Natural Language Processing

*For **sentiment analysis** in the **video game market***

Guilherme Carvalho
Ironhack Data Analytics Final Project

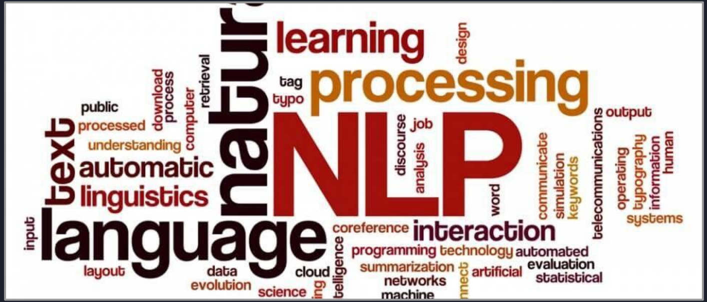
I

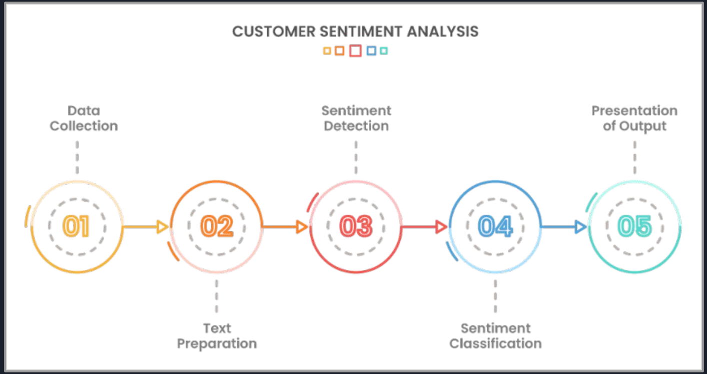
Business Case

Methods:

- **Analysis of user reviews**
  of video game products

- **Natural Language Processing**
  techniques to extracting sentiment and key terms from text




CUSTOMER SENTIMENT ANALYSIS

01 Data Collection

02 Text Preparation

03 Sentiment Detection

04 Sentiment Classification

05 Presentation of Output

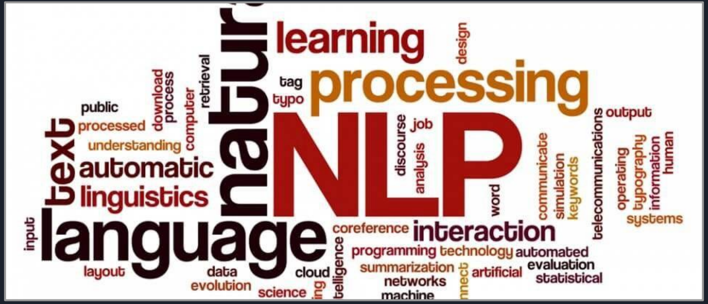
I

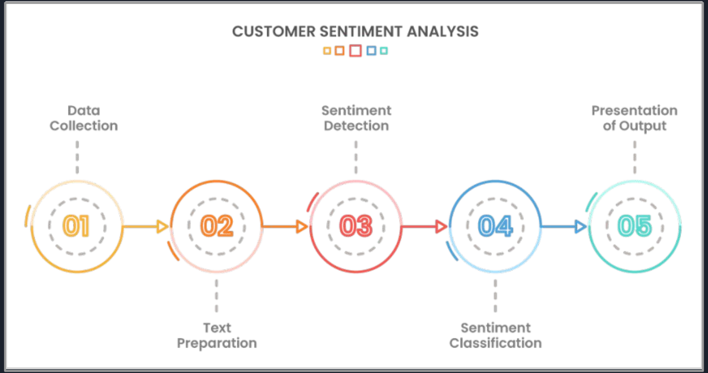
Business Case

Goal:

- ❖ **Identifying products that perform well or poorly** according to user feedback and the factors behind it

- ❖ Useful alongside sales numbers and active player counts for a **deeper analysis of the video game market**

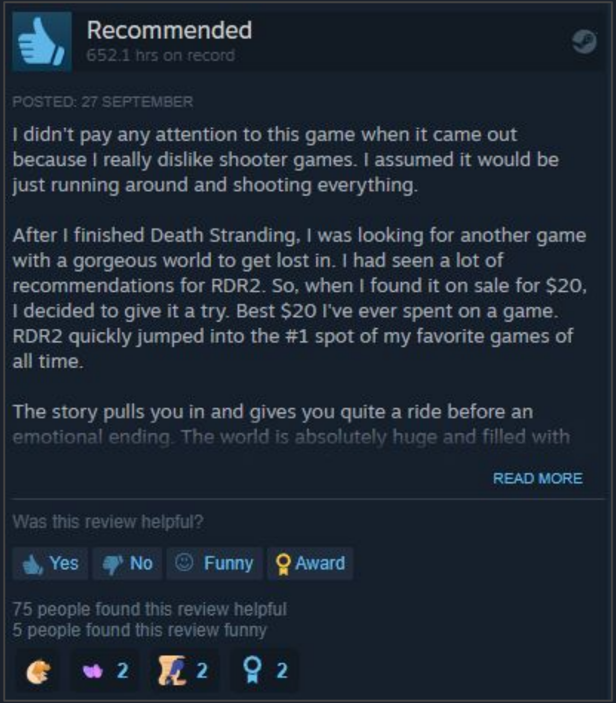



CUSTOMER SENTIMENT ANALYSIS

II

Dataset Overview

- Dataset sourced from **Kaggle**
- User reviews from **Steam**
  Largest video game store
  on PC/Mac/Linux
- **370,000 reviews**
  on **36,000 games**
- Rating system:
  - Recommended / Not Recommended



Recommended
652.1 hrs on record

POSTED: 27 SEPTEMBER

I didn't pay any attention to this game when it came out because I really dislike shooter games. I assumed it would be just running around and shooting everything.

After I finished Death Stranding, I was looking for another game with a gorgeous world to get lost in. I had seen a lot of recommendations for RDR2. So, when I found it on sale for $20, I decided to give it a try. Best $20 I've ever spent on a game. RDR2 quickly jumped into the #1 spot of my favorite games of all time.

The story pulls you in and gives you quite a ride before an emotional ending. The world is absolutely huge and filled with

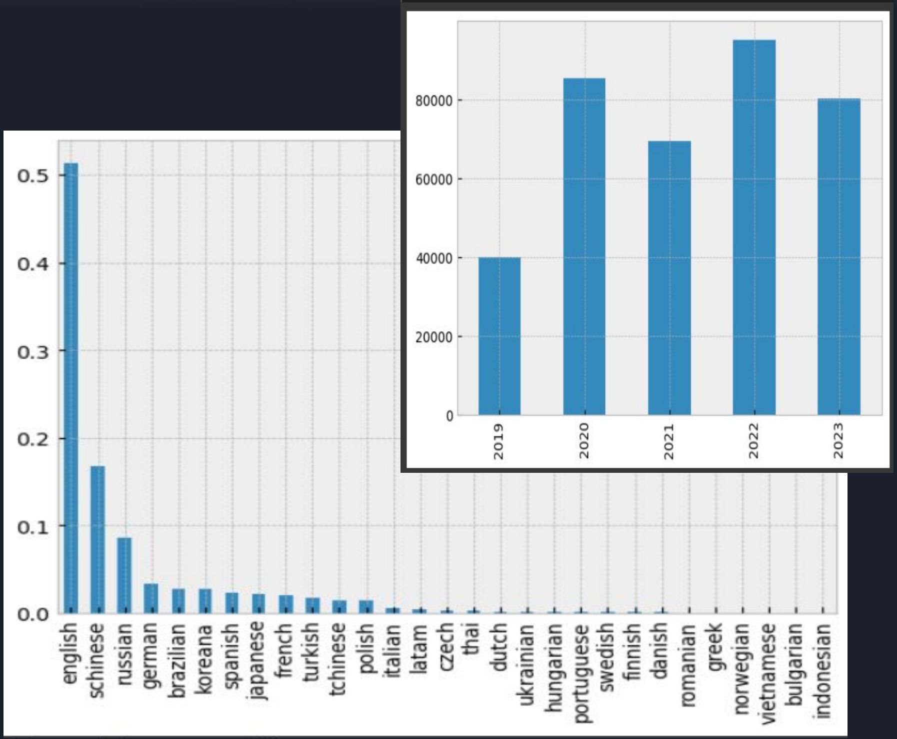
75 people found this review helpful
5 people found this review funny

II

Dataset Overview

- Published between **2019 and 2023**
- **29 languages**
  51% English, 16% Chinese, 8% Russian, 25% other

➢ English only:
190,000 reviews
on **28,000 games**

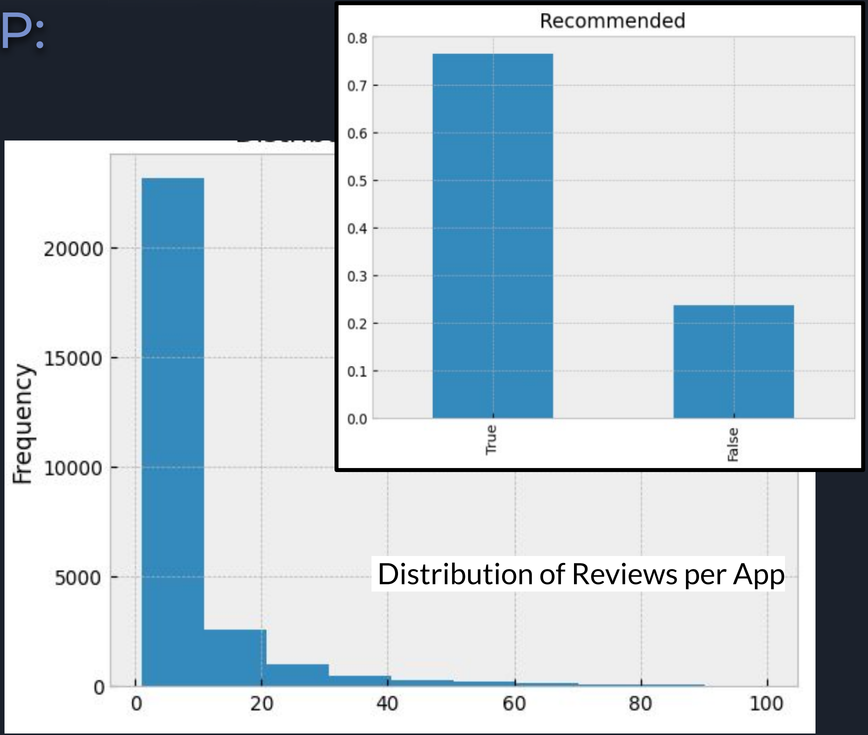


II

Dataset Overview

Feature Selection for NLP:

- ☒ **Review Text**
- ☒ **App name**
- ☑ **Language [English]**
- ☐ Date created / updated
- ☐ Other metrics:
  - Review ID/weight/votes/comments
  - Purchase/gift
  - Chinese market specifics
- ☒ **Recommendation**



Distribution of Reviews per App

III

Procedure Overview

Pre-Processing:

- **Text Cleaning**
  Empty/meaningless text, html tags, ascii art
- **Lemmatization**
  Reducing complexity - improving performance and accuracy
- **Stopwords**
  Eliminating common structural words - irrelevant

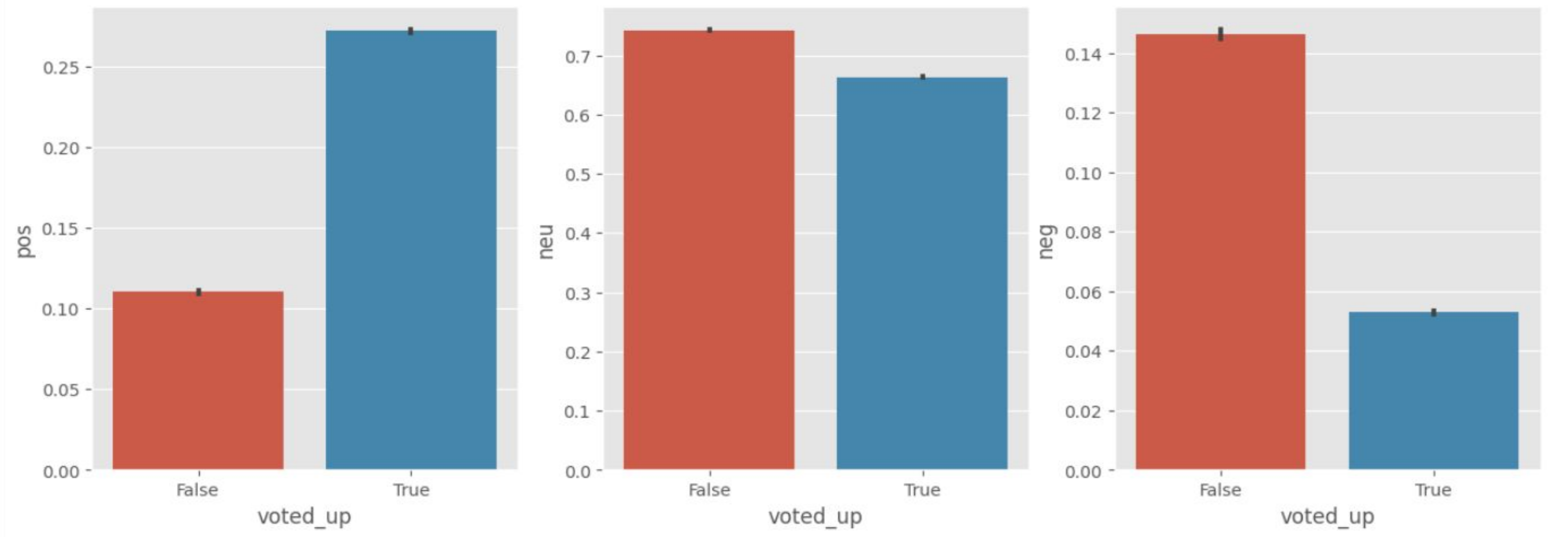
Modeling and Visualization:

- **Sentiment Extraction**
  Rule based analysis - identifying sentiment by based on a set of predefined tags
- **Key-Term Extraction**
  Extracting the most used and most important words
- **Word Clouds**
  Visualization for the extracted data

IV

Sentiment Analysis

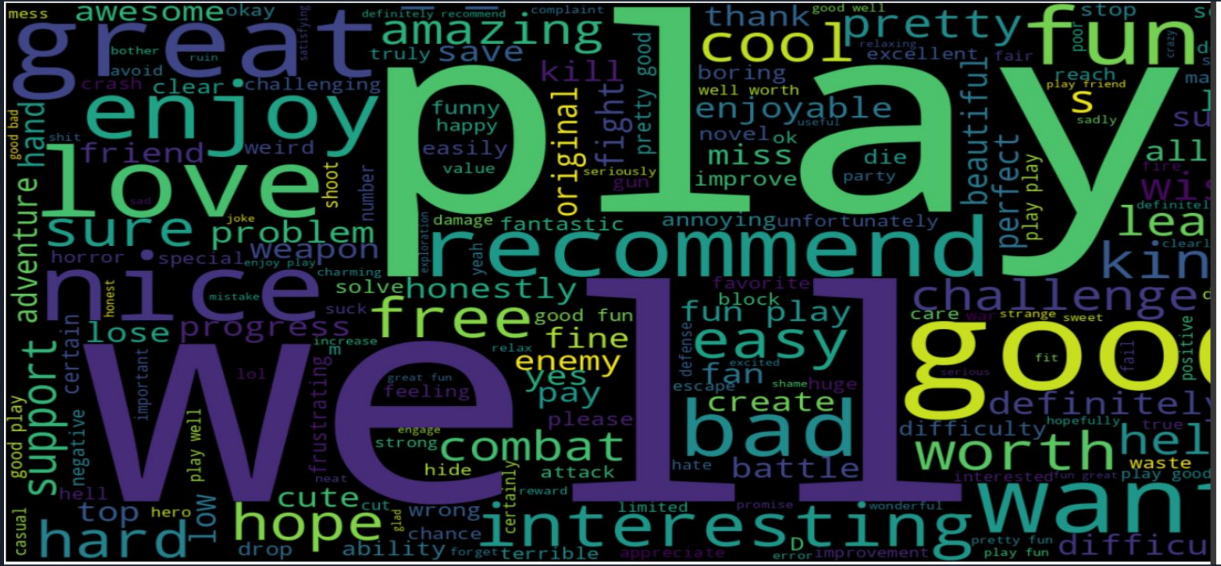
➢ Extracted sentiment vs. Recommendation

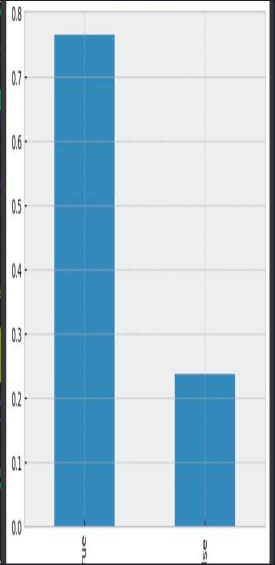


IV

Sentiment Analysis

➢ **Distribution of positive and negative terms across all reviews**

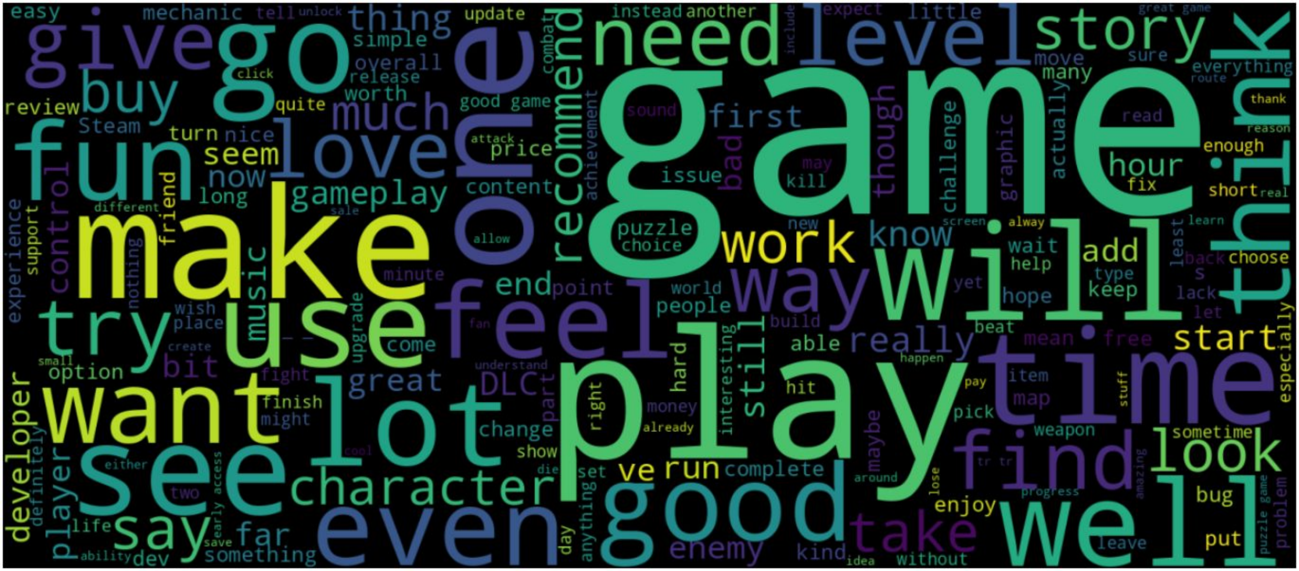




V

Key-Terms Distribution

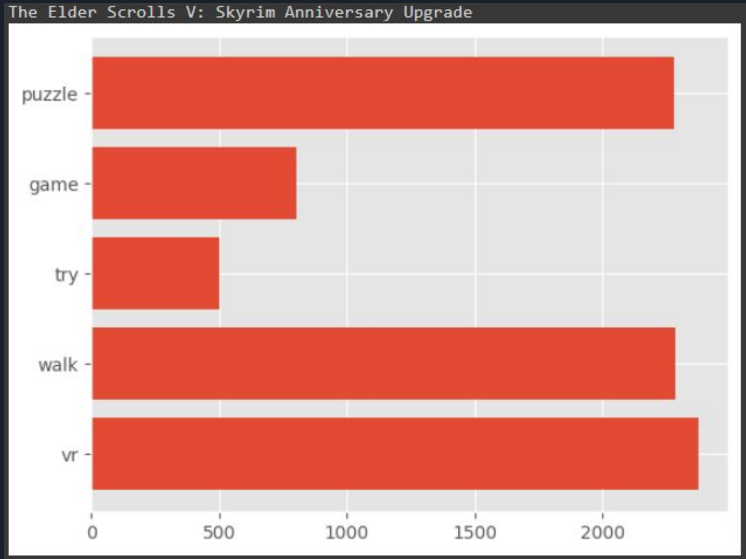
- ➢ Distribution of key terms across all reviews

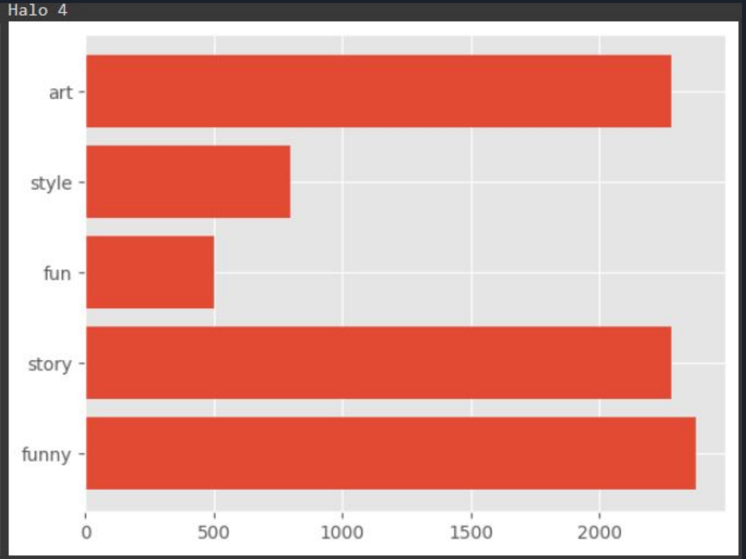


V

Key-Terms Distribution

- ➢ Key terms - Top 5 most common per game (handpicked examples):

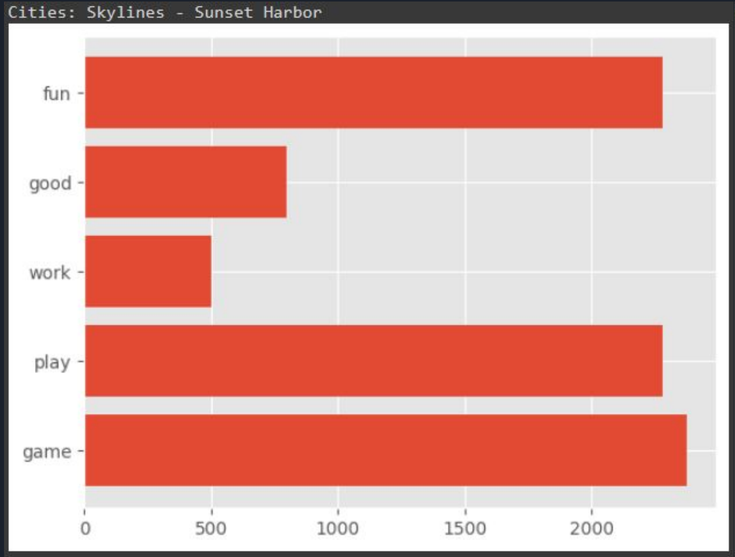


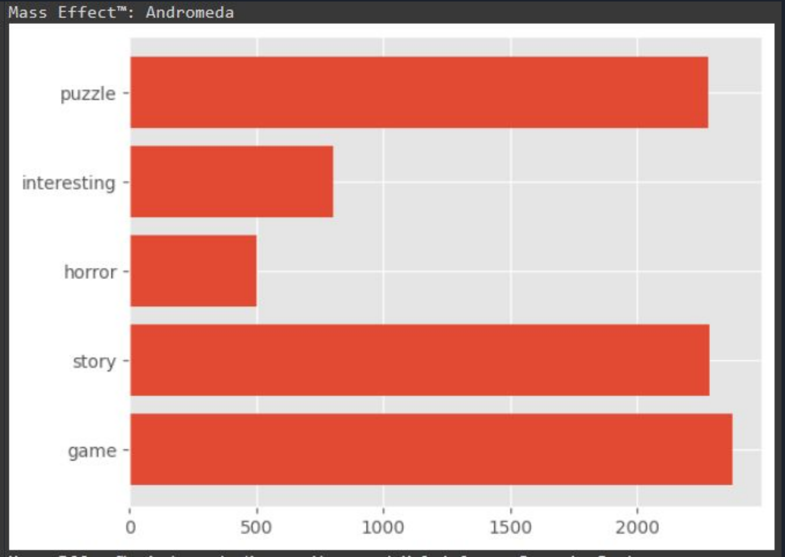


V

Key-Terms Distribution

- ➢ Key terms - Top 5 most common per game (more examples):





VI

Comments about dataset

- *More data about the games present in the set would allow for:*

  ➢ Genre/Category - **grouping games** of a similar nature, **comparing key terms** within groups

  ➢ Sales/Active players - **focusing the analysis** on the **most relevant apps**

  ➢ Date of release/updates - **tracking change** in user sentiment **over time**

VI

Further steps

- **Setting up a Streamlit page** to demonstrate the sentiment and key-terms analysers
- **Expanding the NLP** implementation:
  - **key-term extraction** can be improved by use of **more tools** (ex. entities)
  - to the **other languages** would allow for region based analysis
- **Improving accuracy** by implementing **deep learning models**, such as BERT, which perform **context-aware analysis** in sentiment prediction
- **Sourcing more data** - as previously mentioned to allow for more extensive analysis (Steamworks API)