MINECRAFT: DIGGING FOR INSIGHTS

DATA SCIENCE

During the first semester, we chose to focus on Twitter because there was already an existing Twitter scraper from last year. This meant that we could build on previous work and expand on the previous pipeline.

During the second semester, we chose to focus on YouTube because it is one of the most popular platforms for Minecraft. The previous scraper collected data from specific channels, and we wanted to create something more general to Minecraft as a whole for the platform.

DATA PREPROCESSING

Data preprocessing work was done to clean up and make the data easier to interpret and work with the ML models.

The first preprocessing pass was done to translate emoji to text that could be using in NLP.

The second preprocessing pass was done to identify and remove stop words from the data. The purpose was to further optimize the data for interpretation by removing low-level information and normalizing text.

Lastly, work was done to translate text from other languages into English so that data from other countries could be used and interpreted.

DATA ENGINEERING

DATA ENGINEERING - TWITTER

For Twitter, the data engineering work mainly consisted of working on the pipeline and automating tasks for weekly analysis.

The data engineering team needed to convert and save the data from the scraper, coming in as a JSON format, into a format more accessible for the data science team. This was done by writing a script to automatically parse the JSON and write to Google's datastore where it could be used in the pipeline.

The data engineering team also set out to fix a bug with the Twitter listener where some tweets were being truncated. It was discovered that the listener was reading the Tweet status object returned by the API incorrectly. Fixer was an example of a Tweet status object. In Twitter, where the tweet text was previously extracted, in yellow is where the text was now being correctly extracted.

DATA ENGINEERING - YOUTUBE

For YouTube, we wanted to develop a new scraper that did not rely on the YouTube API. We found the YouTube API limiting due to the daily quota with how much data we can collect and lack of options to search for and find relevant ideas. We decided to build our own scraper using Selenium that could freely scrape through videos without worries about limits.

In the end, we developed two scraper versions. The first version is a general scraper that will use search terms to find related videos. The second version is a specific scraper that will scrape videos from a single channel of interest.

TOPIC MODELING - LATENT DIRICHLET ALLOCATION (LDA)

For topic modeling, we used Latent Dirichlet Allocation (LDA). LDA is an unsupervised clustering technique that is commonly used for text analysis. It's a type of topic modeling in which words are represented as topics, and documents are represented as a collection of these words.

loincat:
- Can find latent topic inside documents
- Unsupervised learning requires a true label, which may not be available
- LDA is easy to train
- LDA gives interpretable topics

Each circle represents a topic that contains related words. The distance between circles is how closely related each topic is. The size of each circle is proportional to the number of words in each topic.

The bars represent the frequency of terms in a topic. This is showing terms for topic 1.

MINECRAFT YOUTUBE CHANNEL TOPIC MODELING RESULTS - MAR 21ST, 2023

With the data from the YouTube scraper, we used the same LDA model that we used for the Twitter data. The work showed a view of the most bought up topics from the comments of the top 50 most recent YouTube videos on the official Minecraft channel using our LDA model. While we started by using comments from general Minecraft videos, we found that there were many creator-specific topics shown up. Instead, we decided to pivot to using videos from the official Minecraft channel as we felt that the videos stood more on topic and we would get less creator-specific comments. However, we were still able to perform topic modeling using comments from general Minecraft videos.

SENTIMENT ANALYSIS

For sentiment analysis, we chose to use BERT (Bidirectional Encoder Representations from Transformers). We chose BERT because, as opposed to other NLP models, BERT models demonstrate state-of-the-art accuracy across various tasks.

TA: Veronica Fulbright

A SPECIAL THANKS TO... MATTHEW SKIRVIN AND THE DATA MINE STAFF