Milestone Report 1

1. Introduction

Often times, it can be difficult for avid gamers to choose what titles they would want to play and enjoy next, as more and more games are released. Building a recommendation system can help to alleviate this problem and provide a number of titles that best suit a gamer, based on their gaming history, or the preferences of like-minded gamers. Luckily, descriptive game content and user critiques provide a rich assessment in reviewing many games, across many genres, themes, developers, etc., and provide information that can help recommend different titles for a user that they will have a high chance of enjoying. This information, along with user submitted ratings, can be utilized in both collaborative filtering (based on other users) and content based filtering (based on game features).

For content-based filtering, natural language processing (NLP) techniques, such as TF-IDF, are implemented to determine similarities between game features and descriptions, and assign recommendations that best suit the user's taste for the type of game they tend to enjoy. For collaborative filtering, model-based approaches, such as Matrix Factorization and Singular Value Decomposition will use machine learning to find user ratings to unrated games, based on the taste of similar users. An example of user-based collaborative filtering: If I like Halo, and you like Halo and Metal Gear Solid, I will be recommended Metal Gear Solid because we both enjoyed Halo. Content-based filtering will make use of the game's features, such as genre, platform, game description, publishers, and developers. An example of this: Because I like Halo, or games made by the developer company Bungie, I will be recommended Destiny and Destiny 2 because they are games developed by Bungie. These techniques will provide an efficient and accurate tool for gamers to determine which collection of games they should play to scratch their gaming itch!

2. Data Acquisition

2.1 Dataset Description

User-submitted reviews for games was collected using GiantBomb.com API queries, the documentation for which can be found here: https://www.giantbomb.com/api/documentation/. The review dataset contains a unique id corresponding to the game the user is reviewing, which was used to query metadata for the game. In total, there were 24023 user reviews, 6561 users reviewing 4223 games, spanning from July 2008 to September 2019. User features included in

the dataset are: date of review, short review summary, full length review, reviewer, and score. The score will be crucial for the collaborative filtering model-based method, in predicting user ratings for unrated games. Game features include: short game description, full length game description, characters, concepts, developers, franchises, genres, platforms, publishers, themes, and similar games. The 'similar games' feature will be crucial, as it is used to evaluate the content based filtering method by matching games recommended by the system.

2.2 Data Wrangling

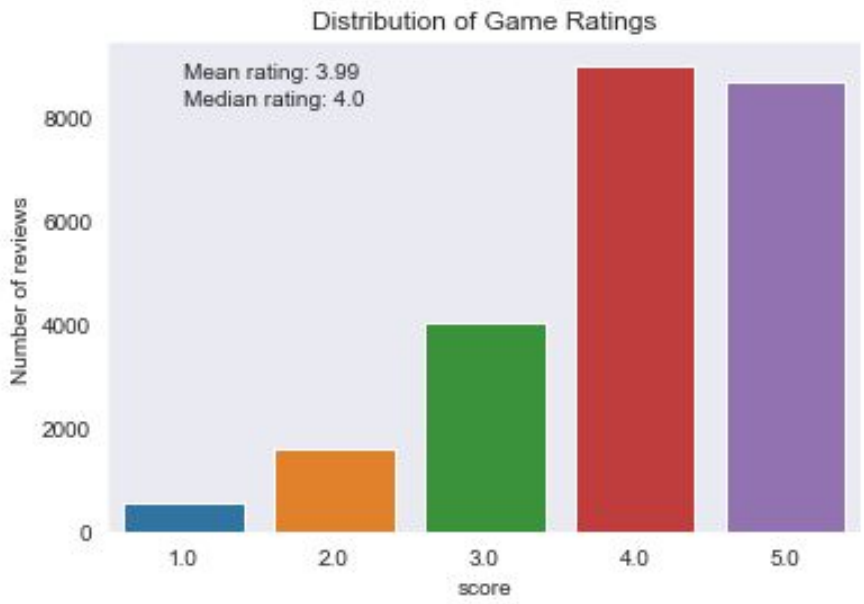
All of the game feature names were nested within lists, which included API links and unique feature ids. Only the names were extracted from these lists, as there can be more than one feature name for each game, as these will be involved in the TF-IDF implementation in finding similarities between game content. Names with spaces between their words, such as the genre "First Person Shooter", were connected with a dash to prevent splitting of descriptive words: "First-Person-Shooter". HTML tags and stopwords (commonly used words) were removed from the user review features and game descriptions to keep the unique and informative words that describe each game. All descriptive feature values were converted to lower-case, to prevent duplicates of different case lettering. Any 's' or 'ies' were removed from word endings, in order to convert plural cases to singular, and lessen the duplicates of plural and singular case letterings There were 1417 reviews missing a value for the 'similar games' feature, approximately 5% of the data, which were omitted.

While all game reviews were given a score from 0 to 5 (whole number), two reviews were given a score of 0.5, for which many of the content and user features were given the value 'test', and as a result, removed from the dataset. All other missing values were replaced with an empty space and ready to be merged to create a 'bag of words' instance, which will be used in the TF-IDF implementation. To ensure that the evaluator feature, 'similar games', list only included games from the reviews, games that were identified to not be part of the list of total unique games (games which cannot be recommended) were removed from each "similar games" list. This process was repeated until each review record's 'similar games' list had at least one game that can be found in the list of total unique games in the dataset. After the removal process, approximately 1500 review records had no games within their 'similar games' feature, and were omitted from the final dataset. After the wrangling process, the final dataset contains 22866 reviews from 3592 users on 4220 games.

3. Data Exploration

3.1 Ratings and Reviews

Looking at the distribution of game ratings:

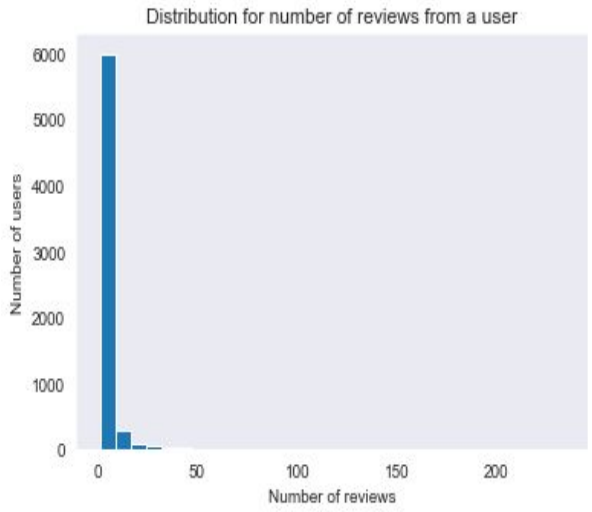


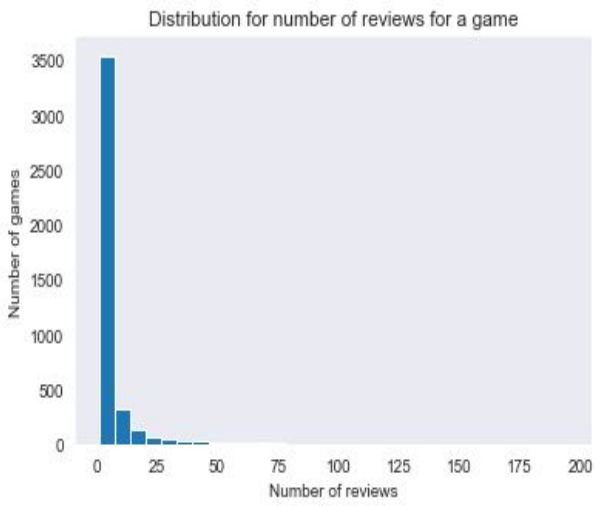
The majority of reviews within the dataset skew heavily to the positive side, with ratings of 4's and 5's. This could be attributed to a variety of reasons, such as:

1. Bias towards popular games
2. Bias towards publishers, developers, or franchises
3. Unwillingness to rate harshly

This positively skewed distribution will affect how the matrix factorization, and prediction of unrated games, will occur, as we expect more positively predicted ratings, based on other user's reviews, as part of collaborative filtering.

Looking at the distribution for the number of reviews as user has, as well as, the distribution for the number of reviews a game has:



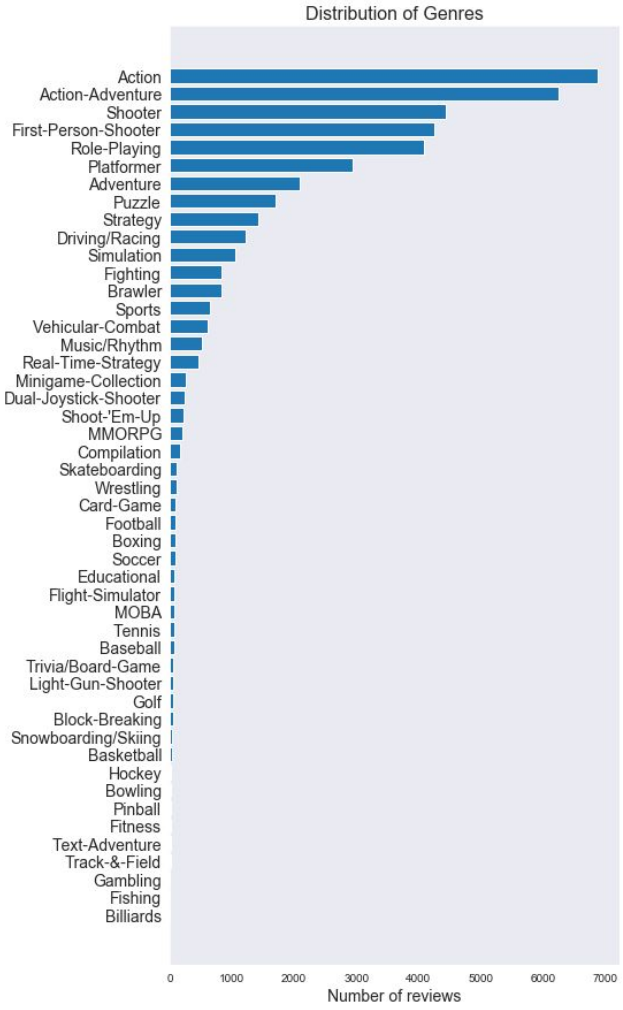


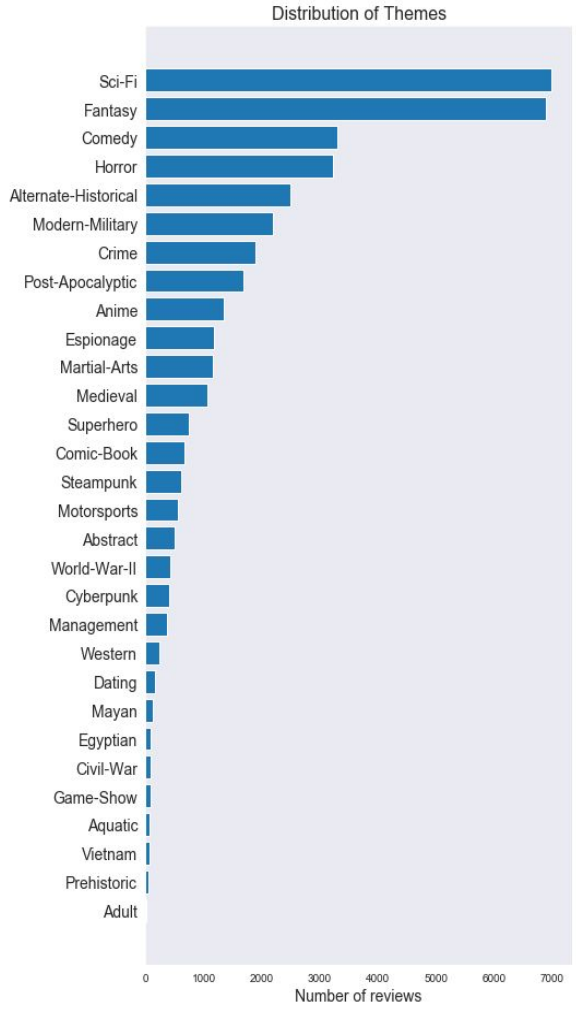
In both distributions, we can observe the presence of a long tail, indicating that many users have only reviewed one game, and many games only have only been reviewed once. This indicates that the data is highly sparse, as not every reviewer has played and rated every game, and we expect many cells in the user-item rating matrix to be empty. Knowing this:

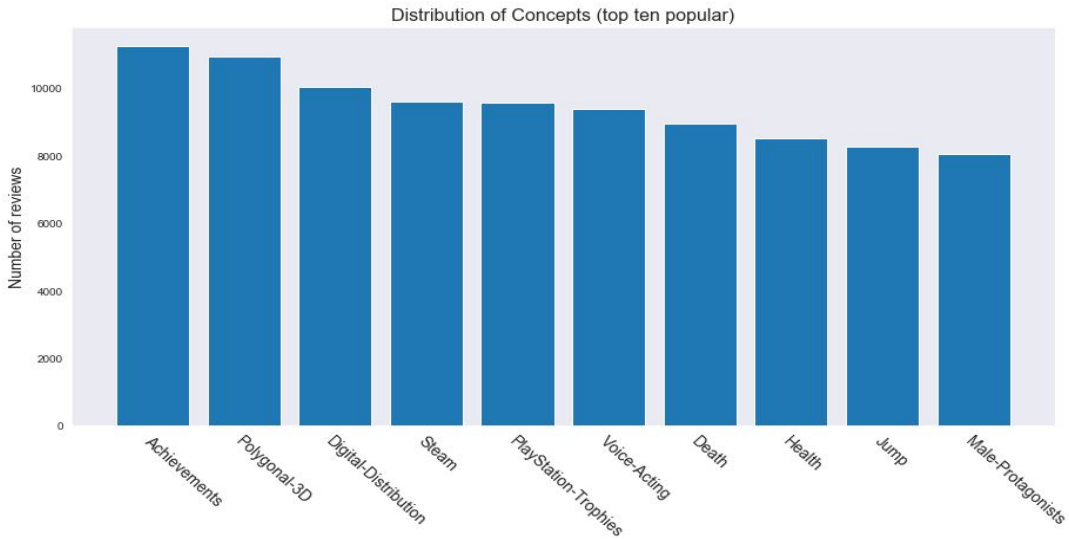
1. We can expect, with the recommendations, that more popular games will tend to be recommended more, whereas games from the 'long tail' section might get ignored.
2. When querying a new video game to the system, the 'cold start' problem may occur, where unless a user has rated these new games, they won't get recommended. The sparsity of the data adds on to that issue.

3.2 Genres, Themes, and Concepts

Most popular features in the reviews:

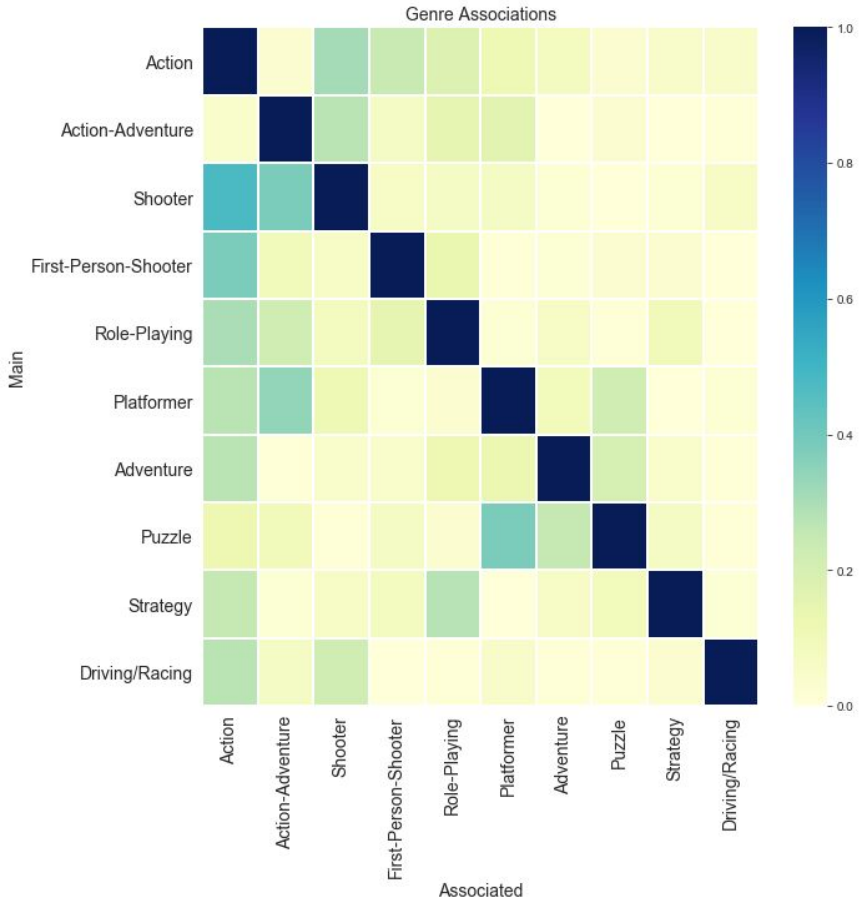

Distribution of Genres
Action
Action-Adventure
Shooter
First-Person-Shooter
Role-Playing
Platformer
Adventure
Puzzle
Strategy
Driving/Racing
Simulation
Fighting
Brawler
Sports
Vehicular-Combat
Music/Rhythm
Real-Time-Strategy
Minigame-Collection
Dual-Joystick-Shooter
Shoot-'Em-Up
MMORPG
Compilation
Skateboarding
Wrestling
Card-Game
Football
Boxing
Soccer
Educational
Flight-Simulator
MOBA
Tennis
Baseball
Trivia/Board-Game
Light-Gun-Shooter
Golf
Block-Breaking
Snowboarding/Skiing
Basketball
Hockey
Bowling
Pinball
Fitness
Text-Adventure
Track-&-Field
Gambling
Fishing
Billiards
0
1000
2000
3000
4000
5000
6000
7000
Number of reviews


Distribution of Themes
Sci-Fi
Fantasy
Comedy
Horror
Alternate-Historical
Modern-Military
Crime
Post-Apocalyptic
Anime
Espionage
Martial-Arts
Medieval
Superhero
Comic-Book
Steampunk
Motorsports
Abstract
World-War-II
Cyberpunk
Management
Western
Dating
Mayan
Egyptian
Civil-War
Game-Show
Aquatic
Vietnam
Prehistoric
Adult
0
1000
2000
3000
4000
5000
6000
7000
Number of reviews


Distribution of Concepts (top ten popular)
Number of reviews
0
2000
4000
6000
8000
10000
Achievements
Polygonal-3D
Digital-Distribution
Steam
PlayStation-Trophies
Voice-Acting
Death
Health
Jump
Male-Protagonists

The majority of games reviewed are in the Action or Action-Adventure genres, and of the Sci-Fi or Fantasy theme. The significant amount of games that contain concepts, as evidenced by the concept frequency plot, are intuitive for an avid gamer, as many games provide in-game incentives such as Achievements (Xbox) and PlaysStation-Trophies (Playstation), be distributed digitally through Steam (popular digital distribution service platform), contain voice-acted non-playable characters, involve a health bar, jumping mechanics, and death sequence with respawn. For content based filtering, we expect that these popular genres, themes, and concepts to have less weight towards determining similarities among games, as most games contain these feature types. It is through associated features, whether it be other genres, themes, concepts, lesser-known features, etc. that similarities can become more filtered to fewer games, and better recommendations can be made.

Looking at the genre associations among the most popular genres:



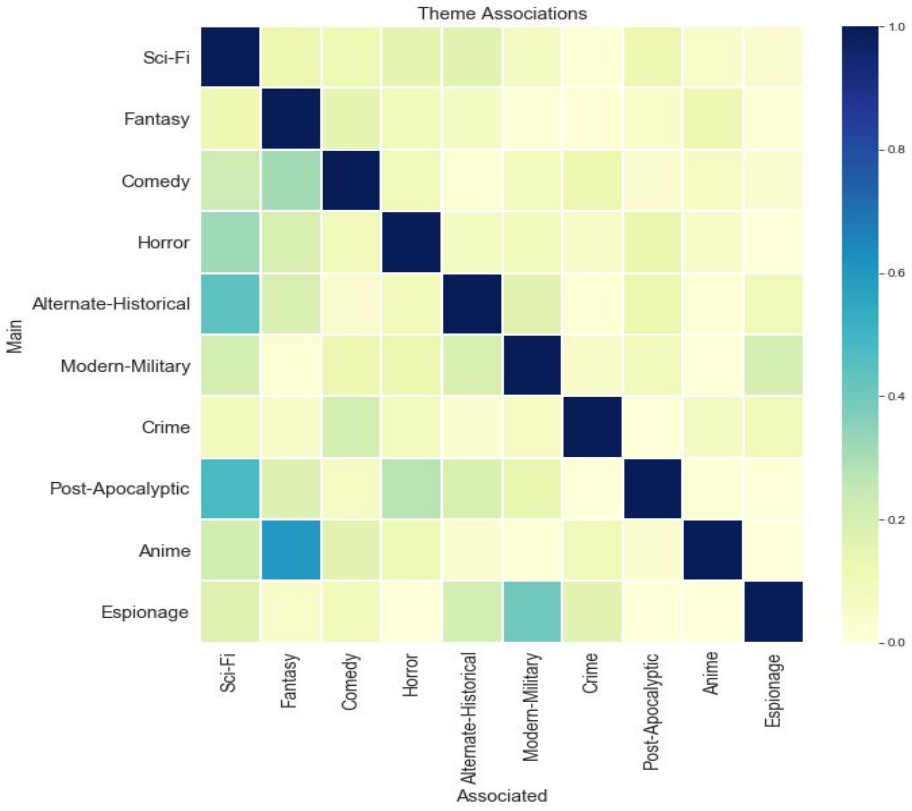
Some examples of genres, with particularly strong associations, included in the plot above and for the less popular genres:

1. Shooter, Flight Simulator, and Vehicular Combat with Action

2. Baseball, Basketball, Fitness, Football, Golf, Hockey, Skateboarding/Skiing, Soccer, Tennis, Track and Field with other sports
3. Gambling with Card Games
4. Real Time Strategy with Strategy

Genres related to sports have moderate to strong associations to each other, which could be attributed to multi-sport games or to a subset of users who play primarily sports games. We expect the recommendation system to recommend titles related to the specific sport first, then widening the list to include other sports. We also expect the recommender system to recommend action games for games related to vehicular combat, shooter, and flight simulator, and card games for games related to gambling.

We look at the same associations in the popular themes:



Some examples of themes, with particularly strong associations, included in the plot above and for the less popular themes:

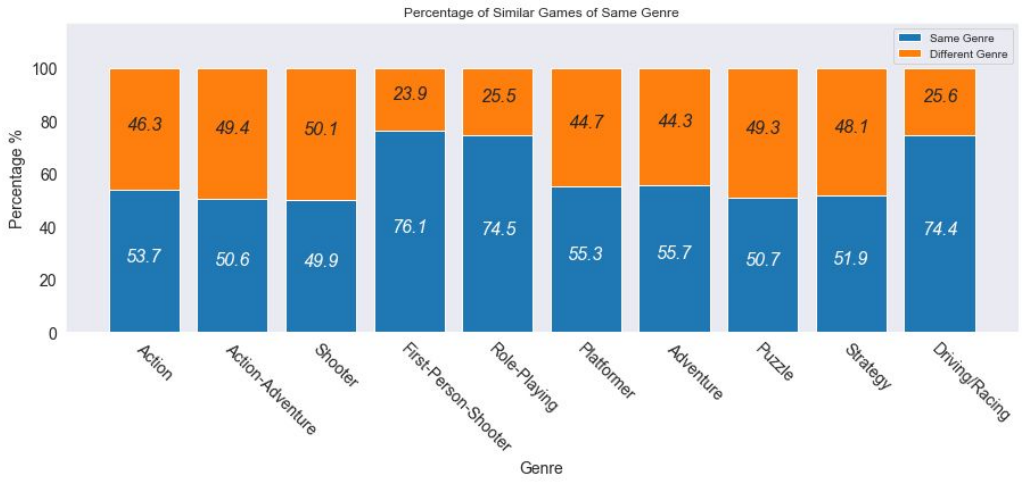
1. Aquatic, Dating, Medieval, and Steampunk with Fantasy
2. Fantasy with Anime

3. Crime with Superhero

Again, using the same logic for the genre associations, we expect the system to recommend games with the same associations as the respective game, then widening the list to the associated theme.

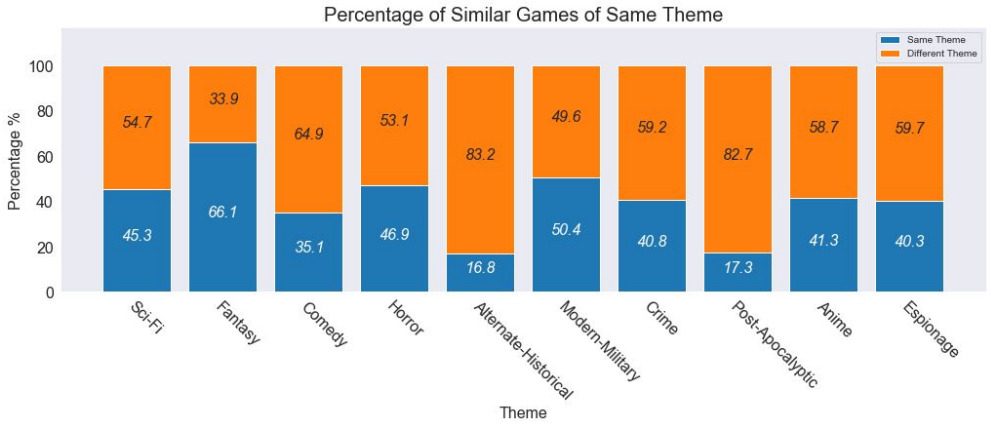
To evaluate the recommendation system, for content based filtering methods, the feature ‘similar games’ provides a list of at least one game that is highly similar to the respective reviewed game. It should be noted however, the recommendation system can still recommend similar games to the respective game, even though the recommender system’s suggestions are not specifically listed in the ‘similar games’ field. The similar games field was created by avid gamers who use GiantBomb and found similarities between games. Reviews may only contain one similar game for the target game, when in reality a game can have various degrees of similarity with multiple other games not listed.

Looking at the percentage of similar games of the same genre (for top then most popular genres) as their respective game:



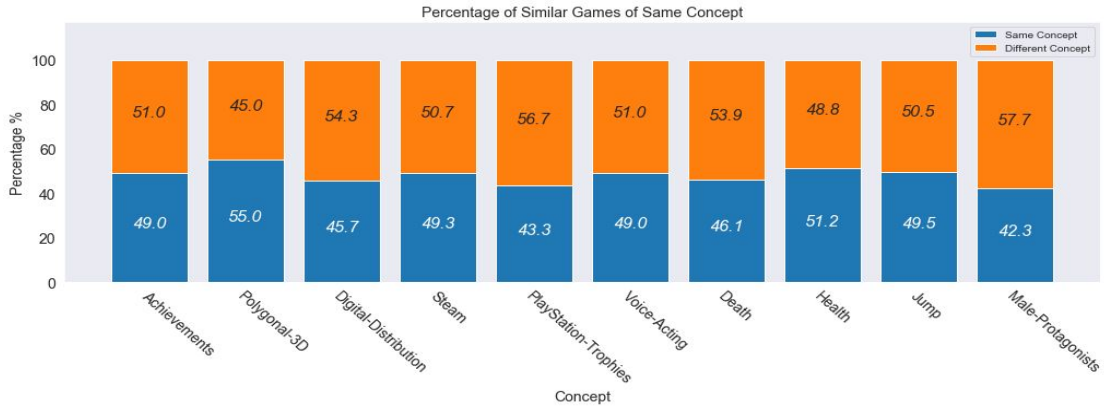
Clearly at least 50%, or more, of the ‘similar games’ are of the same genre as their respective game, especially with First Person Shooter, Role Playing, and Driving/Racing games. This makes sense, as these genres are extremely specific with the type of game, or mechanics of the game, that are characteristic of the genre. Genres such as Action, Action-Adventure, Adventure, etc. are more vague with the type of game mechanics involved. Even still, genre seems to be an important game feature to distinguish games, and will be utilized heavily when assigned importance by the TF-IDF algorithm.

We look at the same thing for themes:



With the exception of the Alternate-Historical and Post-Apocalyptic theme, at least 35%, or higher, similar games contains the same theme as their respective game, especially for Fantasy themed games. It seems that for Alternate-Historical and Post-Apocalyptic themed games, other features will play an important role in determining similarities between games. We can, however, state that genres seems to be a more important factor in determining the games, in the similar games list, over themes.

When analyzing the percentage of similar games that have the same concept as their respective game:

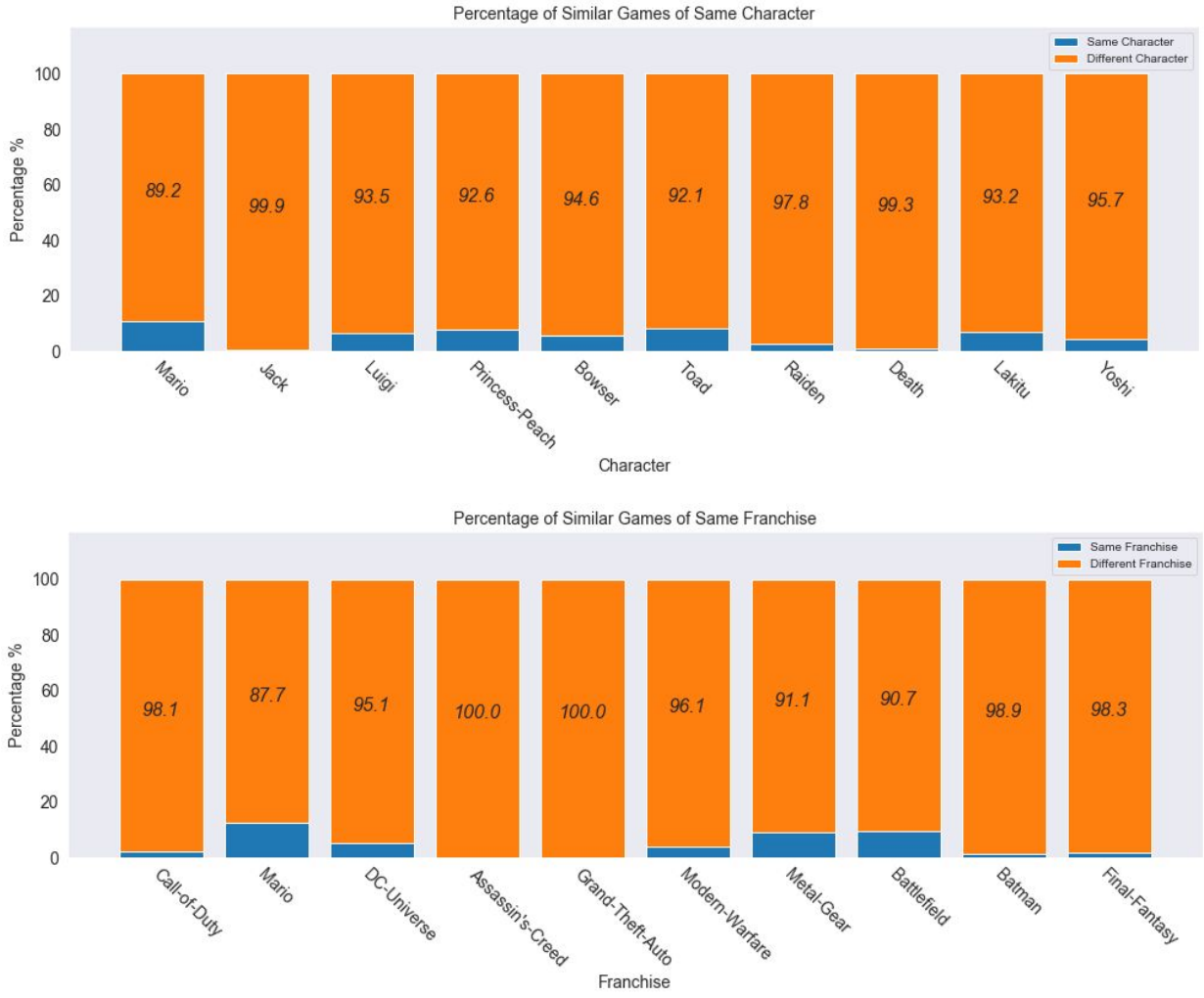


The results are fairly consistent with the top most popular games, approximately 50% of the similar games have the same concept as the target game, which is significant. It should be noted however, based on the distribution of concepts in the games, the amount of games that these popular concepts lie in are extremely large. That leaves a wide array of games that could be recommended based solely on these popular concepts, so additional games features, coupled with these concepts will help narrow the games to be recommended more accurately. However, it can

be concluded that the concepts feature contains more weight than the themes feature, when determining similar games for the target game.

3.3 Characters and Franchises

For the ten most popular characters and franchises reviewed, the percentage of similar games that are of the same character and the same franchise are shown below:

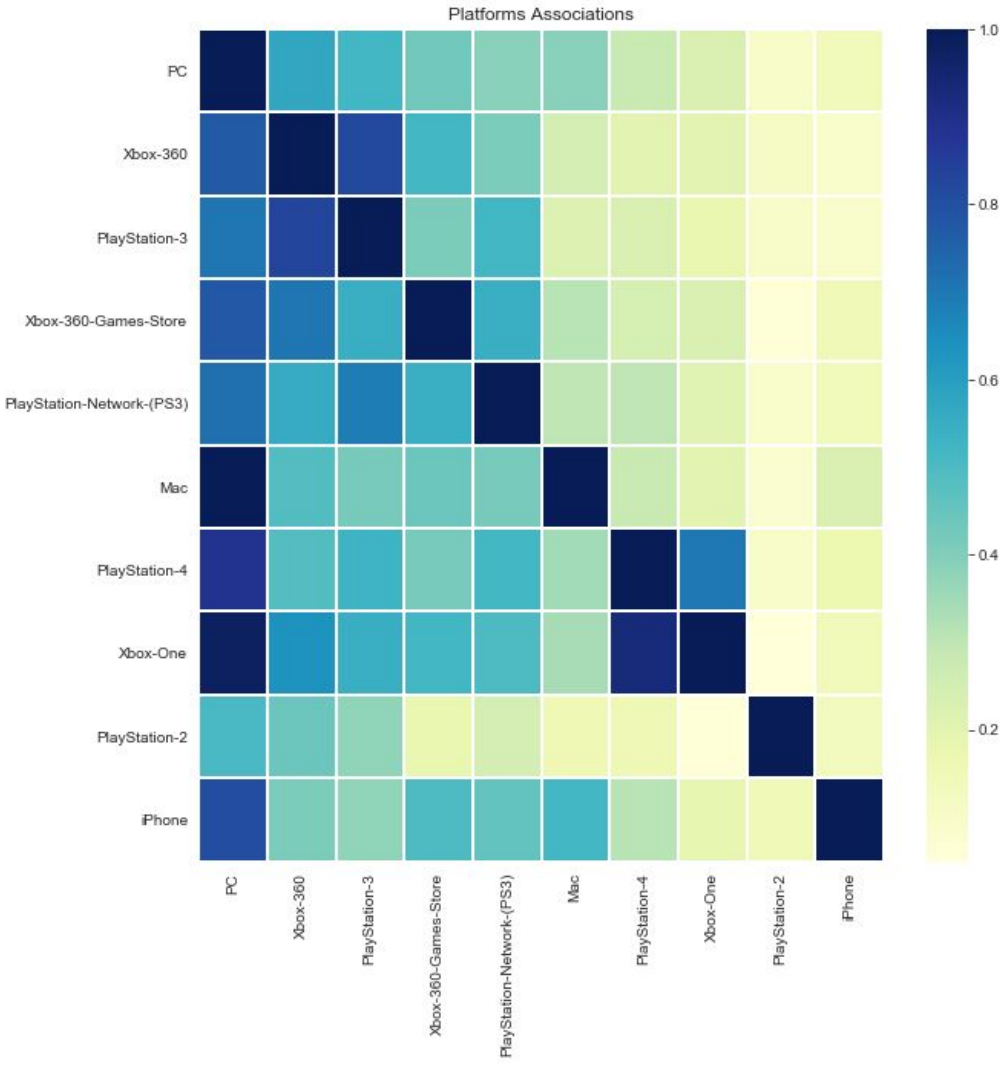


As shown, a significant amount of similar games do not feature the same character as the target game, and is not of the same franchise as the target game. This suggests that the characters and franchises features, for the purposes of the TF-IDF method in content-based filtering, should not be utilized, as these feature values will tend to recommend games of the same franchise and with the same characters, due to the similarities. Therefore, multiple combinations of the aggregation of content based features will be tested for TF-IDF, and we expect the most accurate combination, evaluated by the similar games list feature, will not include the franchise and

characters features. Once again, this does not mean games recommended that are of the same franchise and characters are wrong to be recommended, but for the sake of the evaluation by the similar games feature, they seem to be not important.

3.4 Platforms, Publishers, and Developers

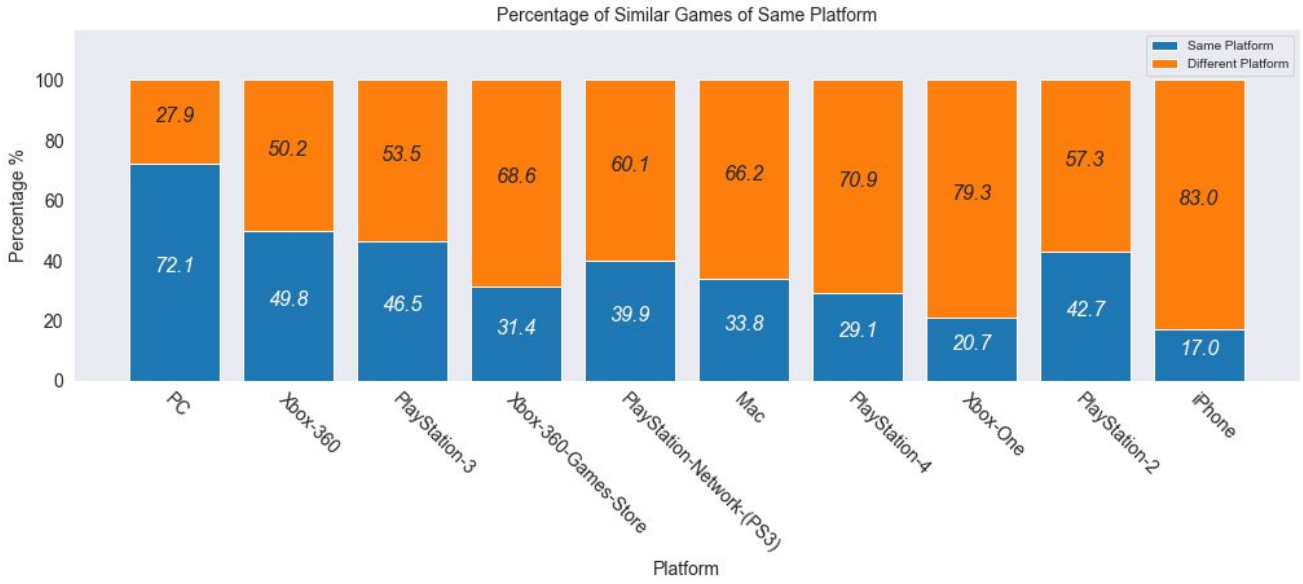
Looking at the association of the top ten common platforms for each game:



This heat-map indicates that many games are cross-platform, especially for games available on PC. This makes sense, as digital distributors are expanding, providing access to games for gamers who don't own, or need to purchase a console. Many games are shared between Xbox (360 and One) and Playstation (3 and 4), with a few console-exclusive exceptions, which also makes sense, as publishers would want the highest possible reach, or audience, to

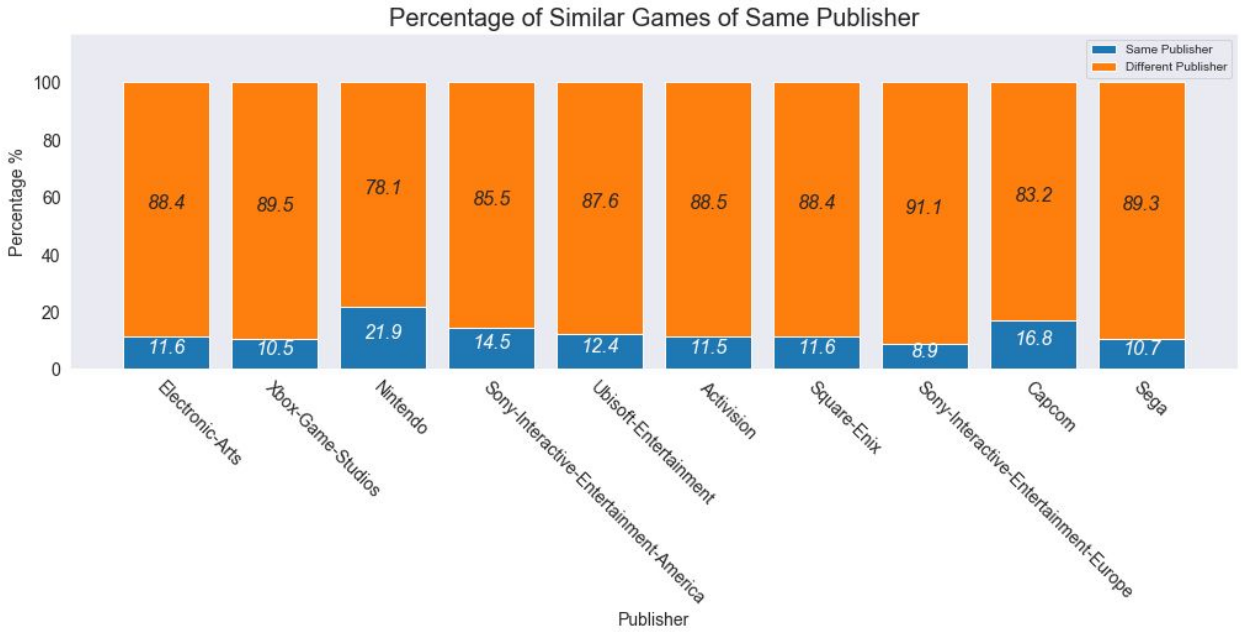
play their games. We expect the recommender system to utilize the platforms feature by recommending console exclusive games for gamers who have enjoyed the same console's exclusive games.

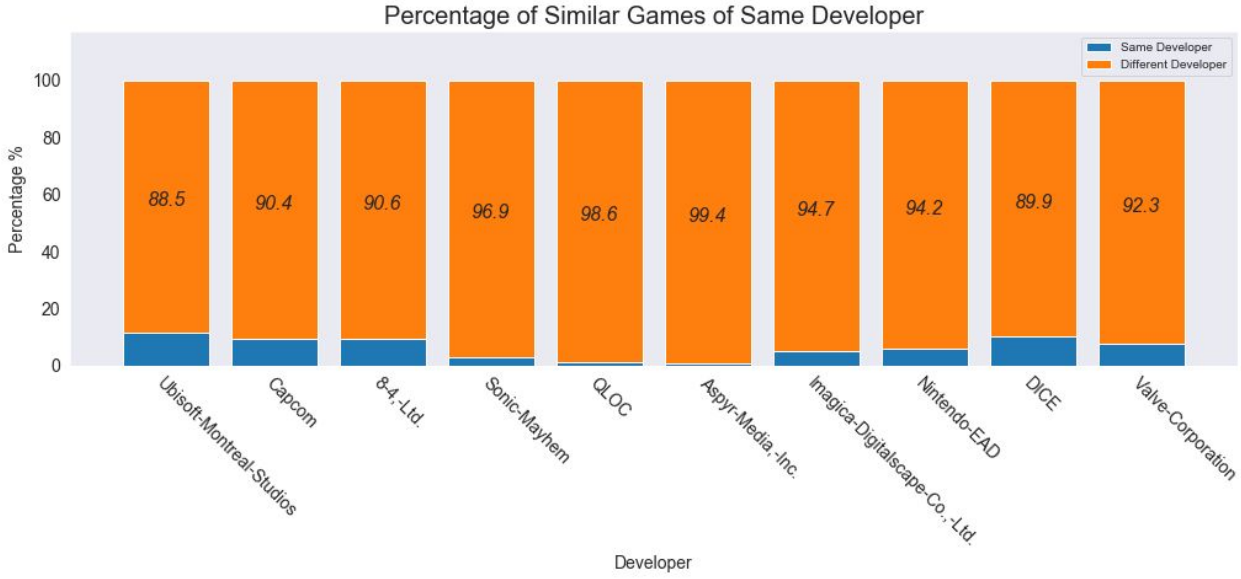
When we observe the percentage of similar games that are of the same platform as the target game:



The similar games tend to be available on the same platform as their target game, especially for PC games, which makes sense, as the majority of games are available on PC. We expect the system to recommend PC games more often for target PC games. For the rest of the platforms, we expect that additional content-based information will help strengthen the accuracy of the system to recommend games that are within the similar-games list, providing an array of games that are both available and not available on the same platform. Therefore, the platform of the games will play a role in determining similarities for the TF-IDF method.

We look at the percentage of similar games that are of the same publisher and developer, for the top ten most popular publishers and developers:

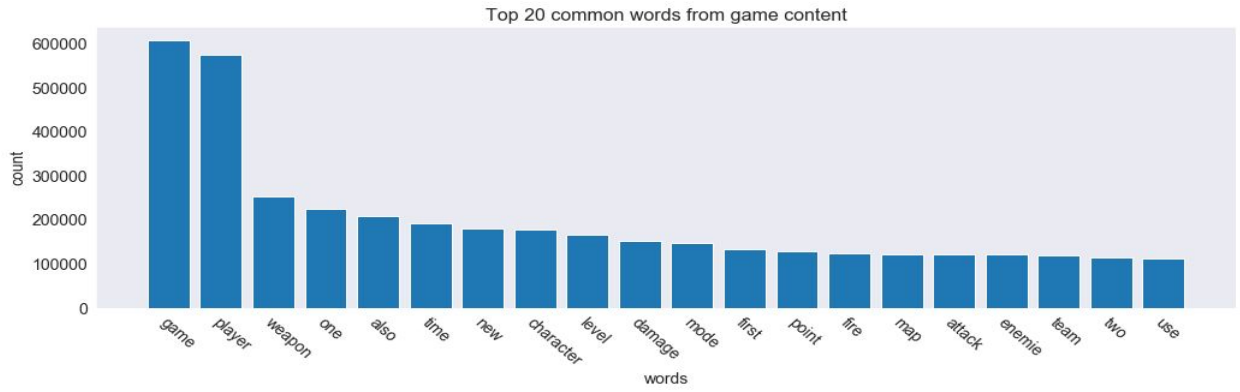


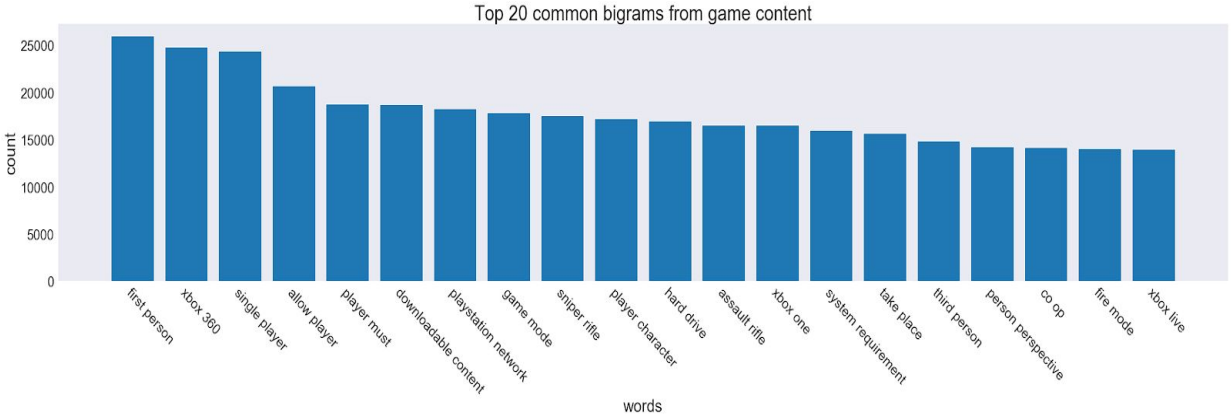


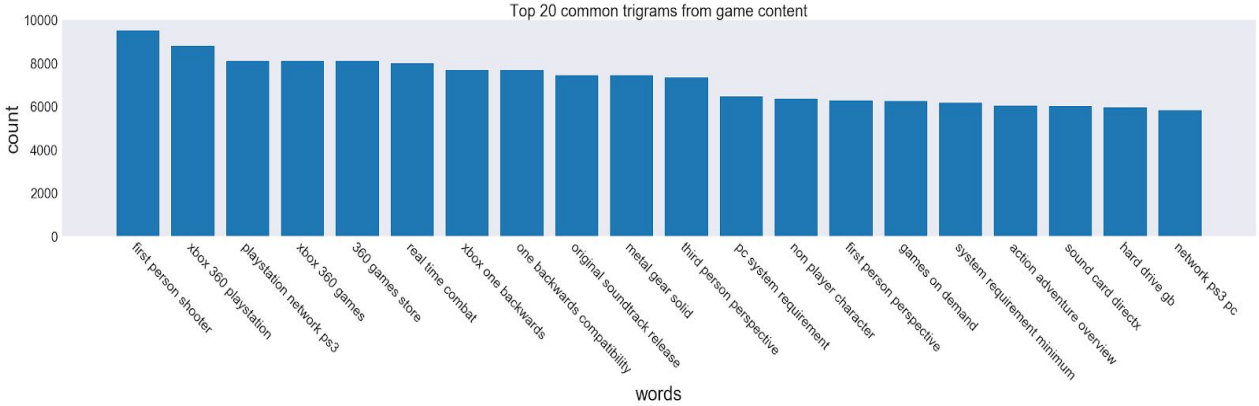
As shown, a significant amount of similar games are not under the same publisher and developer as the target game. This suggests that the publisher and developer features, for the purposes of the TF-IDF method in content-based filtering, should not be utilized, as these feature values will tend to recommend games of the same publisher and developer, due to the similarities. However, just as with the characters and franchises feature, this does not mean that the system recommending games of the same publisher and developer is wrong, but these games will most likely have no matches with the similar games list used to evaluate the system.

3.5 Term Frequency

First, we observe the top terms, bigrams (two words), and trigrams (three words) from the bag of words, which is composed of all of the content-based features:





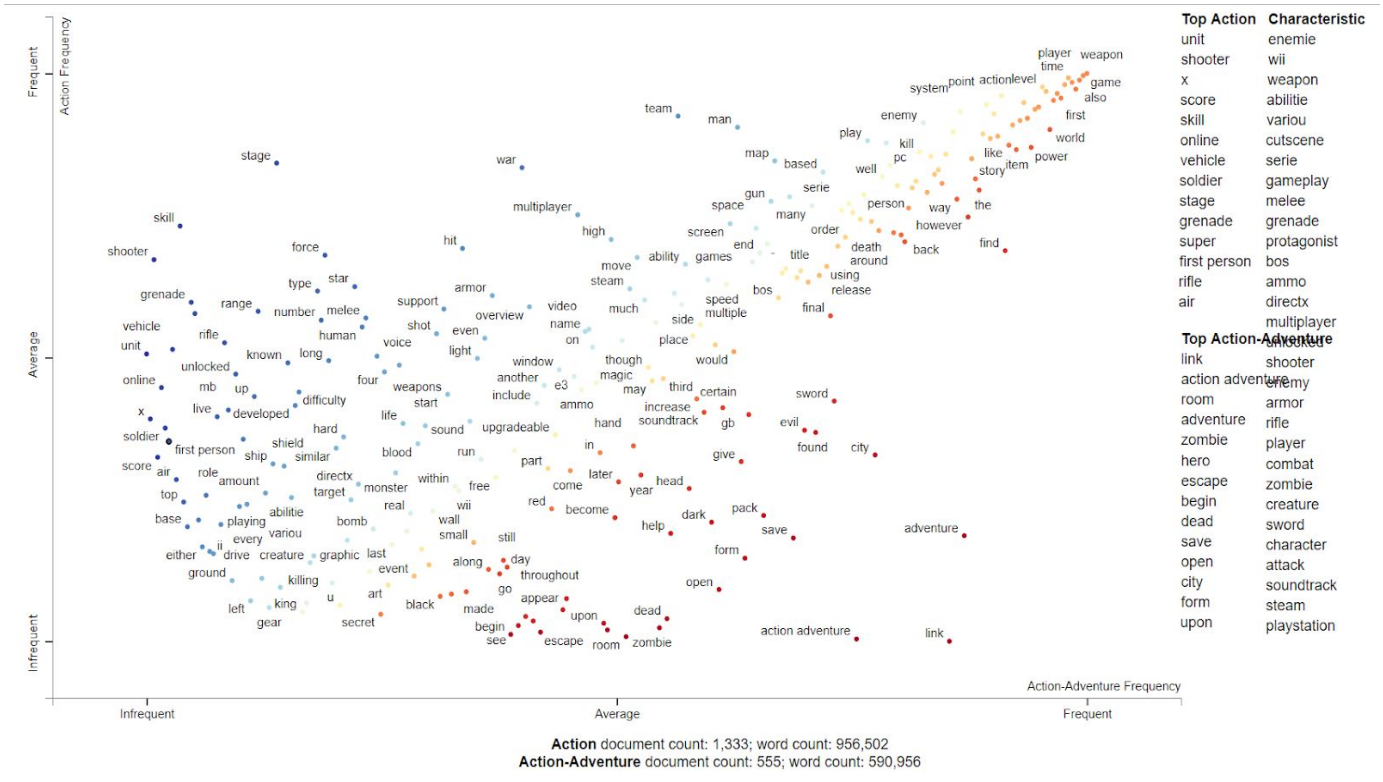


Clearly, each of the popular terms, bigrams, and trigrams refer to common game words and concepts such as “game”, ‘player’, ‘first person’, ‘single player’, “xbox 360 games”, “games on demand”, etc. We expect the system to not assign high weight to these terms and phrases,

when determining what factors will most accurately recommend games that are high similar to the target game, as they are most likely associated with the majority of games in the dataset. Coupled with additional information, such as genres, themes, concepts, etc., will filter out less similar games and provide a more accurate list of similar games for the target game.

Next, because genre and themes seems to be the significant features that would steer the recommendation system towards recommending games that would match the games provided by the evaluating feature list, similar games, we look at how often certain terms are used for games of a certain genre. To observe some examples of how the system will use TF-IDF method to match games of the same genre/theme or to other games of a different genre, based on the terms involved in all of the content-based features, we look at a scattertext plot, which displays term frequency, with a baseline of 500 uses from the bag of words, from two different genres. Specifically, for genre pairs such as Action and Action-Adventure, as well as, Shooter and First-Person-Shooter, we would expect that both genres would relate to a high degree, judging from the name alone, but want to know what specific terms distinguish or relate for both genres.

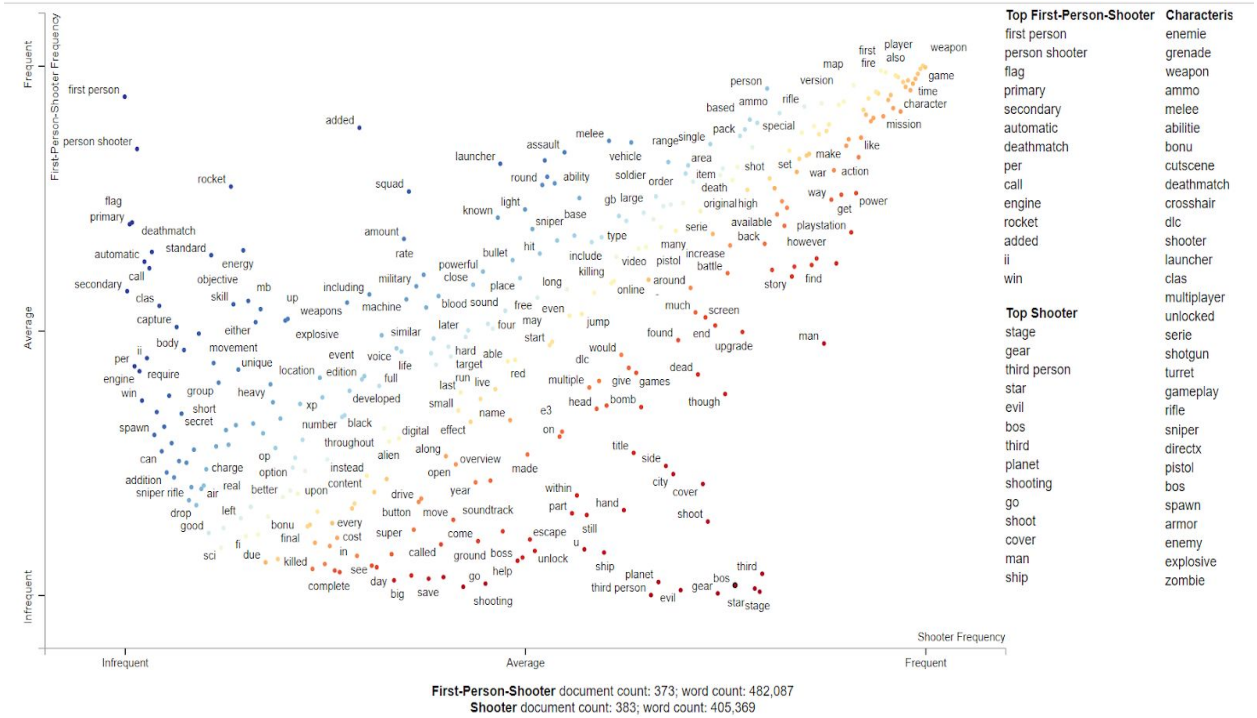
Here is the scattertext plot of the terms for Action and Action-Adventure games:



The visual indicates:

- Terms used frequently among both genres include commonly used game words such as: games, player, system, time, multiplayer, etc. Reference to violence (combat, damage, kill, attack, etc.), weapons (grenade, ammo, shotgun, pistol, etc.), and player mechanics (jump, shoot, cover, etc.) seem to be dominant among both genres. Interestingly, the term ‘action’ is heavily used by both genres, whereas the terms ‘action-adventure’ and ‘adventure’ are exclusive to the Action-Adventure genre.We expect the system to use these terms to recommend both Action and Action-Adventure games.
- Terms more frequently used for Action games, but not Action-Adventure, include shooter and first person (Shooter genre most associated with Action genre), and various war, weapons and vehicular references (rifle, soldier, melee, vehicle, grenade, shield, etc.). We expect these terms to steer the system to recommend more Action games.
- Terms more frequently used for Action-Adventure, but not Action, include fantasy references (hero, zombie, etc.) and player objectives (escape, help, form, see, pack). We expect the system to use these terms to recommend more Action-Adventure games.

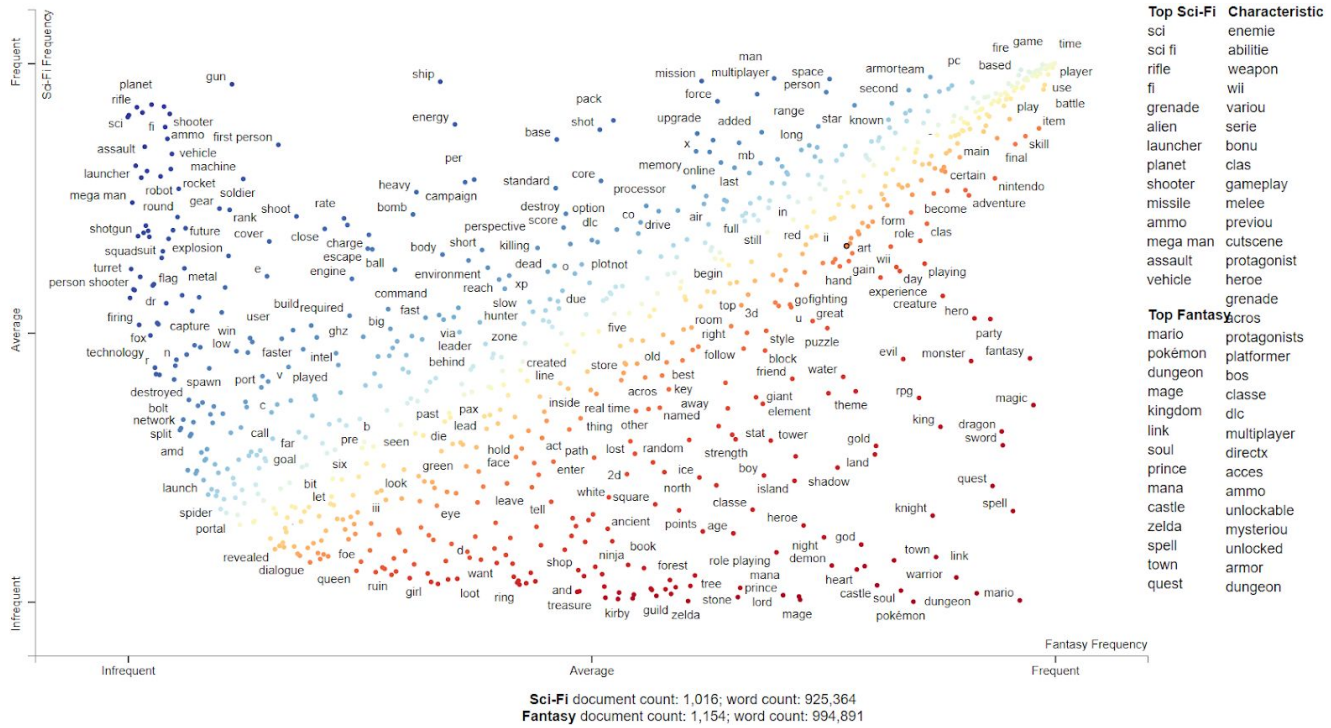
Comparing terms from First-Person-Shooter and Shooter games:



The visual indicates:

- Terms used frequently among both genres include commonly used game words such as: games, player, enemy, gameplay, multiplayer, etc. Large amount of references to violence (combat, damage, kill, attack, etc.) and weapons (grenade, ammo, shotgun, pistol, etc.) seem to be dominant among both genres especially. We expect the system to use these terms to recommend both Shooter and FPS games.
- Terms more frequently used for FPS games, but not Shooter, include specific weapon and game mode terms (first-person, fire mode, fully automatic, deathmatch, etc.). References to popular FPS games such as Halo and Call of Duty are made. We expect these terms to steer the system to recommend more FPS games.
- Terms more frequently used for Shooter, but not FPS, include more references to popular Shooter games such as Uncharted, Metal Gear Solid, Resident Evil. Interestingly, terms such as ‘shoot’, ‘shooting’, etc. are more tied to the Shooter genre rather than the FPS genre. The term “third person” is also widely used, as expected, in this genre, as it is a different perspective shooter than first person perspective. We expect the system to use these terms to steer towards recommend more Shooter games.

Finally we have look at the terms involved in the top two themes, Sci-Fi and Fantasy:



Sci-Fi document count: 1,016; word count: 925,364
Fantasy document count: 1,154; word count: 994,891

- Terms used frequently among both themes include commonly used game words such as: games, player, enemy, playstation, soundtrack, etc. Reference to violence (combat, damage, kill, attack, etc.) and weapons (grenade, ammo, armor, guns, etc.) seem to be dominant among both themes. We expect the system to use these terms to recommend both Sci-Fi and Fantasy games.
- Terms more frequently used for Sci-Fi games, but not Fantasy, include science fiction terms, such as ‘planet’, ‘plasma’, ‘alien’, etc. Popular references to Sci-Fi games such as Halo and Mass Effect are made. We expect these terms to steer the system to recommend more Sci-Fi games.
- Terms more frequently used for Fantasy, but not Sci-Fi, include terms such as ‘dungeon’, ‘dragon’, ‘demon’, ‘knight’, and many other fantasy themed terms. Many references to popular Fantasy games such as Legend of Zelda, Uncharted, Pokemon, and Final Fantasy. We expect the system to use these terms to steer recommendations towards more Fantasy games.