Rachel Phillips

December 8, 2017

Letters & Science 88: Text as Data

How Hard is This Game: A Study of Board Game Instructions

I. Introduction

Board games are a popular pastime with a wide audience. There are no specific cultures, ages, or groups who are more likely to sit down on a Saturday afternoon and play a board game. However, there are specific games that better suit certain cultures, ages, or groups. This project aims to take the first step at building the perfect model to analyze a group of people and introduce them to their perfect game. That first step is predicting difficulty.

It can be very frustrating to gather together a group of friends and set up a game only to discover that the group must first wade through a treacherous twenty pages of instructions before the fun can begin. Is every game with instructions of that length going to be terribly difficult? How can we know, in advance, whether a game will take an hour, or four? How can we know whether participants with little tabletop experience beyond childhood Sorry and Monopoly can cope with the detailed rules and regulations of the latest "drafting worker placement space odyssey" game? By building a model to analyze instructions, we can make predictions to improve players' experience and to improve industry-writing conventions to write rulebooks at a level of detail suitable to the difficulty of their game.

II. Audience

This project is intended for both consumer and industry audiences alike. While the model is usable by consumers, the idea emerged from my personal experiences working in a board game café, called Victory Point Café, in Berkeley, CA. My job activities include organizing games

based on genre and difficulty, analyzing customers' gaming experiences and interests and recommending to them appropriate games, and teaching those games to their party. The work, while both niche and nuanced, is a fascinating compilation of "educated guesses;" my goal with this project is to take both the café and game consumers one step closer to picking the perfect game for them.

III. Research Question

Can I write a classifier to accurately output the difficulty of a game based on the textual content of its instructions?

I undertook this project after witnessing through my work at Victory Point Café the volume of consumers who mistakenly pick a board game that is too difficult or too simple for their preferences, as these uninformed choices often lead to disappointment and diminished quality of play.

For the consumer, an accurate model would mean fewer game returns, more happy players, and a better overall experience. For the designer, it would provide a reflective resource to gauge the quality of a set of instructions, and attempt to prevent the hampering of a simple game by a complicated set of instructions.

This question is specific and covers only one aspect of a board game's ability to entertain a certain party; however, it is the first step towards building a more comprehensive model. Through my exploration of 499 games, I attempted to identify consistent textual features that would correspond to a higher difficulty rating. There are many additional features that could be built into the model. A final goal of the model would be to classify games according to many features, and eventually to be able to recommend ideal games to groups based on their experience, interests, party size, age range, and other preferences.

In considering existent literature that relates to this research, I was unable to find any prior work in the area. I attribute this to the casual nature of this project, which solely seeks to improve the experience of a small fraction of the population that regularly consumes and purchases new board games. However, I am hopeful that an accurate model would help to convince potential consumers who currently feel wary of the structured leisure of board games to try out an easy but interesting new game.

IV. Data Collection and Corpus

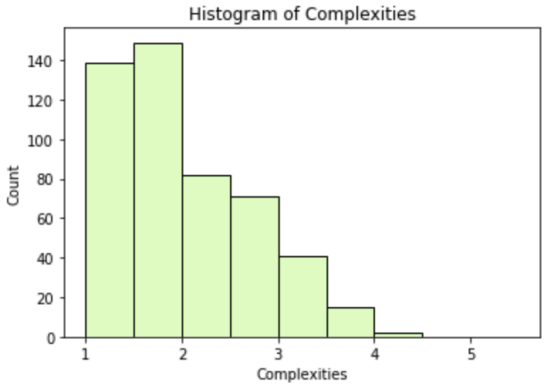
My corpus included all of the games listed on two websites, Board Game Capital and Rio Grande Games. I chose to use these sources because the PDF documents of their instructions had optical character recognition (OCR), which allowed us to easily scrape the text of the documents. Complexity values were scraped from Board Game Geek, a widely accepted authority that has reviews of all well-known (and many lesser-known) board games currently on the market. I then cross-referenced the games based on their titles to correctly correlate complexity values to rules text.

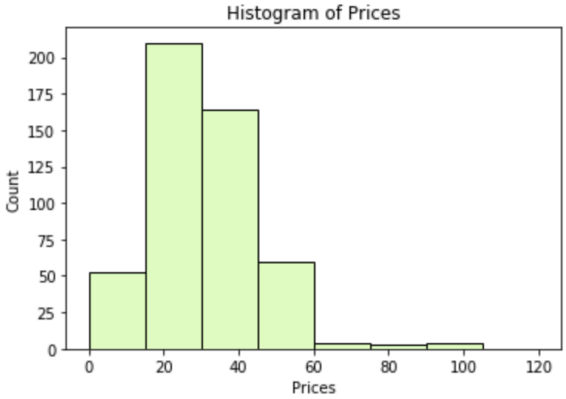
I reformatted the metadata and rules text of each game to make numerical analyses and textual analysis consistent across the dataset. The specific modifications made are available in the Jupyter notebook and are reproducible.

Due to poor OCR on some pdfs, I also made some modifications to the content of the rules text of some games. These changes did not have a significant effect on the content of the text, but removed numerical strings, poorly recognized words, and other irrelevant information from the text strings. I also removed stop words, words common to the English language whose density is typically higher than more unusual words, which I felt would be present in all games; however, I left in typical board game jargon because I did not want to disadvantage games that used

standard wording over thematic wording. [Example: (A) “The players all draw four cards to their hand.” Removing game-specific stop words would leave that sentence as “draw four cards.” (B) “The pirates all seize four doubloons to their ship.” Removing game-specific stop words would yield “pirates seize four doubloons ship.”]

In the future, I feel that the corpus would be strengthened with the incorporation of more data, specifically from websites like Board Game Geek which also have instructional pdfs, to include more modern games and game with higher overall difficulty.





As can be seen from the figures above, both the game difficulties and prices are highly skewed right. Although complexities range from 1 to 5, we have very few values greater than 4. Therefore, most of the games we’re evaluating should be considered “easy,” and in order to create a more accurate classifier, we have skewed borders for easy, medium, and hard. The same is true for the prices, although the distribution is slightly less skewed. More expensive games would likely be more difficult; therefore, it would be good to have more of both in our dataset.

The collection is bounded in that currently it only supports the games available on the two websites mentioned. However, I do not believe that the results of the model will be negated with the introduction of further data. Rather, I feel strongly that introducing more games of a variety

of difficulty and strategy levels will enhance the model and allow us to create more specific classifications.

V. Methods

All of the code for this project is available on GitHub.

The first portion of the program is primarily involved in preprocessing and parameterizing the data. Before I was able to use the data for analysis, I needed to reformat all of the string values into consistent integers (string numbers to integers, hour values to minutes only, etc.). I also wanted to extract from "acceptable player range" the maximum and minimum number of players that a game would support. Finally, I performed some processing of the rules text, which was used to build the classifier. I removed stop words and non-alpha words (sequences of numbers and punctuation, for example), and changed all of the words to lowercase. I also appended some of the metadata, such as number of players and age range, to the text since those would have been features that would be immediately available on the box and in the instructions and I felt that they might contribute to the classification.

For statistical analysis, I created two classifiers – one that used rules text to predict the price of a game, and another that used the rules text to predict the difficulty. I trained the classifier on 300 random games and then tested it on the remaining 199 games. In training, the classifier tries to identify common features of the 300 games that correspond to their classifications. In testing, the classifier tries to recognize similar features in the testing set and makes classifications based on those.

To validate my results, I used a k-fold cross validation with $k = 4$. K-fold cross validation splits the data set into k sections, and then trains with the first $k-1$ sections and tests the kth section. It iterates through combinations of sections so that all possible combinations are tried

(this ends up being *k* combinations). Then it prints the accuracy scores of each attempt and finally a mean score.

I set the random seed before all calculations to ensure reproducibility. The code for the cross validation is publicly available.

VI. Results and Discussion

We were able to train a classifier to modest accuracy. Potential improvements to the accuracy are discussed at length below, in Section VII.

Results of Complexity Classifier

	precision	recall	f1-score	support
easy	0.92	0.94	0.93	121
medium	0.86	0.82	0.84	22
hard	0.85	0.82	0.84	56
avg / total	0.89	0.89	0.89	199

Results of Price Classifier

	precision	recall	f1-score	support
cheap	0.86	0.81	0.84	69
affordable	0.80	0.86	0.83	42
pricey	0.69	0.85	0.77	48
expensive	0.90	0.68	0.77	40
avg / total	0.82	0.80	0.80	199

Shown above are the results of building the initial classifier, training on 300 random data entries and testing on the remaining 199 entries.

Precision is measured as the number of games that were correctly classified in a certain category (i.e. in the Price Classifier results, 86% of the games that were classified as “cheap” were, in fact, cheap). Recall is measured as the number of games in a certain category that were correctly classified as that category (i.e. 81% of the cheap games were classified as “cheap”).

Overall, these charts show that we were able to train a relatively accurate classifier. The attempt to classify on difficulty, however, was significantly more effective. These results are diminished slightly under K-fold cross validation. The mean of the precision scores for price is .59, while the mean of the difficulty scores is .78. The difficulty is significantly easier to predict based on the rules text.

I feel that further exploration of this classifier, testing of different parameters for defining categories (more categories, different breaking points), as well as the addition of more varied

data (see Section VII for more information) would improve the results. However, amidst significant deterrents in the data and inconsistencies, I feel that .78 implies that there might be a correlation between certain textual features and games, a connection that could be explored through further research.

VII. Problems and Improvements

While the accuracy scores are much higher than I expected, they are still too low to consider the classifier entirely accurate. I think that this problem could be minimized with the introduction of more games into the corpus, particularly more modern games of higher difficulty to even out the spectrum of difficulty of our sample.

Additionally, I think that the OCR of the rules text does not give the best picture of how the games' instructions are actually written. For example, here are some excerpts from the processed text:

> "ee ee tay end round player played druid cards round players continue playing lon g druid cards players druid cards round ends note player druid cards left takes con secutive turns played game end druid reached goal players play round player draw s new druid cards adding hand"

> "st yr eggertspiele table contents object game set special rules players derby leagu e game carrots bid racing form horse sense tip game materials ditick rules change horses best game comes"

> "deroulement du jeu deplacement les poules et les cors se deplacent sur les oeufs qui representent le chemin dans le sens des aiguilles plus jeune poule ou le plus j eune commence elle ou il choisit une carte de la la regarde et la montre tous les j oueurs si la carte"

These are three random excerpts from instructions. One contains the same words over and over, the second contains gibberish, and the third is entirely in French. As we can see, a superior of method of game selection and of filtering badly processed instructions would likely improve the model significantly.

VIII. Critical Reflection

My decision to include the metadata as text in the rules text had an effect on my results. I feel strongly that this text was a reasonable addition to the rules text, as all rulebooks have this information on their covers; however, it gives a concrete shape and requirement to all texts to include these features. Additionally, the metadata sentences would be the only ones in English in instruction booklets that are, as in the example above, entirely French.

Additionally, I had to structure my standards for classification specifically to fit my data. If I had more data with a more equal distribution of complexities and prices, then the delineations between easy, medium, and hard games would likely be more clear and simpler to identify. In this case, I simply tried to make relatively even numbers of games in each category to help train the classifier.

This study's results show that it is possible to build a model that relates board game instructions and their textual features to difficulty level, and possibly other characteristics as well. Unfortunately, because games are published and re-published, it is impossible to correlate difficulty with the release date of the game, but I think that further exploration of that subject would reveal that games have only increased in difficulty with time.

This study in its current form tells us little about society; however, with several changes it has the potential to interpret personal preferences and output game suggestions. This process is a job some are paid to do, and the automation of it would affect a small sector of society as well as improve the gameplay experiences of many potential consumers.

IX. Potential New Paths

Unfortunately, this subject is relatively unexplored. Because it lacks significant societal relevance, it is unlikely to be pursued further by researchers. There are very few markets for

improving board game instructions, and so this project focuses on a particularly niche area of literature.

For those in the game industry, writing good instructions is important. Unintelligible or overcomplicated instructions make a game unplayable and alienate it from its intended audience. Therefore, a model such as this could provide good feedback for those writing the instructions. Are the instructions written in a manner that accurately reflects the difficulty of the game? Or, instead, are they written in dense language that renders a children's game a classification of "hard" and "expensive." This project could provide a feedback opportunity for game designers and help their games succeed better in stores and in cafes.

For consumers, a model like this is also important. It's always difficult to invest in a new game without knowing how difficult it will be, how long it will take, and whether or not it will be suitable to one's game-playing group. With refinement, I envision this model growing to take user inputs. For a more personalized classification, customers could input their game preferences/experience, their player count, their age range, or a host of other characteristics, and the model could produce a list of the top 10 games that match these descriptions.

Another challenging aspect of board games is their varying degrees of language dependency. When I am recommending board games at work, I often take into account the degree of English proficiency of the party, as most games are targeted towards an audience that is fluent in English. Games with heavy terminology or wordy cards are challenging for guests who don't feel confident with everyday English usage, and for whom board game jargon is even more foreign. The model could be extended to predict a score of English dependency, which could be used by customers to evaluate whether a particular game, regardless of its other classification, would suit them.