Ethan
3/18/21
NLP Semester Project

Amazon Reviews: A Foray into Fine Foods

Abstract

In this analysis, we explore a data set of Amazon Fine Food Reviews from October 1999 to October 2012. Given the extensive nature of the dataset, one can't help but wonder if the language used in online reviews has changed over time, especially on a platform as prevalent as Amazon. Changes in technology and the eventual rise of smartphones coincide with the time period of the data, and we hypothesize that the technological and culture impact of technology can also be seen through the language used on online platforms. This is to say, shorthand used in smartphone communication may have a negative impact on the lexical complexity of the reviews. The subset of the data used in this analysis specifically focused on "coffee" and "tea" reviews from 2007 to 2012 as these categories were the most popular within the dataset. The main methods used in this paper are TTR and R to measure lexical diversity, as well as tf-idf and cosine similarity to measure how similar reviews are between years. After looking at the results, we fail to reject our null hypothesis and conclude that for the "coffee" and "tea" reviews, there is no evidence of change in language over time. Across years, language used in reviews were found to be extremely similar which was quite surprising. It is worth noting however, that average review length somewhat increased between 2007 and 2012 for both categories which could be another possible avenue of interest.

Introduction

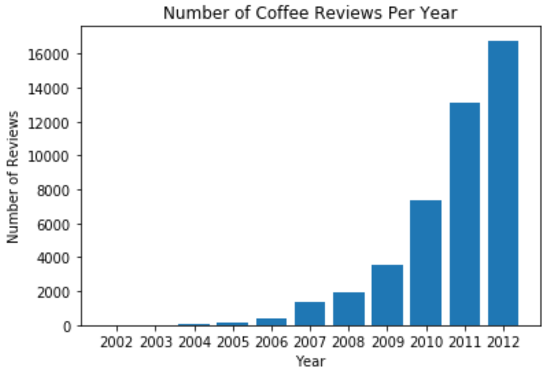
The dataset used in this analysis can be found here: https://www.kaggle.com/snap/amazon-fine-food-reviews. The early 2000's was a pivotal period in technology and culture. The internet was beginning to become more connected and more and more individuals adopted the use of mobile phones. With the introduction of smartphones at the end of the 2000's, shorthand became more commonly accepted as a means of communication. Because this dataset contains a collection of reviews throughout the 2000's, it's possible that some of the shorthand that was becoming popular on phones has translated to the online review space. We hypothesize that lexical diversity in Amazon reviews has decreased over time.

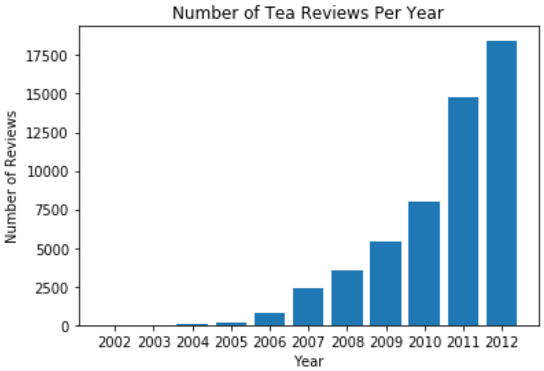
Similar work has been done by Michael Hine in his paper, "The Effects of Text Complexity on Online Review Helpfulness[1]", which looks at lexical diversity in Amazon reviews to determine helpfulness. In regards to lexical diversity, Hine rejects his hypothesis that "an increase in online review lexical diversity will decrease review helpfulness." In fact, lexical diversity was the only measure that predicted online review helpfulness. In this paper, instead of TTR, estimation algorithms were used to calculate lexical diversity. While we don't have access to such tools in this analysis, lexical diversity has shown to be an important metric in the space of Amazon reviews.

[1] Hine, Michael J. (2014) "The Effects of Text Complexity on Online Review Helpfulness," Communications of the IIMA: Vol. 14 : Iss. 1 , Article 3. https://scholarworks.lib.csusb.edu/cgi/viewcontent.cgi?article=1353&context=ciima

Methods

An analysis of the entire dataset was performed before focusing on specific categories within fine foods. This analysis yielded that the most popular terms overall in the dataset were “tea” and “coffee.” From this, we can infer that most of the reviews focused on these two categories. This would make sense given that these reviews are from the “fine food” section of Amazon. Since product names were not provided, it was decided that tracking these two categories would be the best course of action. Below, we can see the number of reviews per year for each category.

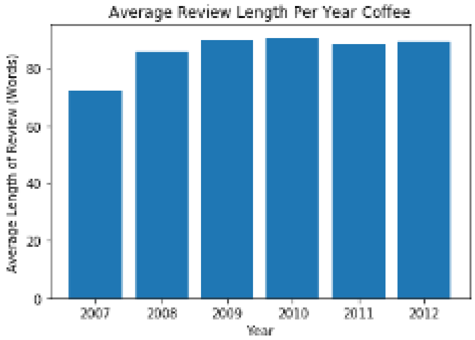


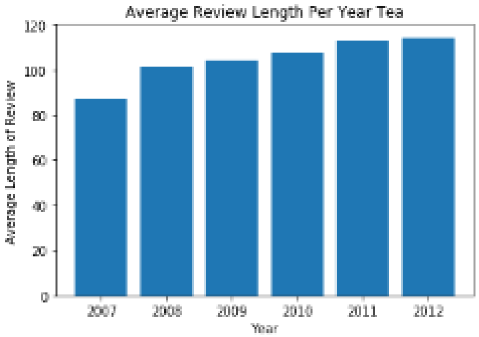


Since the number of reviews per year before 2007 were extremely low, this analysis focused only on the reviews present between 2007 and 2012 to avoid having skewed data. After being split per year, the reviews had duplicates removed as well as html formatting and punctuation removed as part of the cleaning. The reviews were also changed to lower case since case-sensitivity was not considered important here. Punctuation was decidedly not going to add any additional information to the analysis. Stopwords from the nltk package were removed after tokenizing. This list of stopwords was chosen due to its robustness and the fact that it is not as penalizing as other stopword libraries. Since Amazon reviews are generally somewhat brief, it would not be ideal to heavily remove words. Stemming was also avoided since it was not deemed necessary for this analysis. The only additional stopwords added were either “tea” or “coffee” since the dataset was already subsetted on these words as well as some additional spam words that were relics of formatting. The word “not” was also removed from the stopword list to account for negative sentiment. Both unigrams and bigrams were included in the analysis to make sure that words such as brand names and negative sentiment were captured. To check the most frequent words, tf-idf scoring was used as a more robust method of checking commonly occurring words. The methods of comparing lexical diversity were TTR and R. While TTR was the original metric chosen for this analysis, R was also included since it factored in the number of tokens present, making it more robust. Cosine similarity was chosen as the main metric for comparison also due to its robustness and usefulness in comparing texts, relative to other measures of distance. In the calculations of TTR, R, and cosine similarity, all the reviews for each year were joined into one large document for the year and then the calculations were performed on each per-year text.

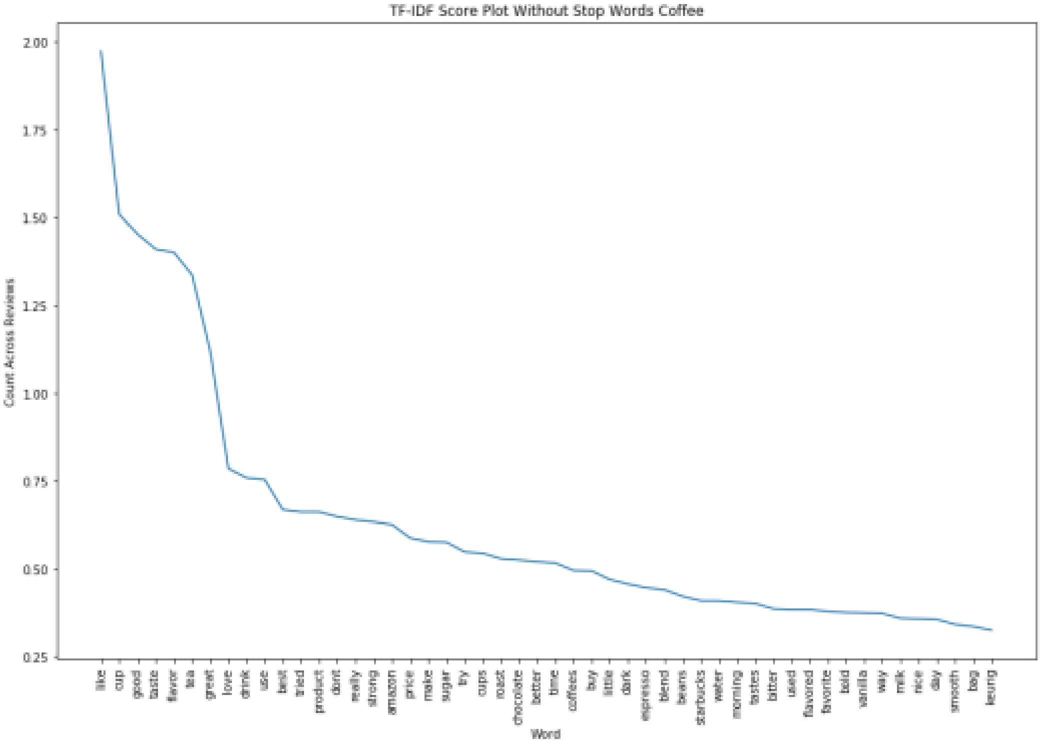
Results

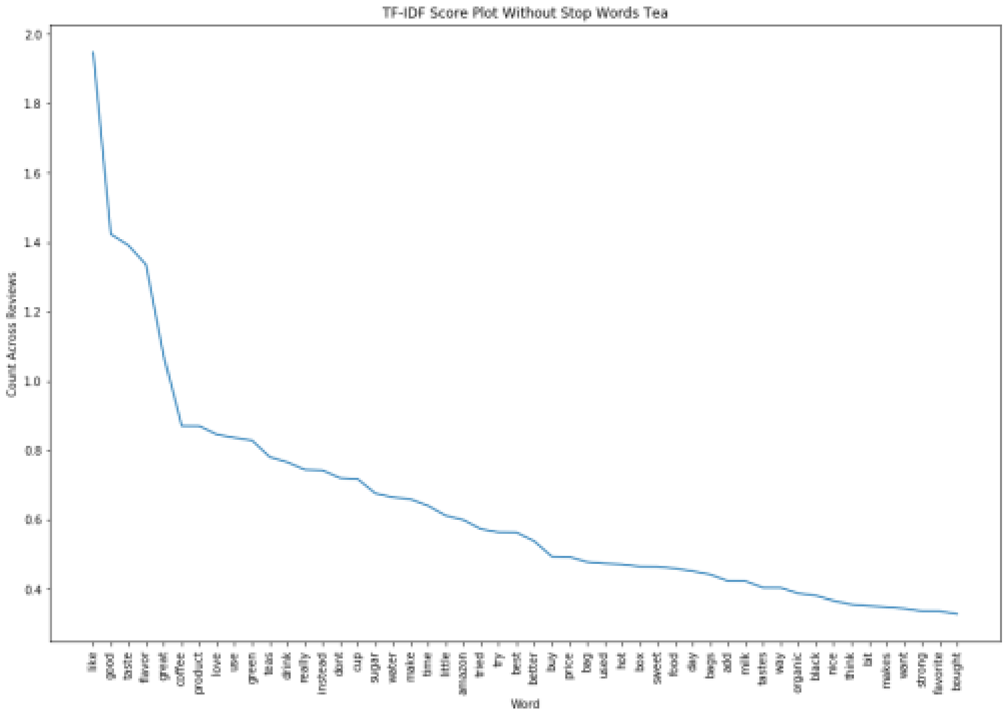
As part of the preliminary analysis, mean review length was calculated. The plots are below:



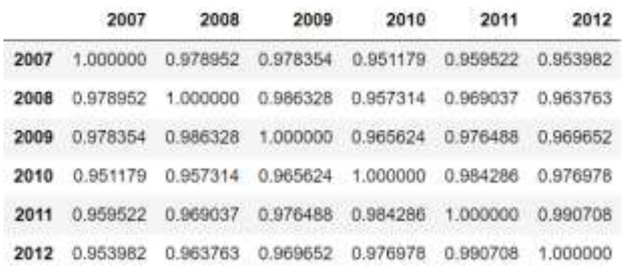


Here we can see a slight increase in mean review length over the years. One possible explanation for this would be the fact that the number of reviews were increasing over the years. After the dtm was created, tf-idf scores across all years was calculated to check the most commonly used words. As expected, many of the words were in relation to both tea and coffee, including words like "cup", "flavor", and "starbucks".

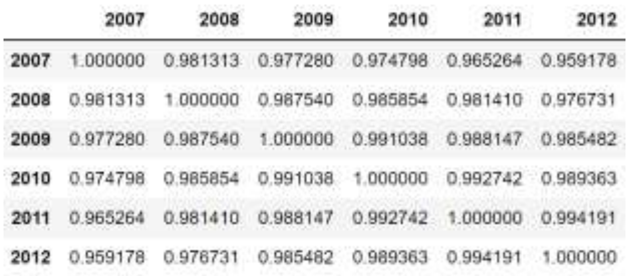




We see that across both categories, the most common words stay relatively constant. Cosine similarity was calculated next. Coffee is present on the left while tea is on the right:

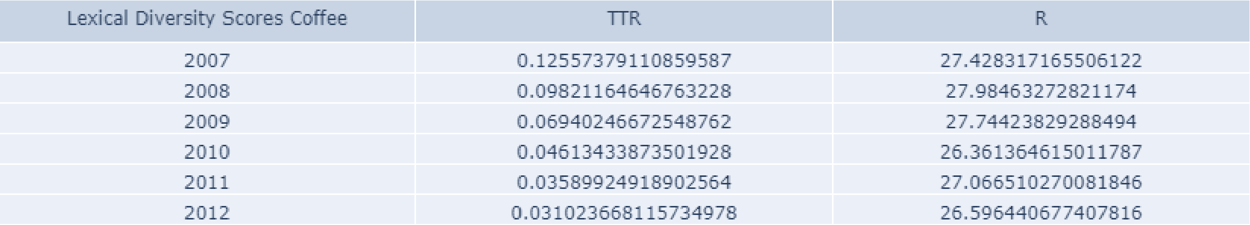


	2007	2008	2009	2010	2011	2012
2007	1.000000	0.978952	0.978354	0.951179	0.959522	0.953982
2008	0.978952	1.000000	0.986328	0.957314	0.969037	0.963763
2009	0.978354	0.986328	1.000000	0.965624	0.976488	0.969652
2010	0.951179	0.957314	0.965624	1.000000	0.984286	0.976978
2011	0.959522	0.969037	0.976488	0.984286	1.000000	0.990708
2012	0.953982	0.963763	0.969652	0.976978	0.990708	1.000000

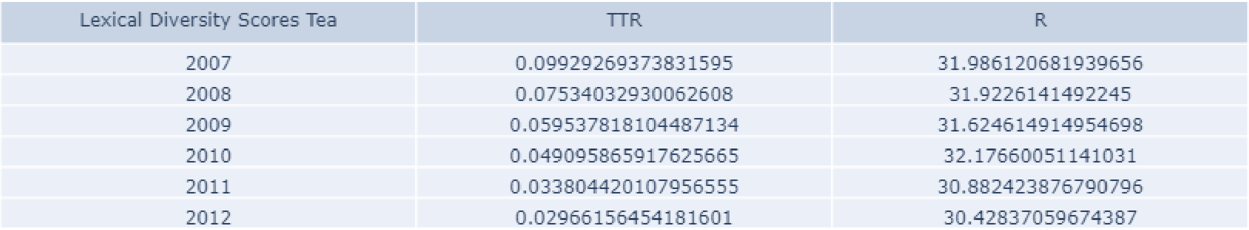


	2007	2008	2009	2010	2011	2012
2007	1.000000	0.981313	0.977280	0.974798	0.965264	0.959178
2008	0.981313	1.000000	0.987540	0.985854	0.981410	0.976731
2009	0.977280	0.987540	1.000000	0.991038	0.988147	0.985482
2010	0.974798	0.985854	0.991038	1.000000	0.992742	0.989363
2011	0.965264	0.981410	0.988147	0.992742	1.000000	0.994191
2012	0.959178	0.976731	0.985482	0.989363	0.994191	1.000000

Across years we see that the reviews overall are quite similar. Even between 2007 and 2012 the score is still .959 which is very high. Finally, we check the TTR and R scores of the reviews between years. Looking at the TTR scores for both categories, we see a decrease in lexical diversity, however, TTR does not take into account token length which will clearly be higher for later years since there are more reviews. The R score on the other hand, which penalizes number of tokens, yields results that show no difference in lexical diversity in between years.



Lexical Diversity Scores Coffee	TTR	R
2007	0.12557379110859587	27.428317165506122
2008	0.09821164646763228	27.98463272821174
2009	0.06940246672548762	27.74423829288494
2010	0.04613433873501928	26.361364615011787
2011	0.03589924918902564	27.066510270081846
2012	0.031023668115734978	26.596440677407816



Lexical Diversity Scores Tea	TTR	R
2007	0.09929269373831595	31.986120681939656
2008	0.07534032930062608	31.9226141492245
2009	0.059537818104487134	31.624614914954698
2010	0.049095865917625665	32.17660051141031
2011	0.033804420107956555	30.882423876790796
2012	0.02966156454181601	30.42837059674387

Here we fail to reject the null hypothesis of difference in lexical diversity between years. One possible explanation for this would be that the coffee and tea are not categories that would have a large change in the demographic of consumers and overall as a category, it is pretty stagnant. Categories that appeal to younger generations such as electronics, gaming, and popular culture items like toys might be areas where we could observe a greater change over time in lexical diversity.

Discussion

In conclusion, we were not able to find a meaningful change in lexical diversity between the years 2007-2012 within our subsetted data. While somewhat surprising, this would fall somewhat in line with the results of Hine. Since lexical diversity is representative of review helpfulness, a decrease in lexical diversity over time would also correlate with a decrease in review helpfulness over time which would not make a lot of sense logically, since a heavy decrease in useful reviews would be indicative of other problems such as bots. Interestingly, the average length of review increased between 2007 and 2012 which could be due to the number of reviews but other factors could also be affecting this.

One of the clear limitations of this analysis was the fact that data before 2007 was not usable due to the scarcity of reviews during those years. It would have been nice to compare the results from even earlier in the 2000's since technology was not as prevalent then as it was by the end of the dataset. Another limitation of the analysis was the lack of product names. With a specific product it would be much easier to track reviews over time instead of filtering reviews only containing a specific key word.

One of the reviewers suggested running an analysis on the extremely extensive dataset found here: https://nijianmo.github.io/amazon/index.html. I think this could easily be combined with the data used in this analysis to create a more extensive time series analysis. It might also be worth looking into measures of lexical complexity, although this would be quite difficult to gauge given simple subject matter such as Amazon reviews. Finally, some level of sentiment analysis could be done by implementing Hu & Liu's sentiment lexicon. This would be particularly useful in determining changes in customer sentiment for specific products over time.