Cookbook Consumer Behavior Analysis from Amazon.com

Xing Fang, Yating Liu, Zi-Qi Liu, Tianrui Wang
Instructor: Professor Rong Liu

MOTIVATIONS AND OBJECTIVES

What key elements made up those best sellers in the book industry is one of the most concerned topics for publishers. For long, customer reviews have been a doorway to analyzing consumer behavior and buying patterns for product introduction and improvement. In this web mining project, we analyzed different aspects of the best-selling cookbooks on Amazon.com to obtain consumer preferences and patterns and help the publishers in their book production lifecycles.

INTRODUCTION

This project used unsupervised & supervised approaches, analyzed different aspects of the best-selling cookbooks on Amazon.com, derived valuable consumer behavior patterns, obtained consumer preferences and patterns to help the publishers in their book production lifecycles.

LITERATURE REVIEWS

A lot of research has been done in analyzing amazon product reviews and ratings. After reviewing a few pieces of literature, we summarized some key takeaways on this research topic. Feature extraction could be one of the difficulties to crack, as Sohial, Siddiqui & Ali (2016) implied that negative words may be used in favor of a product. They introduced human intelligence to categorize the features, which could be a reference in the working process of our project. Another difficulty could come from the Sentiment Analysis, as we want to extract positive and negative features of the books by analyzing the sentiment of the reviews. Currently, most research on Sentiment Analysis focus on classification, for example, the sentiment analysis was done by Srujan, et. al. (2018) classified the amazon book reviews into 7 categories of emotions.

DATA

Web Scraping

We used Selenium Webdriver to scrape the data of 1,095 books from 'Cookbooks, Food & Wine' category on Amazon.com. The data was scraped into two parts:

1. Title, Author, Star, Review Amounts, Price, Intro, Keywords
2. Reviews – Title, Star, Review Title, Comment, Sentiment by Amazon

Exploratory Data Analysis

We used words from the book title to generate the following WordCloud. It is clear that the top3 used words are “diet”, “guide”, and “healthy”.

Figure 1 WordCloud of Title

We also plotted frequency charts for the words in book titles and Top words: Title and Intro are similar

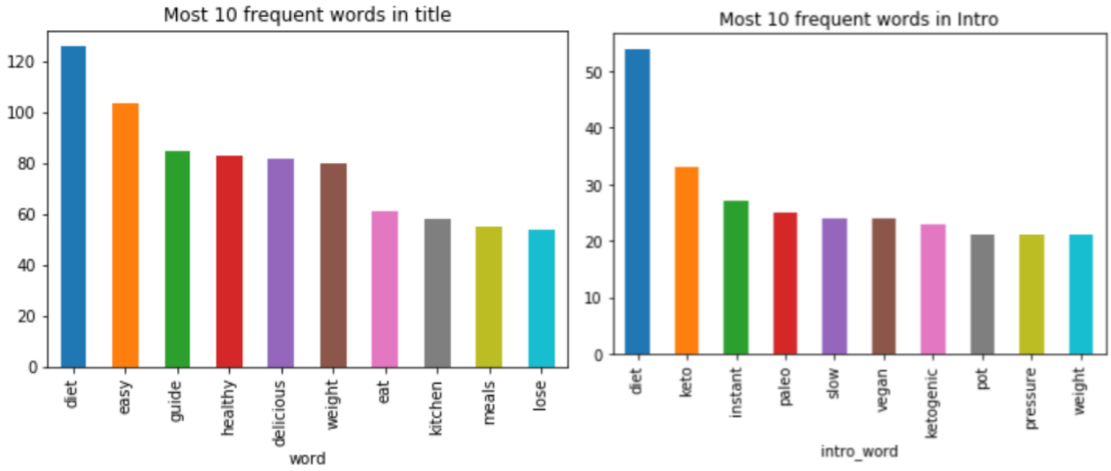

Figure 2 Comparison of Frequent Words in Title and Intro

From the correlation matrix we could see that positive correlations include: (hardcover_price, star), (kindle_price, star), (hardcover_price, kindle_price); and negative correlations include: (book_rank, star), (book_rank,review_amount).

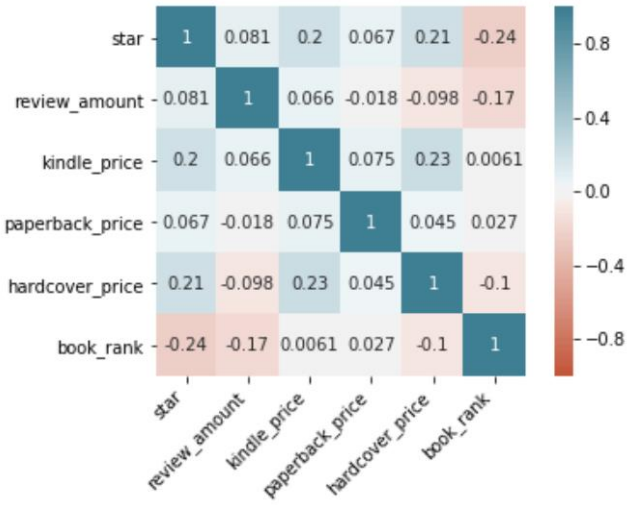

Figure 3 Correlation Plot of Related Variables

After reviewing 10,685 reviews, we find out Amazon labels the sentiment based on the ratings of star. Not we expectation Amazon will use the review content to label the sentiment.

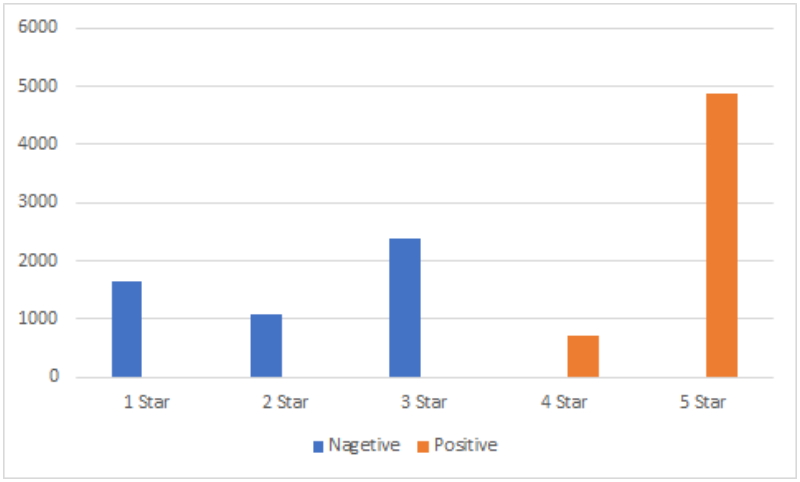

Figure 4 Star Distribution in Positive and Negative Reviews

METHODOLOGY

We approached our problem from two ways, using both unsupervised learning and supervised learning. On the one hand, we implemented Clustering Analysis with K-Means and Latent Dirichlet Allocation (LDA) to divide the large corpus of information on the book as well as on the reviews. On the other hand, we did sentiment analysis based on the reviews and their sentiment labels generated by Amazon.com.

Clustering Analysis

K-Means Clustering Based on Introduction + description

Firstly, we used stop words to filter meaningless words like cooking, food, recipes and something like that. Secondly we used silhouette scores to determine the optimal number of clusters. In silhouette scores Y means the mean distance to the other instances in the same cluster, X depicts mean distance to the instances of the next closest cluster. So with higher scores, the clustering results would be better. Based on our plot, we can see with increment of more number of clusters, effect of clustering will be better. However with high number of clusters, the metric is computation expensive as the coefficient is calculated for every instance. Considered with more numbers, the features of each cluster is not so obvious. Finally we chose local optima 3.

$$Silhouette\ Coefficient = (x\text{-}y)/\ max(x,y)$$

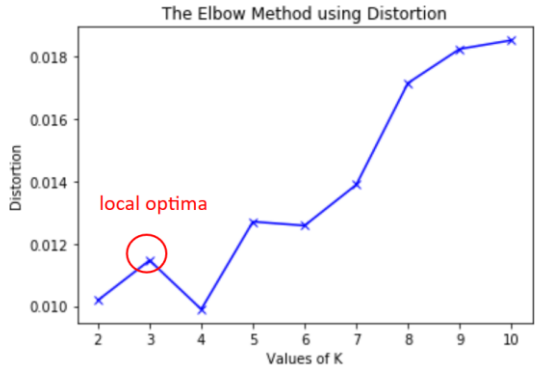

Figure 5 Silhouette Score for optimal number Clustering

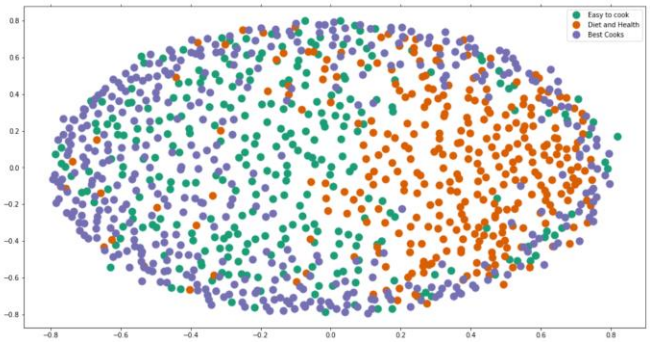

Figure 6 Multidimensional Scaling of

We use the top words of each cluster, conclude cluster 0 is Easy to cook; cluster 1 is Diet and Health; cluster 2 is Popular Chefs.

Figure 7 WordCloud of Each Cluster's Title

Topic Modeling on Book Reviews

In order to cluster the reviews into topics and derive meaningful insights from the corpus, we used Latent Dirichlet Allocation (LDA), which is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. Both sklearn and gensim packages are used in the modeling process. LatentDirichletAllocation & gensim.models.ldamodel.LdaModel to do Topic Modeling. We used a total of 6000+ reviews with an average of 50 reviews per book.

In Used GridSearchCV to search for the best parameter for n_component:
Although the result showed that 2 components result in the best Log Likelihood & Perplexity
We decide to use n = 3 for convenience and consistency with the K-Means Clustering.

```
Best Model's Params:  {'n_components': 2}
Best Log Likelihood Score:  -211628.0623783079
# Log Likelyhood: Higher the better
Model Perplexity:  1102.8682382842942
# Perplexity: Lower the better. Perplexity = exp(-1. * log-likelihood per word
```

Figure 8 Gridsearch Results for the LDA Model

After deciding the parameters for the LDA model, we trained two models for positive and negative reviews to see if there are major differences in the most frequent words. Below are the results.

	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10	Word 11	Word 12	Word 13	Word 14
Topic 0	good	make	use	ingredient	just	look	time	great	buy	try	lot	star	page	really	think
Topic 1	read	good	write	story	just	buy	know	author	say	information	life	really	want	think	people
Topic 2	wheat	diet	eat	alternative	non	just	product	bread	good	try	free	plan	day	weight	make

	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10	Word 11	Word 12	Word 13	Word 14
Topic 0	diet	eat	weight	lose	good	day	work	lot	plan	healthy	pound	make	just	week	fat
Topic 1	read	life	good	eat	feel	make	diet	love	year	change	just	way	time	really	know
Topic 2	make	good	great	love	use	easy	try	just	buy	ingredient	time	bread	delicious	keto	look

Figure 9 Top 15 keywords for each topic

It seems useless to separate positive reviews from negative reviews and train two topic models on top of each, since top words for negative reviews also includes a lot of positive words. Therefore, we trained another LDA model on all the reviews using gensim. The results are shown below.

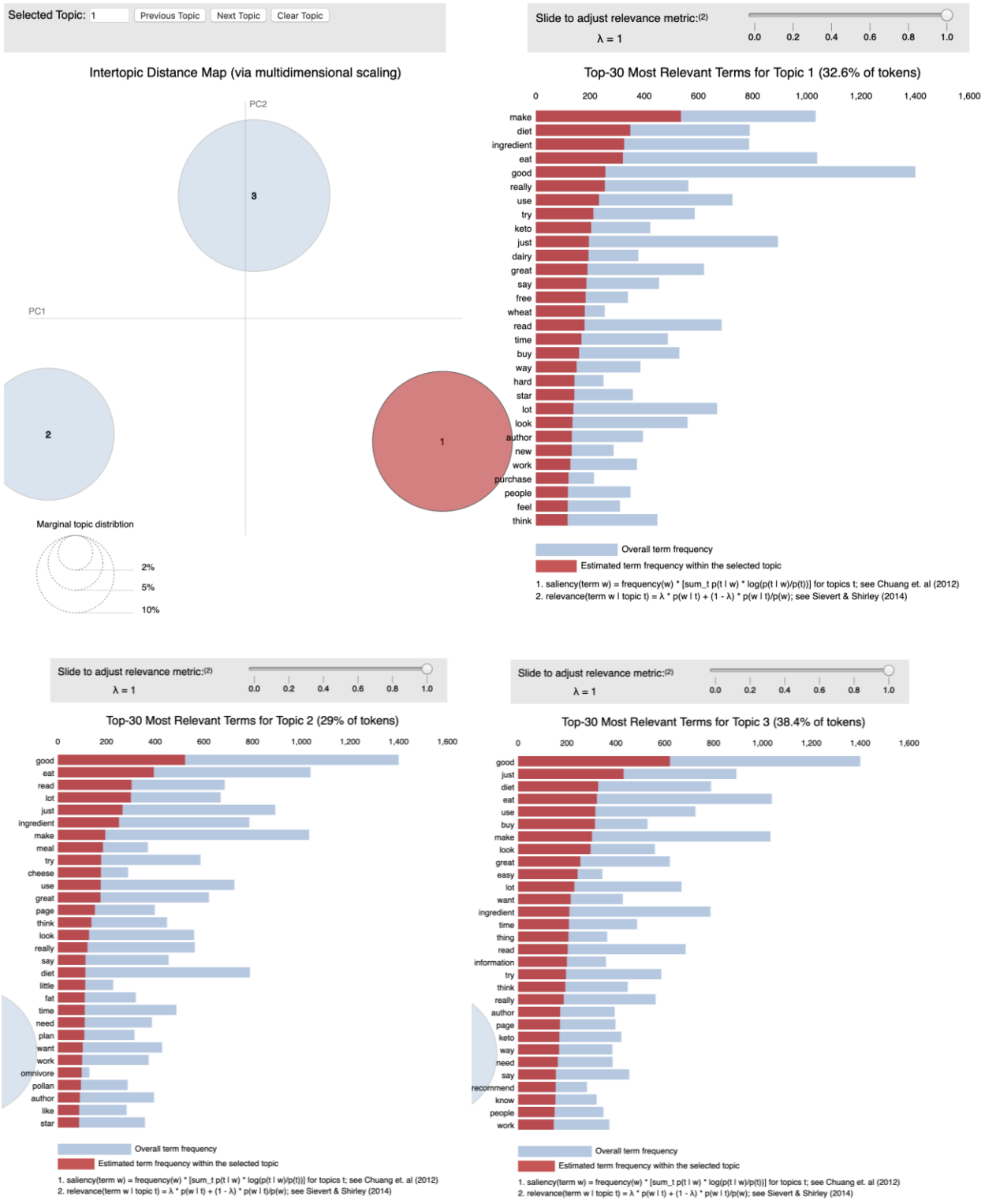

Figure 10 Topic Modeling Results using Gensim

Scanning through the top-30 most relevant terms for all the topics, we could observe that Probably not a good idea to do Topic Modeling directly on all reviews, as reviews on different books will be mixed together.

Sentiment Analysis

Base on clusters - Naive

	0: Easy-to-Cook	1: Diet & Health	2: Famous Chefs	The Whole Dataset
Sentiment Accuracy	0.70	0.81	0.65	0.69

Sheet 1 Accuracy Score of Naive Sentiment Analysis

Initially, we did a naive sentiment analysis, which counts positive words and negative words after tokenization, then we determined the sentiment of reviews and compare the result with labels defined by Amazon. To evaluate the model, we test the accuracy with 3 clusters and the whole dataset. As you can see, the cluster of Diet and Health got a best accuracy as 0.81 while the accuracy of the whole dataset just as 0.69.

Base on whole dataset - Absolute Proportional Difference

threshold	0.0001	0.001	0.003	0.004	0.005	0.006	0.009	0.01
accuracy	68.53%	68.58%	68.61%	69.37%	68.78%	68.48%	66.79%	65.6%

Sheet 2 Accuracy Score of Each Threshold

We tried a different way to test the sentiment which is absolute proportional difference. Sentiment is calculated as Sentiment = (P − N) / (P + N + 0). During the process, the difficult part is determining the best threshold. We tried several thresholds as above, the highest accuracy is 69.37% with threshold as 0.004. The accuracy is very similar to accuracy of naive and. So we thought that Amazon sentiment tags may not be as accurate as we expected.

Compare label difference and manually relabel confused reviews

In this step, we compared results based on naive method with Amazon's results. And check the different labels, manually relabelled the confused reviews, remove reviews that are difficult to differentiate.

By checking the results which have different sentiment tags, I found Amazon's sentiment tags is not so accurate, like below reviews. Some people think it's a good book, but some parts of it are

not so good; or receipe is good, but not fit for them. All in all, the book may not fulfill their expectations, but it is still one good book.¶

Based on our report insights, we want to use sentiment analysis to evaluate one book. So although those reviews are tagged as negative, they finally agree that it is a good book. Then that kind reviews should be labeled as positive which can finally improve the results of book evaluations

#Yet another recipe book. We have thousands of them, and this is just another one. Is it useful? Yes, of course it is, but it's nothing special in the scheme of things.
#It's a pretty good book, there are some ingredients in it that are hard to find but other than that pretty awesome.
#Would have been nice if there were photos.
#Some good recipes but uses Splenda too much.
#I like the smoothie recipes, but my husband and I didn't stick on the cleanse past 3 days. I can't drink my meals every day. He also found them to not be as tasty (not very sweet, but he's a sweet tooth) which didn't help our motivation much. I think she gives some good principles, but this cleanse was not for us. Also, I would have liked to see more scientific-based information or at least references for some of the health information she provides.
#As a seasoned cook, I thought this would be a good resource to have to help our family avoid processed food. However, I found the recipes extremely basic for me. I felt I got better recipes from the authors website.
#I bought this because it was super discounted. I vaguely follow the blog. Author has a sensible moderate approach to food with no crazy out there ingredients. Good basic recipes inside, easy to make, but nothing I haven't seen elsewhere or tweaked my self. I don't have kids but the school lunch ideas were probably the most useful as I often get bored with the food I pack for work. Was worth it for the price I paid.
#there are parts I love about this book such as part one. Which talks about how to make changes and shopping for whole real food items but as far as the recipe portion I was not at all impressed. I love love love to cook and bake for my family but I could honestly have found these on Pinterest or created these myself. All in all, I gave three stars for the whole front half of the book because there was some great info there.
#Enjoy and gave as gift however nothing new.
#Good ideas and intentions, but the recipes are a little too basic.
#Not as I expected, but still good.

Below reviews are those obviously positive, but tagged as negative by Amazon.

#Pretty good

#Useful jar recipes of seasonal gifts for friends and family... The soups are worth the price of admission alone!
#This book has some good tips for how to pack and package homemade mixes in jars. If you're looking for creative ways to give cookies for the holidays, check out this book.
#It was ok. Wasn't too impressed.
#Informative and useful to understand the program. Geared mostly to suit the Western diet

So Amazon's mentiment labels are not so accurate, and we relabelled those confused reviews. And based on that, we ran CNN to do the sentiment analysis again.

Run sentiment analysis on CNN

We built one CNN model which concatenated unigram, bigram and trigram. And set manually labelled reviews and same sentiment result reviews given by Amazon and naive as groundtruth. From the plots below, we can see that the model has an excellent performance.

After that, we used the trained model based on relabelled data to predict test data, and set Amazon sentiment as the Y values. So that we can see the difference between relabelled one and Amazon tags.

The accuracy is 69.34% which shows that Amazon reviews have about 70% correct rate which is not too high. So in the future, if we want to derive some valuable information from reviews' sentiment, Amazon's sentiment tags are not so reliable. Furthermore, for Amazon, we recommended a more fancy algorithm to assign sentiment to each review.

Layer (type)	Output Shape	Param #	Connected to
main_input (InputLayer)	(None, 500)	0	
embedding (Embedding)	(None, 500, 100)	1000100	main_input[0][0]
conv_unigram (Conv1D)	(None, 500, 64)	6464	embedding[0][0]
conv_bigram (Conv1D)	(None, 499, 64)	12864	embedding[0][0]
conv_trigram (Conv1D)	(None, 498, 64)	19264	embedding[0][0]
pool_unigram (MaxPooling1D)	(None, 1, 64)	0	conv_unigram[0][0]
pool_bigram (MaxPooling1D)	(None, 1, 64)	0	conv_bigram[0][0]
pool_trigram (MaxPooling1D)	(None, 1, 64)	0	conv_trigram[0][0]
flat_unigram (Flatten)	(None, 64)	0	pool_unigram[0][0]
flat_bigram (Flatten)	(None, 64)	0	pool_bigram[0][0]
flat_trigram (Flatten)	(None, 64)	0	pool_trigram[0][0]
concate (Concatenate)	(None, 192)	0	flat_unigram[0][0] flat_bigram[0][0] flat_trigram[0][0]
dropout (Dropout)	(None, 192)	0	concate[0][0]
dense (Dense)	(None, 192)	37056	dropout[0][0]
output (Dense)	(None, 1)	193	dense[0][0]

Total params: 1,075,941
Trainable params: 1,075,941
Non-trainable params: 0

Figure 11 CNN Model

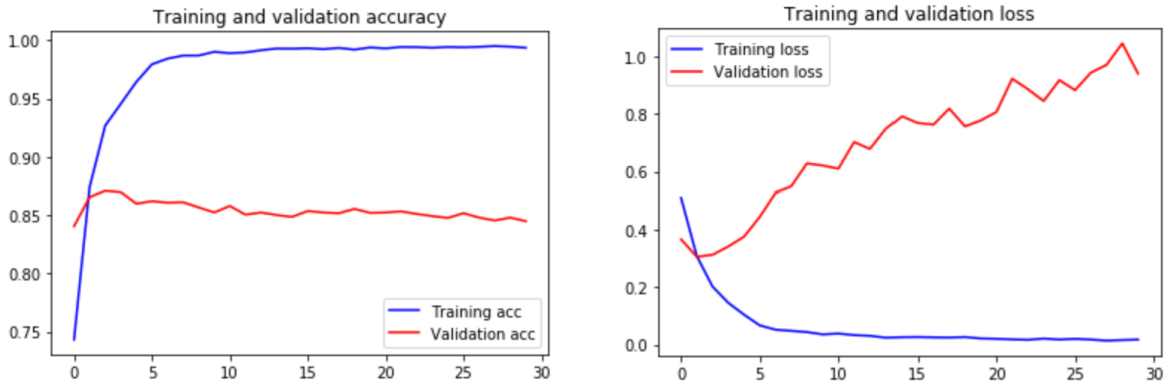

Figure 12 Accuracy of CNN

Word2Vec Modeling

Aiming to build up relationships between words to derive insights, we trained our own word2vecs based on the corpus derived from the reviews. We first concatenated all the review sentences for each cluster, and then cleaned the sentences. Using those cleaned sentences, we were able to train the word2vec model using gensim package. And then we visualized the word vectors using t-SNE in 2D plane. It could be observed that there are no structural differences in the three models.

```
1 my_word_2_vec(0)
```

```
Raw Corpus contains 965,846 characters
The punkt tokenizer is loaded
We have 11,229 raw sentences
We have 11,229 clean sentences
The dataset corpus contains 179,511 tokens
The vocabulary is built
Training finished
Model saved
```

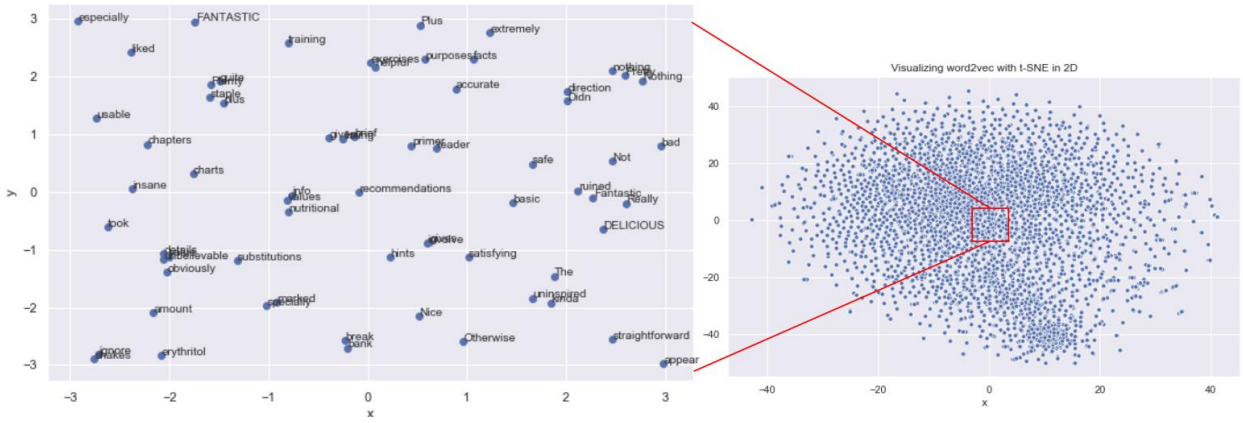

Figure 13 'Easy to Cook' Cluster

```
1 my_word_2_vec(1)
```

```
Raw Corpus contains 1,627,166 characters
The punkt tokenizer is loaded
We have 18,017 raw sentences
We have 18,017 clean sentences
The dataset corpus contains 300,598 tokens
The vocabulary is built
Training finished
Model saved
```

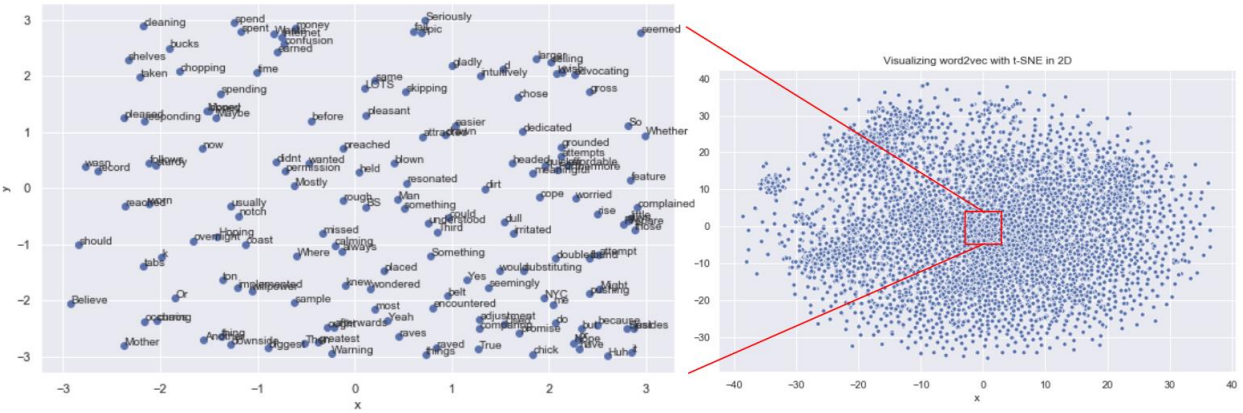

Figure 14 'Diet & Health' Cluster

```
1 my_word_2_vec(2)
```

```
Raw Corpus contains 1,465,886 characters
The punkt tokenizer is loaded
We have 16,172 raw sentences
We have 16,172 clean sentences
The dataset corpus contains 270,775 tokens
The vocabulary is built
Training finished
Model saved
```

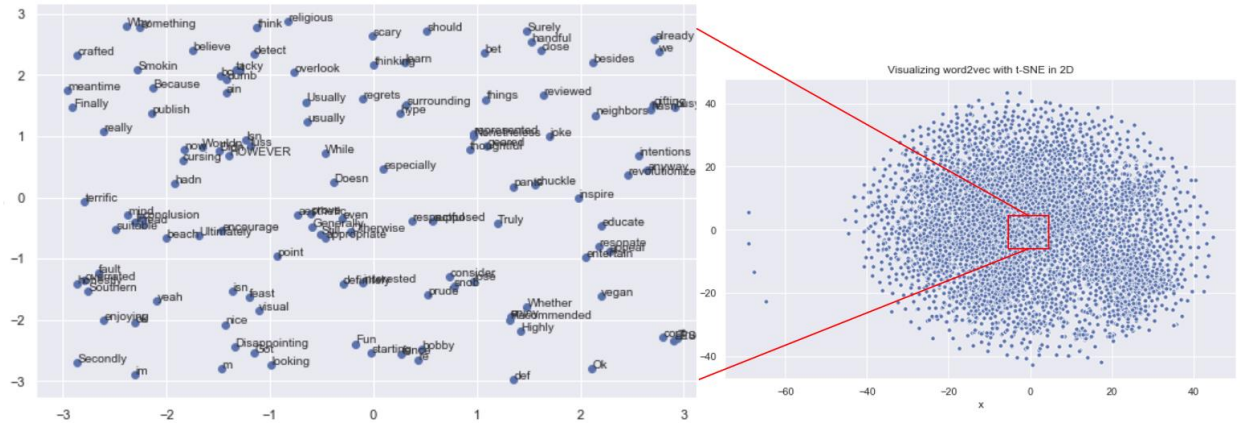

Figure 15 'Popular Chef' Cluster

Word2Vec Similarity Distance

Based on the Word2Vec Similarity Distance, we targeted several customer groups and searched for their similar words.

Implications on the Age of Book Audiences

When we search for words similar to 'husband' and 'like', we found an interesting phenomenon that the 3 clusters may indicate 3 age groups. As shown in the figure below, husbands of customers in cluster2 maybe young. In cluster1, omni means omnivorous and triglycerides, picky, seems like that kind of middle-aged men who may have hyperlipidemia, like eating, but also like to complain. Husbands belong to cluster0 are more likely to old men who are supportive of their wives, like rambling but still like complaining.

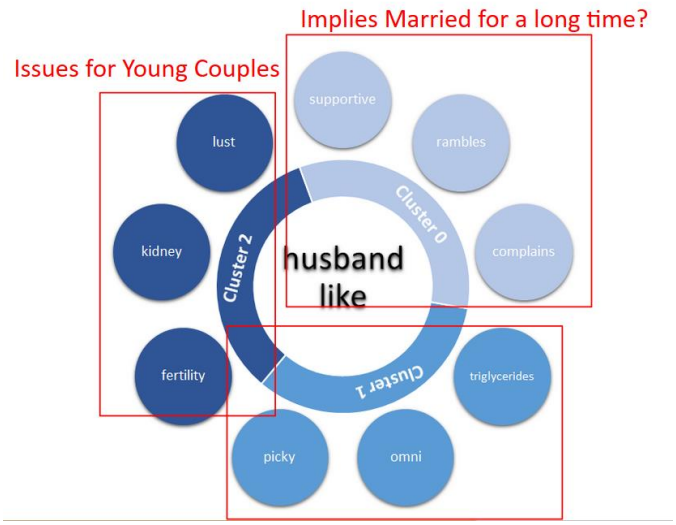

Figure 16 Similar words for 'husband' & 'like'

Implications on Reader's Culture/BackGround

We search words similar to 'chinese' to find out what kind of diet are Chinese interested in or people are more interested in what kind of Chinese food. Figure 16 shows the results. Dumplings, barbecue, ribs, exactly fit our impression of Chinese food. What surprises me is how much Chinese love pickles. Figure 17 shows that Asians attach great importance to staple foods and they still love kimchee. 'Illustrated' may indicates that cookbook with pictures are more popular.

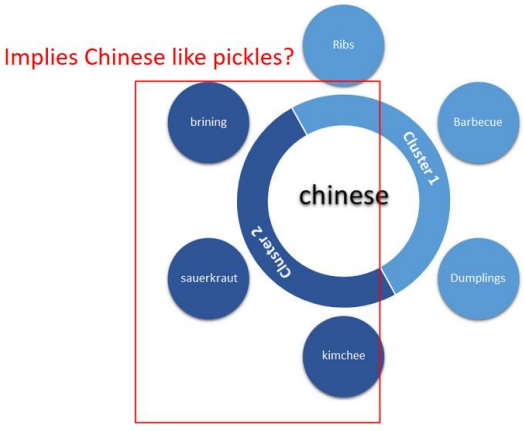

Figure 17 Similar Words for 'chinese'

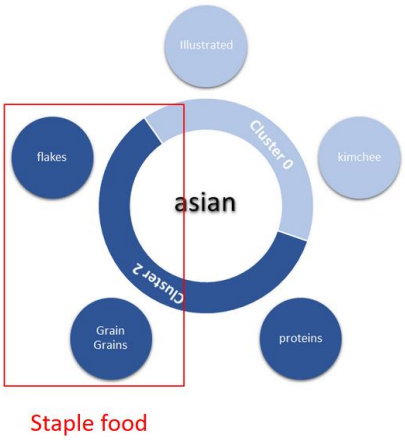

Figure 18 Similar Words for 'asian'

Implications on Cookbooks for Kids

For toddlers, their food should be easy to swallow and digest, like mush. Cluster2, which means recipes of Famous chief may share more complex recipes, like macaroons. In cluster0, easy to cook, moms focus more on how to let their children eat something may be good to them but they don't like, like carrots or something. For words similar to 'kid' and 'meal', 'Reheat' and 'simpler'

may indicate that mothers are more likely to learn recipes of simple and ready-made foods. In cluster2 (famous chefs), the recipes are going to be more complex, like casseroles.

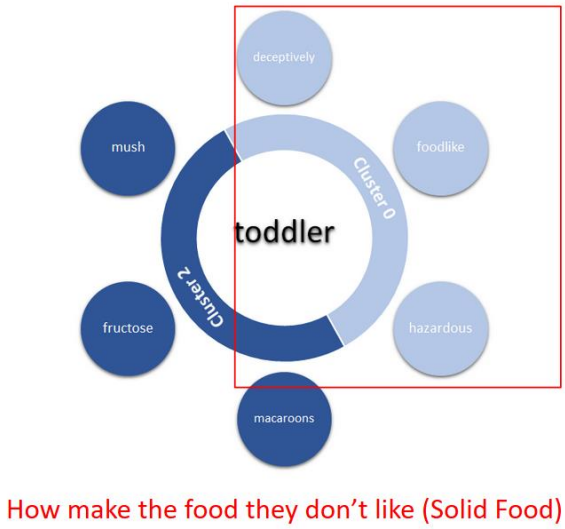

Figure 19 Similar Words for 'toddler'

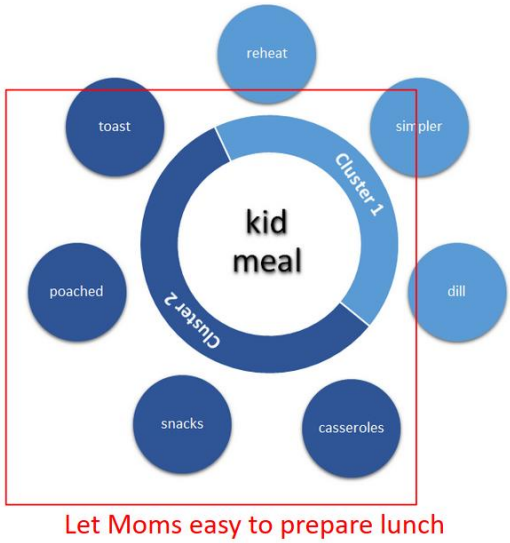

Figure 20 Similar Words for 'kid' & 'meal'

Implications on Special Diets

We're trying to look at the dietary needs of certain populations, for example, diabetes and vegetarian. Diabetes needs to control their intake of nutrition each meal, such as they should not eat a lot of starch in breakfast. As a result, it's difficult to make meals for diabetes. In "Easy to cook" cluster, the result shows they need easy to prepare and nutritious foods. For vegetarians, some of them want to lose weight or stay healthy, so cluster 1 shows they care about the food weather sensitivities and toxic. Also, they want to try different recipes, so cluster 2 shows they want to learn more recipes by famous chefs.

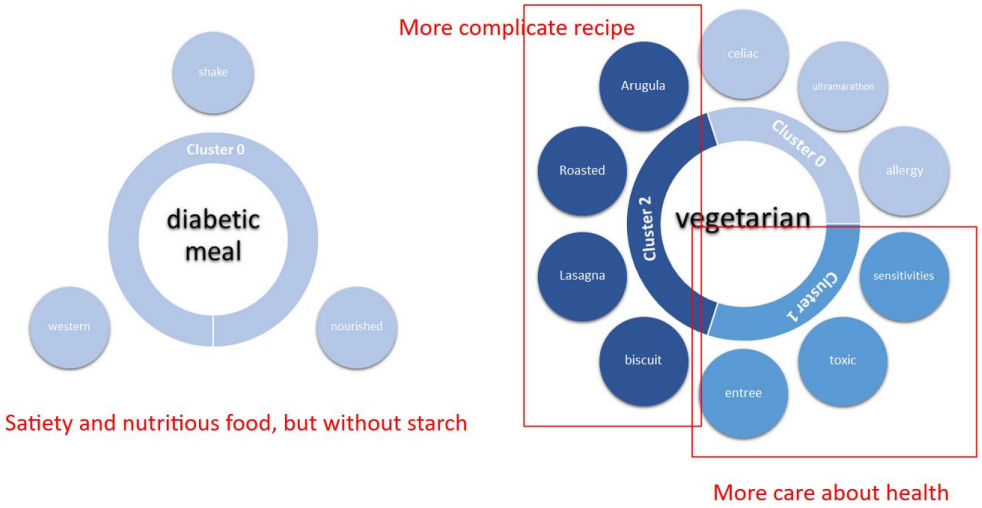

Figure 21 Similar Words for 'diabetic' & 'meal'

Figure 22 Similar Words for 'vegetarian'

Implications on Work vs Life

We also want to know what is the difference between workdays and weekends, and the result shows readers do like cooking simple meals on weekdays, and cooking more complex food on weekends. It also reflects the breakfast, they want easy and fast cooking food for the first meal in the day.

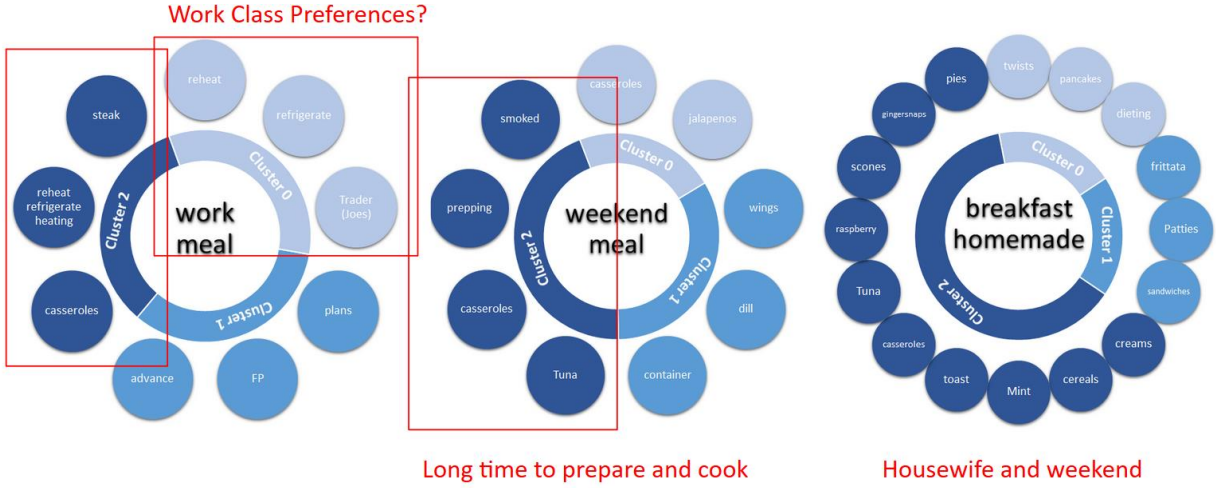

Figure 23 Similar Words for 'work' & 'meal'
Figure 24 Similar Words for 'weekend' & 'meal'
Figure 25 Similar Words for 'breakfast' & 'homemade'

CONCLUSION

From the results of the cluster analysis, we could roughly divide all the books in our datasets into three clusters, and summarize them with: "Easy-To-Cook", "Diet & Health", "Famous Chefs".

Comparing the original sentiment labeling by Amazon.com and our sentiment analysis, we could conclude that the logic behind Amazon's labeling rules is too simple. In some situations, Amazon's sentiment tags are obviously wrong. And in some cases, reviews show that the book is good but can't meet some personal requirements. For those reviews, Amazon tagged them as negative, but for the book, it should be labelled as positive. So for sentiment analysis part, we can't rely too much on Amazon sentiment labels. Negative reviews tend to focus on these aspects: Authors, Price, Waste-of-time

From the word2vec models we were able to derive meaningful insights by calculating similarity distances between words. We could see that housewives are an important part of the cookbook

readers, whereas different ages of the group have different preference in terms of cooking recipes and styles, corresponding to the book clusters. Understanding the cultural makeup of the readers is important, as one might think asians loves pickles, but do they?
Special groups of target readers/beneficiaries, such as toddlers/kids, have special needs in terms of nutrition, whether an easy recipe can fulfill these requirements remain a question.
Taking into account of the work-life balance of the readers is also an important takeaway for the Cookbook publishers, e.g. Work Class prefers frozen food or food that is easy to cook in minutes.

FUTURE WORK

Although we are able to derive some interesting insight from the word2vec model, there is currently no systematic way to gain such knowledge. We may need to search for relevant literature to develop a general and systematic approach to obtain these insights automatically from the models. Also, we might want to use part of speech tagging to identify important features of the text and train our model more carefully.

In terms of sentiment analysis of the reviews, we found that Amazon.com labeling is not entirely correct and accurate. We can expand our research purpose into deriving a good strategy for sentiment labeling.

There are two ways to label text, manual labeling and automatic labeling. Here we used manual labeling, which is more accurate and effective. But it cost lots of time, in the future we could try to build a model to label documents automatically based on documents we have labeled.

REFERENCES

Elli, M. S., & Wang, Y. F. (2015). Amazon reviews, business analytics with sentiment analysis. Retrieved from https://pdfs.semanticscholar.org/bbb4/b549cae71fb74680764fd3fe4d72b705f4f4.pdf

Kessler, W., Klinger, R., & Kuhn, J. (2015). Towards opinion mining from reviews for the prediction of product rankings. *In Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 51-57)*. Retrieved from https://www.aclweb.org/anthology/W15-2908

Shrestha, N., & Nasoz, F. (2019). Deep Learning Sentiment Analysis of Amazon.com Reviews and Ratings. Retrieved from https://arxiv.org/ftp/arxiv/papers/1904/1904.04096.pdf

Sohail, S. S., Siddiqui, J., & Ali, R. (2016). Feature extraction and analysis of online reviews for the recommendation of books using opinion mining technique. *Perspectives in Science, 8, 754-756.* Retrieved from https://www.sciencedirect.com/science/article/pii/S221302091630218X

Srujan, K. S., Nikhil, S. S., Rao, H. R., Karthik, K., Harish, B. S., & Kumar, H. K. (2018). Classification of amazon book reviews based on sentiment analysis. *In Information Systems Design and Intelligent Applications (pp. 401-411).* Springer, Singapore. Retrieved from https://link.springer.com/chapter/10.1007/978-981-10-7512-4_40

Tan, W., Wang, X., & Xu, X. Sentiment Analysis for Amazon Reviews.