Group 5: Emma Kramer, Chuhan Peng, Yujie Fu, Zhoulin Zhuo
Text Mining
Final Paper
5/7/2021

Cards Against Humanity Amazon Reviews

Executive summary

Our project utilized a dataset of Amazon.com product reviews. This dataset was originally very large so we narrowed the focus of our analysis to just be of one product from the toys and games category. We analyzed the reviews of the popular game Cards Against Humanity. The objective of our project is to analyze the Cards Against Humanity reviews in order to discover important insights that the company can use to increase the sales of Cards Against Humanity. We analyzed the Cards Against Humanity reviews using multiple text mining methods. We first performed data preprocessing on the data to normalize it. Then we performed sentiment analysis as well as topic modeling. We found that most of the game reviews were positive and our sentiment analysis models were able to accurately predict which reviews were positive with an accuracy rate of 91%. We also separated the positive reviews into four topics and the negative reviews into three topics and named those topics based on the most prevalent words in them.

We found that most people have overall positive feelings about Cards Against Humanity but there are some people that have issues with the game. Cards Against Humanity has many devoted fans, people frequently wrote “fun”, “funny”, “love" and “party game”. But, Cards Against Humanity should also be aware of the negative comments that people leave about the game - namely that it is “expensive”, “rude”, problematic and “dull”. Based on our analysis we made a few key recommendations for Cards Against Humanity. First, they should expand on their current advantages and minimize their disadvantages. Second, they need to compete against competitors better. Finally, they should utilize the keywords we discovered to improve their rank on the search engines.

One of the main shortcomings of our project and analysis is that we are only working with a small number of product reviews in total. Additionally, only a very small number of the reviews are negative reviews which skews our data.

Data description

For our final project data we found and utilized an online dataset of Amazon.com product reviews collected by Julian McAuley, an Associate Professor of computer science at UCSD. The original complete dataset contains Amazon product reviews including 142.8 million reviews spanning from May 1996 until July 2014. We chose to use a subset of this data that was just reviews of toys and games sold on Amazon. As a sample of reviews of toys and games sold on Amazon, our data does not have a huge sample but contains all the necessary information for topical analysis. The size also makes it easy and fast to run code. However, since the data is only a small sample while the original dataset is really huge, the sample will not correspond to the whole original dataset. So, there may be random bias existing in our analysis.We then narrowed this dataset down even further and specifically looked at reviews for the game Cards Against Humanity. We analyzed 309 reviews of Card Against Humanity in total. The dates these reviews were posted on Amazon ranges from 2013 to 2014.

The dataset contains many variables including ‘reviewerID, ASIN, Reviewer Name, Helpful, Review text, overall, summary, and review time. ReviewerID is the first variable and it is a unique identifier for each person who left a review on Amazon. ASIN stands for Amazon Standard Identification Number and it is a unique number used to identify each product sold on Amazon. Reviewer Name is the person who left the review’s name. Helpful is the number of people who voted that someone’s review of a product was helpful. Review text is the actual full text that a person wrote in their review. ‘Overall’ is the score or amount of stars (on a sale of 1-5, 5 being best 1 being worst) that the reviewer gave the product.

Summary is a brief summary of the review. Review time is the date that the review was posted. We only used the variables ASIN, review text, overall and review time for our analysis.

There are some possible shortcomings of this dataset. There is definitely selection bias in this dataset because only certain people decide to leave a review on a product they bought on Amazon. Usually people only leave a review if they are extremely satisfied or unsatisfied with the product. Therefore, the average sentiment score calculated from reviews could be biased. We also do not know that people are not receiving an incentive from the company to write a review on Amazon. In the case of our data, which is reviews of Cards Against Humanity, it appears that the reviews are very positively skewed, mostly overall scores of 4s and 5s (Figure 1). Another concern is that we do not know if the chosen 309 reviews between 2013 and 2014 are really representative of the whole customer base. We could gain insights of customer reviews in this one year from this dataset, but we are not sure that these insights could be applied to the company at a more general level. It would be better to expand the timeframe and include more review data to reach a more general conclusion of customer reviews. Another issue could be that because these reviews are left by random people there could be many spelling and grammar issues in the review text.

Project objectives

The objective of our project is to analyze the Cards Against Humanity reviews using text mining tools in order to discover insights that the company can use to increase the sales of Cards Against Humanity. One of our main objectives is to first determine if each product review is mostly positive or negative by evaluating the sentiment of the words in the review. Then we will use topic modeling to find the key themes and topics being discussed in the product reviews. We will use these insights to help the company better understand customers' preferences so that they can improve overall customer satisfaction. We also want to uncover the most common positive words in the product reviews and then repurpose these key words and use them in the product description or advertising to encourage the customers to buy the game. We will also use the topics of the negative and positive reviews to suggest how Cards Against Humanity can improve future versions of the game.

Methodology

In order to accomplish the above objectives, we analyzed what comprises the overall positive or negative sentiment score by grouping only positive score reviews and negative score reviews and then performed topic modeling. By doing this, we could better understand in general what factors lead to positive reviews or negative reviews and these insights would be very helpful for the company to improve its product.

To begin our text mining analysis the first thing we had to do was data pre processing. We used normalization to change the json data to the data type that we can conveniently analyze with. After normalizing the data we performed vectorization. Text vectorization is the process of feature extraction from text data, that is the process of creating variables for each observation, where an observation is a text document. We'll consider the bag-of-words and TF-IDF representations of text. We used bag of words to help us to get a matrix that clearly shows which words are frequent and which words are unimportant by looking at the count numbers. However, there are some distracting features like numbers and product names in the matrix, so we decided to filter these feature names by deleting the features that count numbers less than ten. In addition, there are some frequent words that do not make sense. We will use the stop words method later to filter these kinds of words.

Stemming algorithms work by cutting off the end or the beginning of the word, taking into account a list of common prefixes and suffixes that can be found in an inflected word. Lemmatization, on the other hand, takes into consideration the morphological analysis of the words. To do so, it is necessary to have detailed dictionaries which the algorithm can look through to link the form back to its lemma. (Reference1) We were ultimately not able to use stemming and lemmatization because our text was in an array. Stop words method helps us to filter unimportant or meaningless features. There is a package of pre-defined stop words. However, this package may miss some features that are specific for our case. To avoid this situation, we decided to remove some features by using the stop words package. And then we use TF-IDF to filter some features that have little weights. In addition, we look at our rest features and redefine our stopwords to remove meaningless ones.

Sentiment Analysis

Sentiment analysis is a cutting-edge technique used in natural language processing, text analysis, computational linguistics, and biometrics. Not only has a large number of implementation fields, but sentiment analysis can also help identify, extract, quantify, and study affective states and information.

Amazon customers share their experiences and thoughts about the products by leaving reviews. With the large volume of user generated opinions there is a need by Amazon or companies that have products on Amazon to analyze these reviews in order to improve customers satisfaction and drive better business decisions. The reviews customers were leaving does not contain a label for their sentiment (e.g.," positive" or "negative"). Lexicon-based sentiment analysis is the most popular solution when researchers are interested in customers' sentiment without labels. Lexicons contain many words, and the words are assigned scores for positive/negative sentiment. Some advanced lexicons possibly contain emotions like joy, anger, sadness, and so forth.

AFINN and VADER are two lexicons chosen in analyzing the sentiments of Cards Against Humanity reviews. AFINN was first developed by Finn Årup Nielsen. But the author of VADER is still unknown, and VADER stands for Valence Aware Dictionary and Sentiment Reasoner. Moreover, researchers always encounter a difficulty to choose an appropriate lexicon when there is no label for customers' sentiment. The best way to solve this issue is to try different lexicons and compare the results. If all lexicons have similar prediction results, it's safer to use either one to predict the sentiments of customers. Potential risk is minimized.

Analysis steps:

1. Randomly split data into training (209 observations) and testing (100 observations)
2. Fine-tune a best lexicon-based sentiment analyzer (Due to limited sample size, Fine-tuning is meaningless)
3. Using charts to help analyze

To help lexicons choose processes, we create dummy sentiment labels based on the overall ratings. If the overall rating is greater than 3 (4 or 5), the label is "positive". If the overall rating is not greater than 3 (3, 2 or 1), the label is "negative". If we try different thresholds ranging from -5 to 5 for AFINN-based scores and -1 to 1 for VADER-based compound scores, although we have the highest prediction accuracy, all sentiments will be labeled as "positive". All misclassified cases are False Positives. Then, we can conclude the fine-tuned model is not reliable. Besides, business decisions should be based on trusted models. Thus, all the two lexicon-based models have a threshold 0.

Now let us compare the confusion matrices (Figure2, Figure3). There is no substantial difference between the performance of two models, which implies choosing either one is acceptable. To further support this conclusion, actually positive and actually negative cases are plotted (Figure4, Figure5).

Supervised machine techniques like SVM have been tried to analyze the sentiment of dummy labels. However, its performance is disappointing, and SVM is based on the dummy labels generated from overall ratings, which may cause potential bias. Therefore, SVM is not considered.

VADER is finally chosen to label the sentiment of customers. The first reason it's chosen is that VADER is also good at dealing with slang. The second reason is that VADER can also analyze the sentiment from emotions and even punctuation. Finally, VADER is still updated. AFINN is developed early, and AFINN might be too old-school to have an accurate prediction on new texts.

Sentiment analysis results summary

1. **VADER**
   The Afinn lexicon based sentiment analysis prediction accuracy rate is 0.91. The recall sensitivity is calculated by true positives divided by true positives plus false negatives. This says what share of truly positive reviews in the data were identified as positive by the algorithm. Our recall sensitivity was 91/94 = 0.97. Precision is calculated by dividing true positives by true positives plus false positives which tells us what share of reviews identified as positive by the algorithm are truly positive. Our precision was 91/97= 0.94.

2. **Afinn**
   The Afinn sentiment analysis prediction accuracy rate is also 0.91. The recall sensitivity was also 91/94 = 0.97. The precision was also 91/97= 0.94.

Topic Modeling

Topic modeling is another method our group applied into this dataset in order to discover the abstract "topics" that occur in a collection of the text subject. Topic model is a type of statistical model used as a text-mining tool for discovery of hidden semantic structures in a text body.
For our project topic modeling could help us to better understand what customers are commenting on this particular product. The result of our topic modeling would give the company who produces Cards Against Humanity a better suggestion about how they could improve their product in order to attract more customers. Before we applied topic modeling into our dataset, we first did a sentiment analysis which divided the text variable into two groups, one included all the positive comments and another contained all the negative comments. Reviews with a score above 3 were deemed positive and reviews with a score 3 or lower were deemed negative.

Analysis steps:

1.Topic Model (via Latent Dirichlet Allocation) on the subsetted Toys and Games dataset (Cards Against Humanity reviews)
2.Visualize topics in a text corpus
3.Evaluate and discriminate between topic models

During the topic modeling process, we set doc_topic_prior(alpha) equal 0.25 and topic_word_prior(beta) also equal 0.25. Since alpha equals beta and both of them are less than 1 which means the sample means and variance would also be close between alpha and beta. And at the beginning we choose to set n_componets equal to 8 which will return 8 topics. However, the topics we had are duplicated and included some unimportant information and 8 topics is too much to name them at the same time to keep each of them unique. So we finally decided to use 4 topics for the positive reviews and 3 topics for the negative reviews.

Topic Modeling results summary

Even after splitting up the reviews into a positive and negative group, many common words were still similar from topic to topic in both the positive and negative groups. It seems like people say many of the same things about Cards Against Humanity in their reviews. For the positive group (as Figure 5 shows) the first topic is "Comedy", the second topic is "Family", the third topic is "Party" and the fourth topic is "Rule Specific". We name each topic by looking at the top 10 most salient terms for each individual topic. On the other hand, for the negative review group (Figure 6) we defined those three topics using the same process. We named the first topic as "Apples to Apples Fans", the second topic as "Unsatisfied", and the third topic as "Offended ".

Conclusion & Future strategies

Improvement of data collection

One of the main shortcomings of our analysis is that there are only a small number of negative reviews and an even smaller amount in the test sample. Another shortcoming of the models could be possible overfitting.

Recommendation:

1. Give customers rewards if they leave a review. Then, we could avoid the shortcomings discussed above if we have more reviews. More reviews improves prediction accuracy as well as analysis correctness.

Expand advantages and avoid disadvantages

Based on the topic modeling result, we believe tha Cards Against Humanity is doing a pretty good job positioning their product and targeting the right customers because people generally have the same comments about the game whether the review is positive or negative. Cards Against Humanity clearly has many devoted fans, people frequently said “fun”, “funny”, “love" and “party game”. But, Cards Against Humanity should also be aware of the negative comments that people leave about the game - namely that it is “expensive”, “rude”, problematic and “dull”. Also, the words “offend” and “offensive” came up a few times in both the positive and negative reviews so clearly some people like the inappropriate nature of the game while other people are opposed/ offended by it.

Recommendations:

1. Make a video focusing on the offensive element of Cards Against Humanity, and put the video on the product page.
2. Make an announcement to customers. Let customers know they can win discounts if leaving a review.

Compete against competitors

Another important insight from our analysis is that many people said “apple” in their comments which is probably a reference to the game Apples to Apples which is a card game that Cards Against Humanity competes directly against. It seems like some people prefer Apples to Apples over Cards Against Humanity because Apples to Apples is less inappropriate and therefore more kid friendly.

Recommendations:

1. Design some new similar products like expansion packs or update the game with a new version of Cards Against Humanity based on the topic modeling. For instance the company could release a party pack for Cards Against Humanity with more mature or funny cards.
2. Make the cards waterproof or glow in the dark which could better satisfy some of their customers.
3. Make a pack that is more PG and targeted for families with kids that is more age appropriate. This could help them compete with Apples to Apples.

Ranking improvement on the search engine (e.g, Amazon, Google)

Another suggestion would be for the company who designed Cards Against Humanity to add keywords such as “party game” and “fun” into their product description on Amazon so their product would be recommended to customers who search those keywords on the browser.

Recommendation:

1. Add customers' most-used words (e.g,“party game”, “fun”) into the title of the products.

Appendix

Link to data: http://jmcauley.ucsd.edu/data/amazon/
http://snap.stanford.edu/data/amazon/productGraph/categoryFiles/reviews_Toys_and_Games_5.json.gz

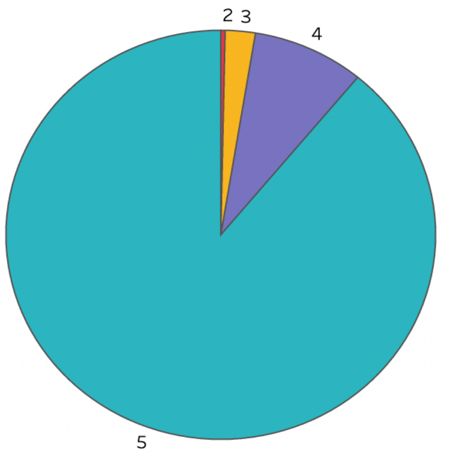

Figure 1: Pie chart of overall ratings of Cards Against Humanity

Predicted:	negative	positive	All
True:			
negative	0	3	3
positive	6	91	97
All	6	94	100

Figure 2: AFINN confusion matrix

Predicted:	negative	positive	All
True:			
negative	0	3	3
positive	6	91	97
All	6	94	100

Figure 3: VADER confusion matrix

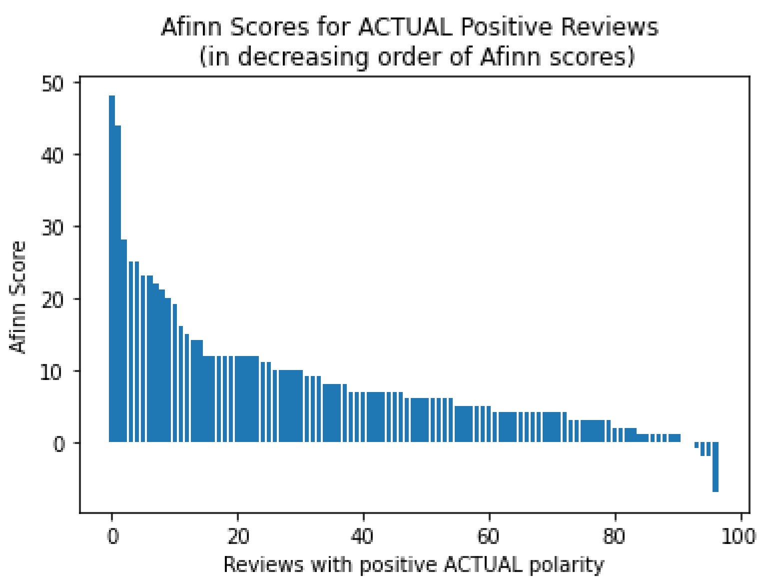

Figure 4: AFINN scores for actual positive reviews

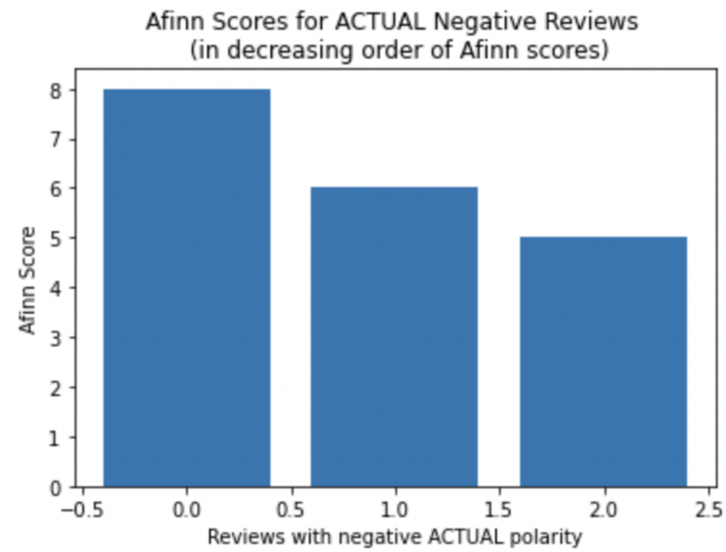

Figure 5: AFINN scores for actual negative reviews

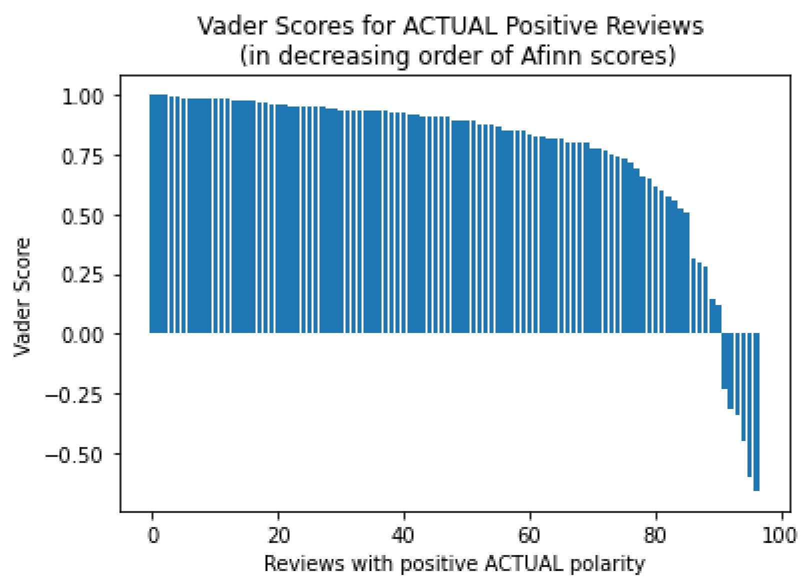

Figure4: VADER scores for actual positive reviews

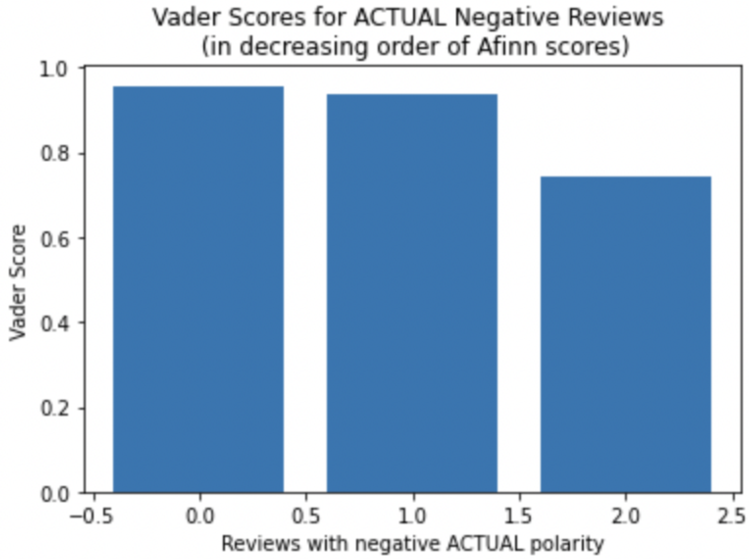

Figure 4: VADER scores for actual negative reviews

Figure 5: Positive topic modeling result for topic 1

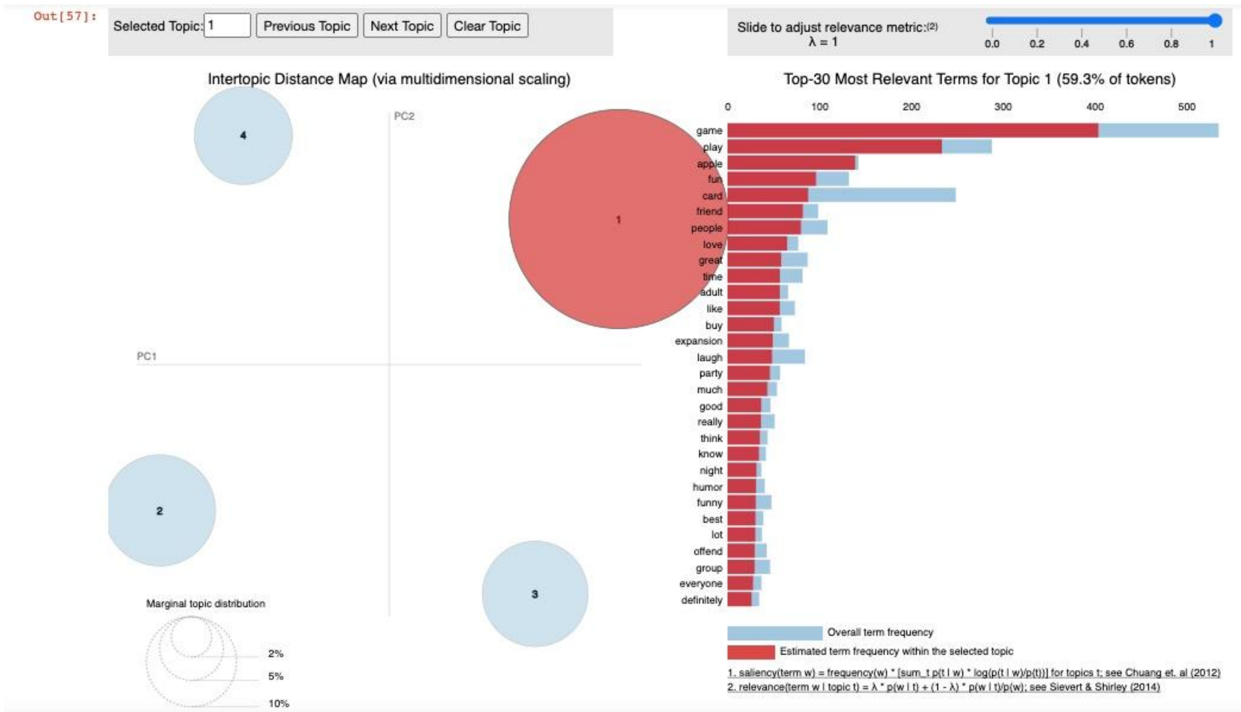

Figure 5: Positive topic modeling result for topic 1

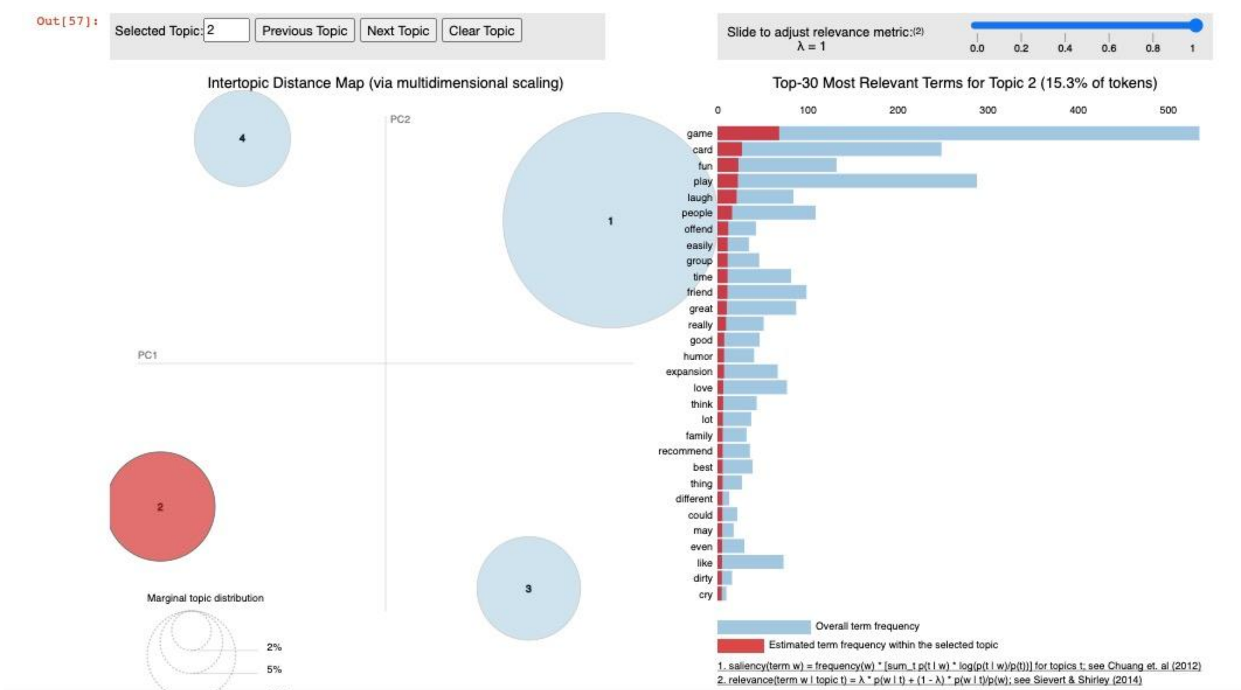

Figure 5: Positive topic modeling result for topic 2

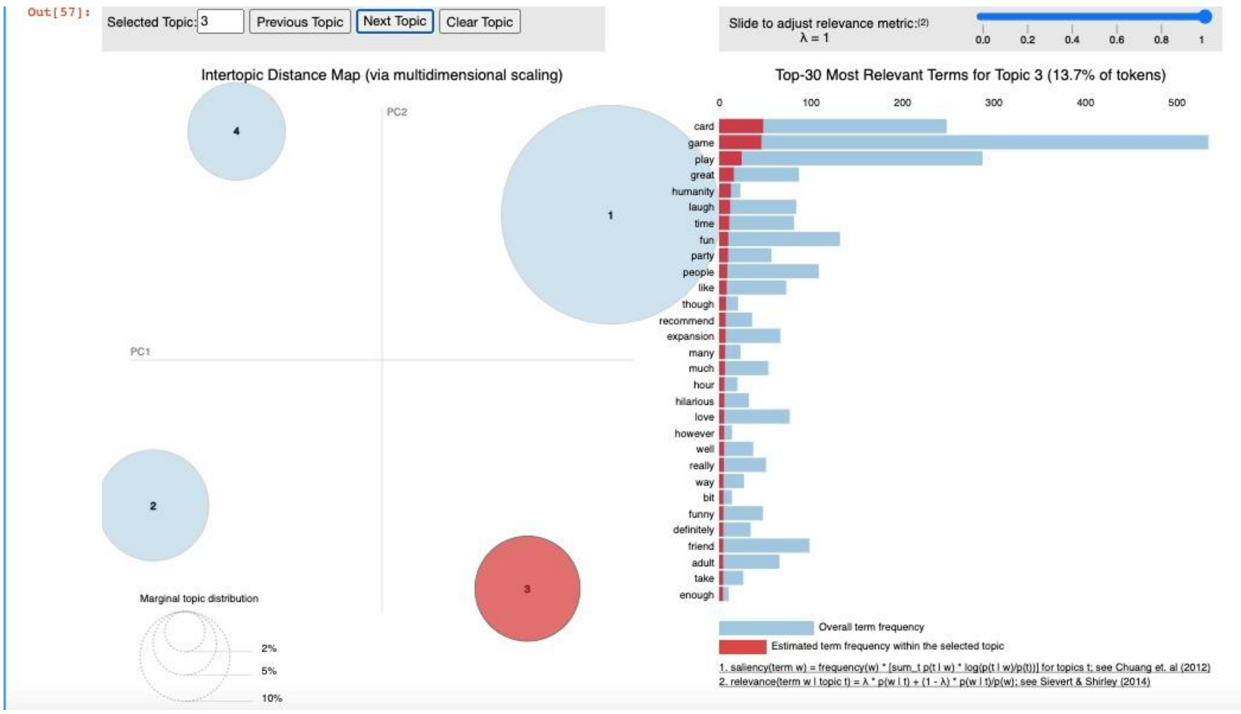

Figure 5: Positive topic modeling result for topic 3

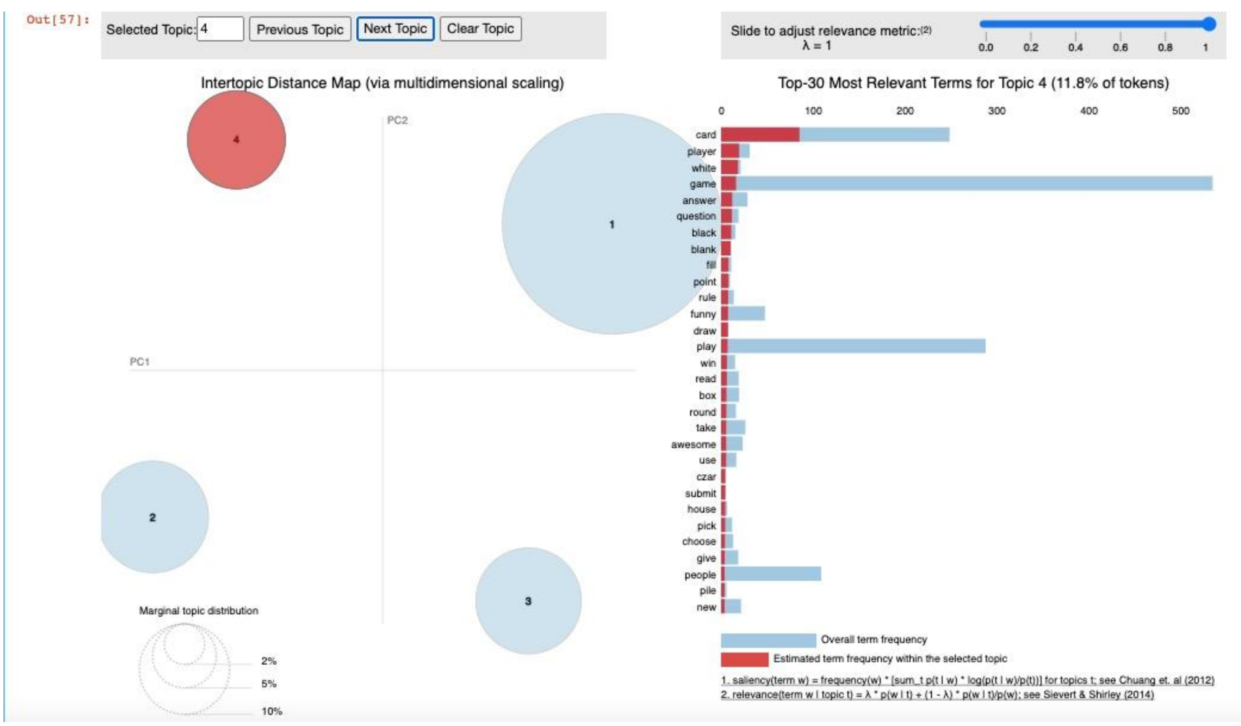

Figure 5: Positive topic modeling result for topic 4

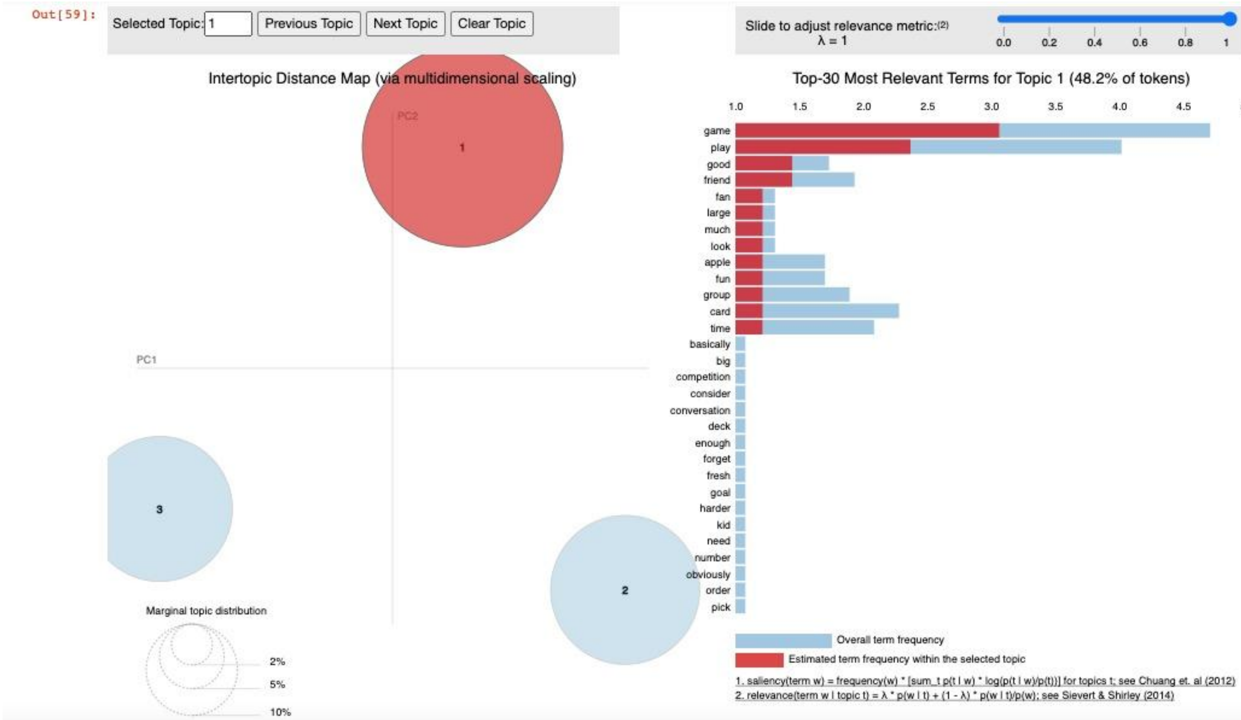

Figure 6: Negative topic modeling result for topic 1

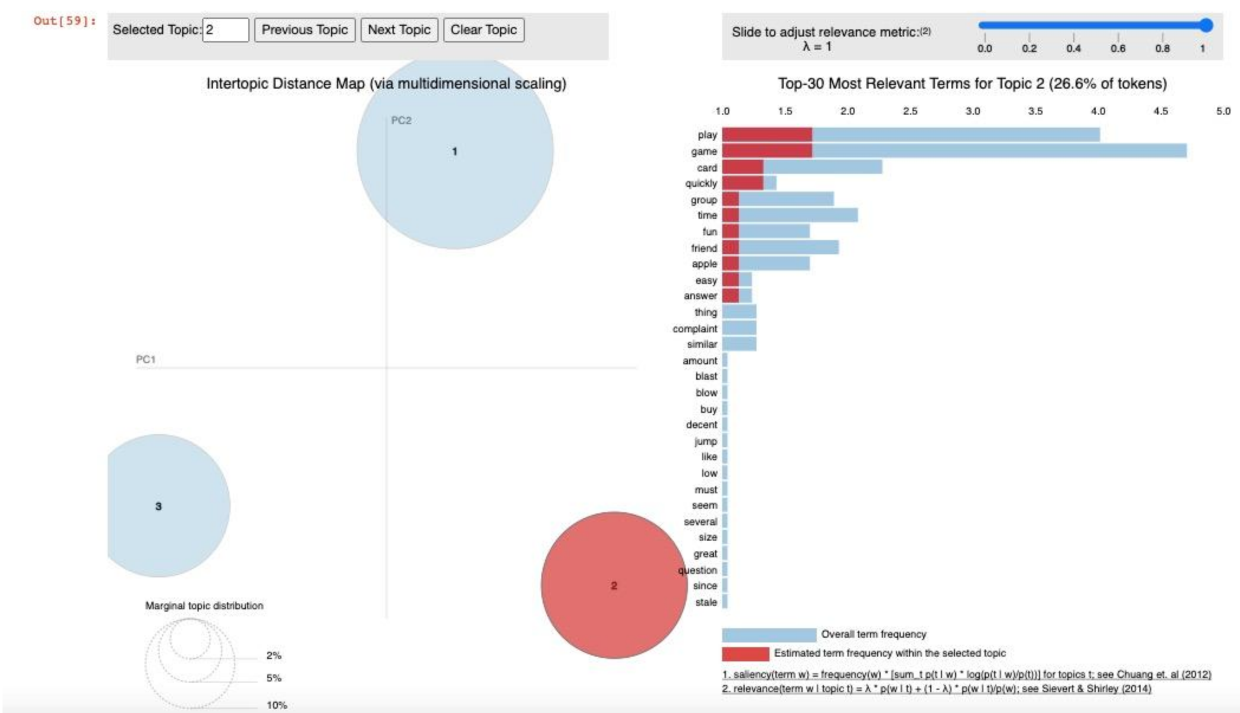

Figure 6: Negative topic modeling result for topic 2

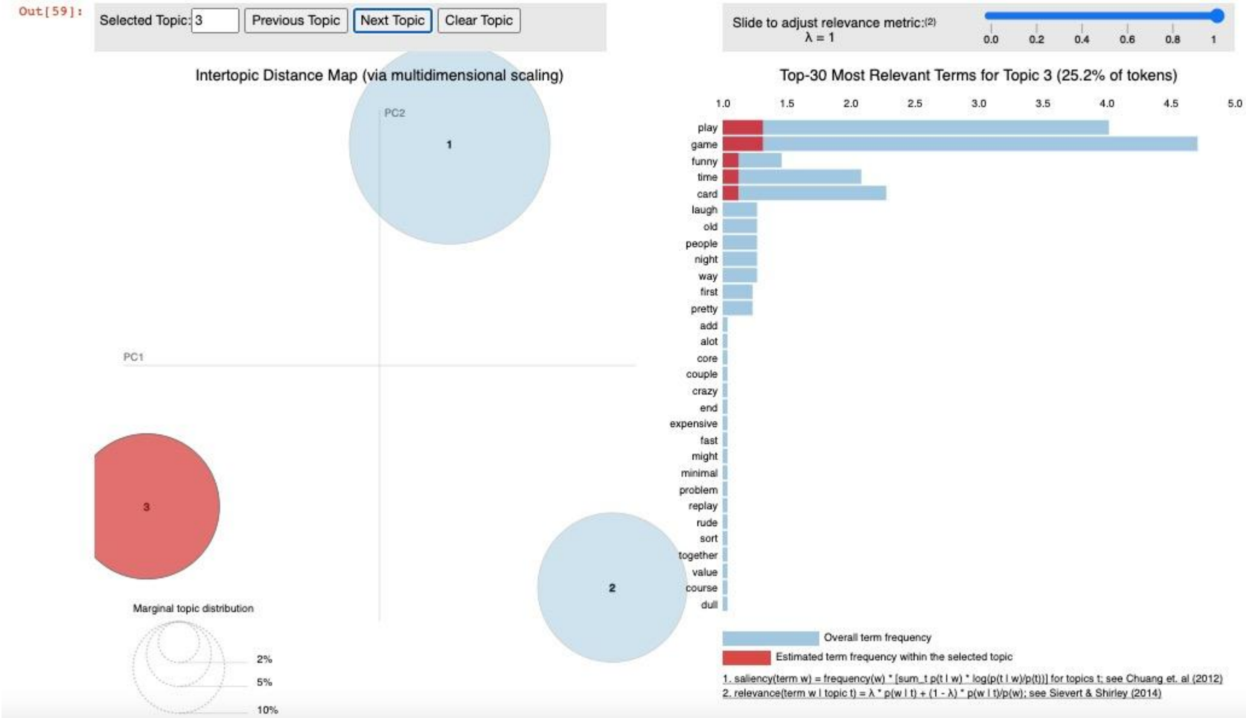

Figure 6: Negative topic modeling result for topic 3

References

Reference 1:

Bitext. (n.d.). What is the difference between stemming and lemmatization? https://blog.bitext.com/what-is-the-difference-between-stemming-and-lemmatization/.