

What is a web scraper?
- A web scraper is a program that can collect data from web pages automatically.

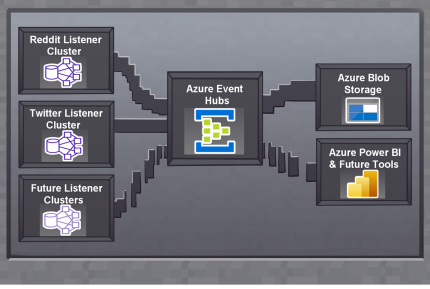
Details on web scrapers include:
- One must build their own program to describe how to collect data
- They are a more flexible data collection process
- They are how we collected data from Reddit and are working towards making one for YouTube

Moving to Azure:

The architecture of our social media listening system on Microsoft Azure differs from the previous architecture of the listening system on Purdue's Scholar computing cluster.

About Scholar:
- On Scholar, each listener is independent.
-It creates a fractured data ecosystem.

About Azure:
-The Azure architecture uses Azure's Event Hubs to consolidate listeners.
-Each listener exists on its own cluster.
- Listeners send messages in a centralized Event Hubs namespace.
-Multiple services can subscribe to this Event Hubs namespace.
-In the future, tools such as Power BI can be used for real-time dashboarding.
-Event Hubs have an extensible and flexible design.



The Patch Notes:

Since we moved to Azure a lot of things have changed, mostly for the better!

Moving to Azure has helped bring a lot of benefits to us, including:
- Increased uptime to 99.95% by using Databricks virtual machine Job Clusters (Standard_DS3_v2 VMs).
- Added redundancy in case one virtual machine fails
- Centralized event scraping with Azure Event Hubs acting as a data highway for all Reddit posts, tweets, etc.
- Built an extensible messaging system that receivers can subscribe to dashboard, push events into a database, etc.
- Transfer of data to Microsoft employees simplified through Azure Blob Storage.
- Removed Herobrine.

MINECRAFT DATA SCIENCE

SPONSORED BY: MICROSOFT MINECRAFT

Reddit

Reddit is a forum that has subreddits which are forums specific to certain things.

r/Minecraft is a subreddit that has 5 million members.

Reddit sorts posts through tags and is more moderated than Twitter.

Members of r/Minecraft are hardcore users who enjoy modding and often have prior knowledge of programming.

Posts made to the subreddit are scraped once within 30 seconds of creation and again a day later.

Twitter

Twitter is a social media platform where people post "tweets" of up to 280 characters.

People tweeting about Minecraft tend to represent more casual gamers, or are people calling attention to Twitch streams.

Any tweet with"Minecraft" in the tweet or hashtags is streamed in real-time.

Twitter tends to involve popularity and followers more than Reddit due to the way both are set up.

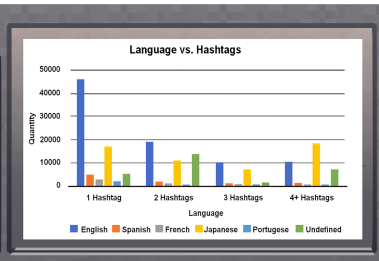
Future Plans

We hope the Minecraft Data Mine team during the 2021- 2022 academic year can utilize our tools to develop new insights. Currently, our plans for next year are to:
-Continue building relationships with the Microsoft Minecraft team.
-Increase collaboration.
-Understand different cultural behaviors of players to understand trends and perspectives.
(Ex. A player from the US vs. China)

Our Scope:

When it comes to sampling sizes, it's only natural to want a large pool. However, we can't monitor everything, so we had to make some choices to narrow down the scope and scale of our project:
-Currently, we are focusing mainly on the West and on data that is in English.
-Language barriers have caused us some issues.
-For now, we are sticking to English BUT we hope to broaden our datasets.
-Some team members have experience with different languages.
-We have looked into Chinese in specific due to Minecraft China edition.

Some important notes about languages:
-Our Twitter scraper can differentiate languages from each other.
-The Minecraft subreddit is moderated to be English only.

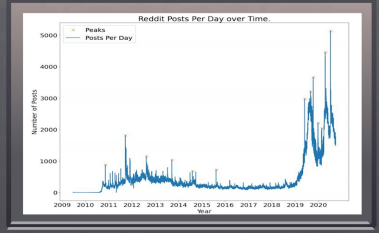


The figure above shows the relationship between the number of hashtags and the language of them:
-English dominates the top for most of the batches.
-Japanese tweets have a higher amount for 4+ hashtags.
This could be the result of Japanese Twitter users utilizing more hashtags than English and other language speakers.

Next To Address: YouTube

We mainly focused on Twitter and Reddit data because we deemed those two social media websites as essential for Minecraft data. However, one primary audience we have been neglecting is younger players. Youtube fills the gap that we have when it comes to our data. Therefore, we've been working on a Youtube scraper as well. Here are some critical points about Youtube:
-YouTube is a platform where users can post and comment on videos.
-These videos vary in length and content.
-About 2 billion logged-in users visit YouTube each month.
-Minecraft-related content in YouTube can range from short individual videos to an entire series of videos.

Any Minecraft-related video is collected. General data regarding the video (e.g., Title, date uploaded, hashtags) is collected along with the video's transcription and the video's comment section. The tools used to gather this data are still in development.



The above graph shows the number of Reddit posts between 2009 and 2020~. We see several spikes in posts, most of which we found to equivalate to game updates and the buzz around them.

A SPECIAL THANKS TO...

OUR MENTORS FROM MICROSOFT:
FRANCISCO RIUS AND BHRIGU SHREE

In Conclusion:

In the past two semesters, we discovered the seasonality of players' activities and discovered the similarity in trends between in-game behaviors and social media feed. We successfully transitioned from python-based listener to Azure, which built strong foundation for future exploration in the coming semesters.

References:

Brownlee, Jason. "17 Statistical Hypothesis Tests in Python (Cheat Sheet)." Machine Learning Mastery, 27 Nov. 2019, machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/.

Lenadroid. "Tutorial: Stream Data into Azure Databricks Using Event Hubs." Microsoft Docs, docs.microsoft.com/en-us/azure/databricks/scenarios/databricks-stream-from-eventhubs.

"Selenium with Python." Selenium with Python - Selenium Python Bindings 2 Documentation, selenium-python.readthedocs.io/.

"Tweepy Documentation¶." Tweepy Documentation - Tweepy 3.10.0 Documentation, docs.tweepy.org/en/latest/.