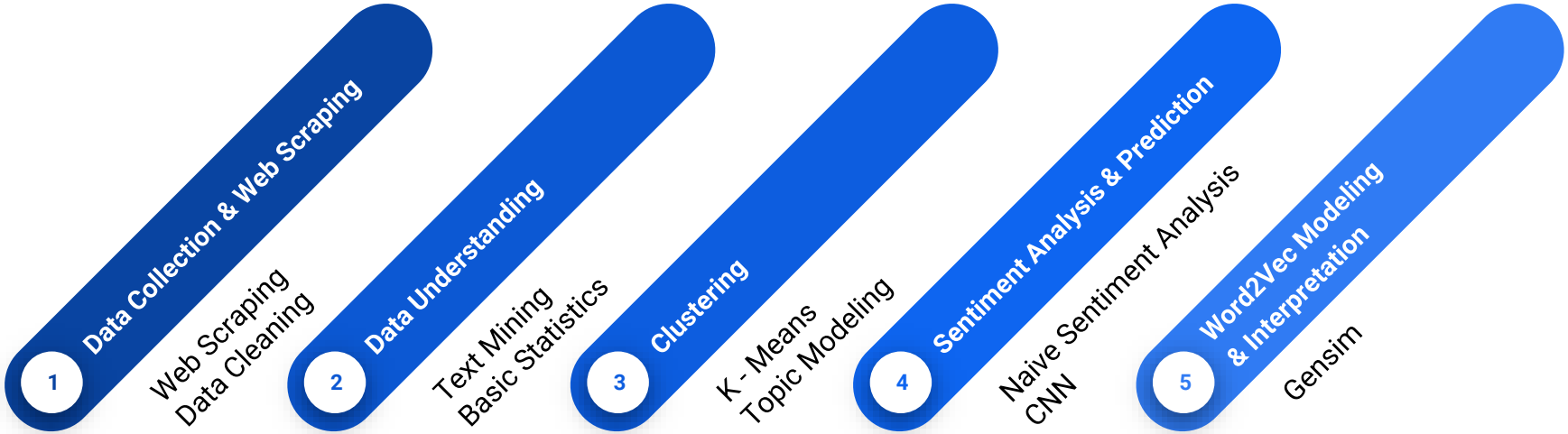
Cookbook Consumer Behavior Analysis From Amazon

Introduction & Problem Statement

- **What key elements made up the 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.
- 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

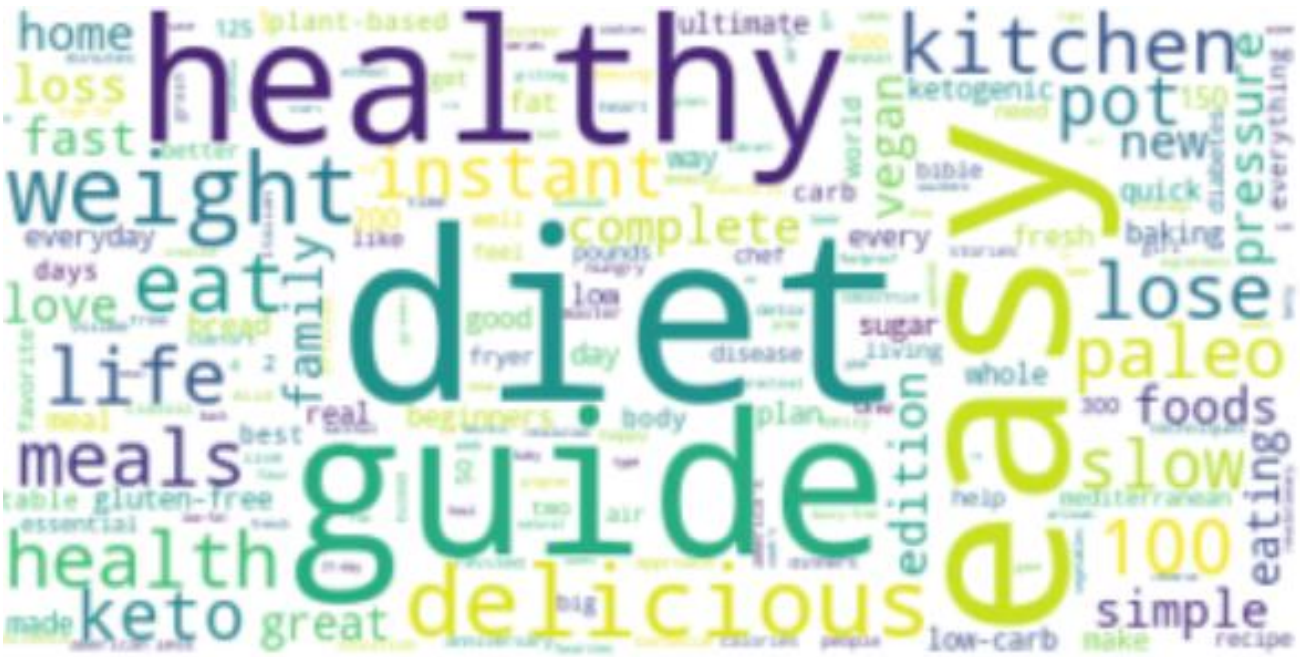
Project Process



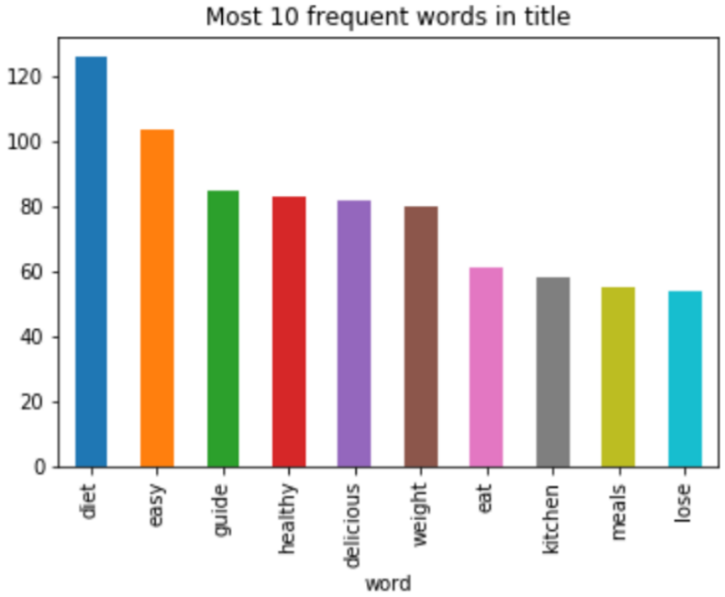
Collect Data: Web Scraping

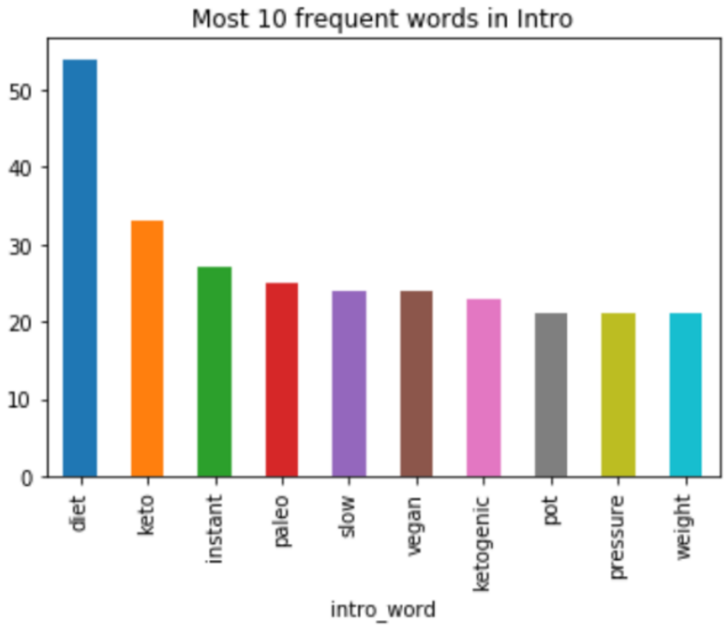
- We used Selenium Webdriver to scrape the data of 1,095 books from ‘Cookbooks, Food & Wine’ category on Amazon.com.
- There are 9 main categories we are interested

WordCloud of Book Titles



Top words: Title and Intro

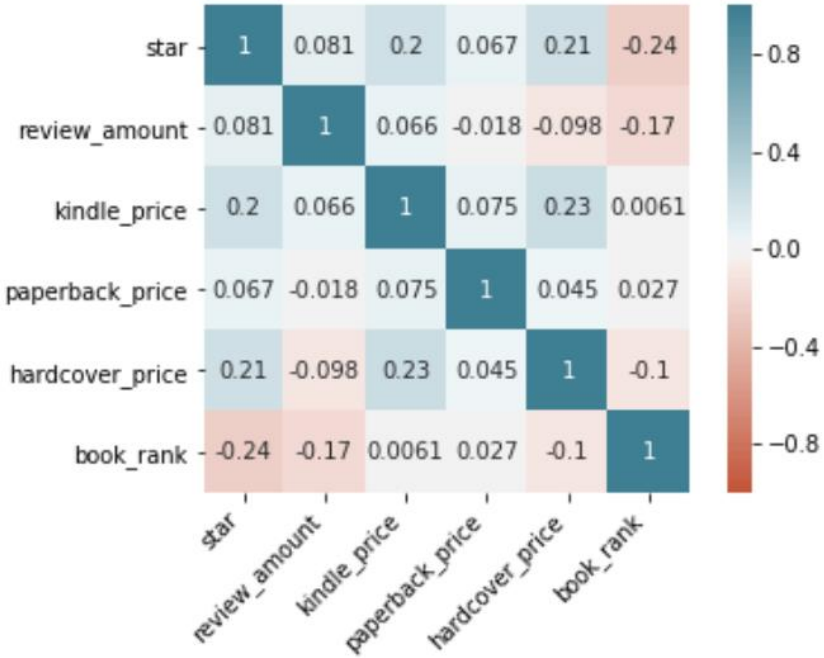




Variables Correlation

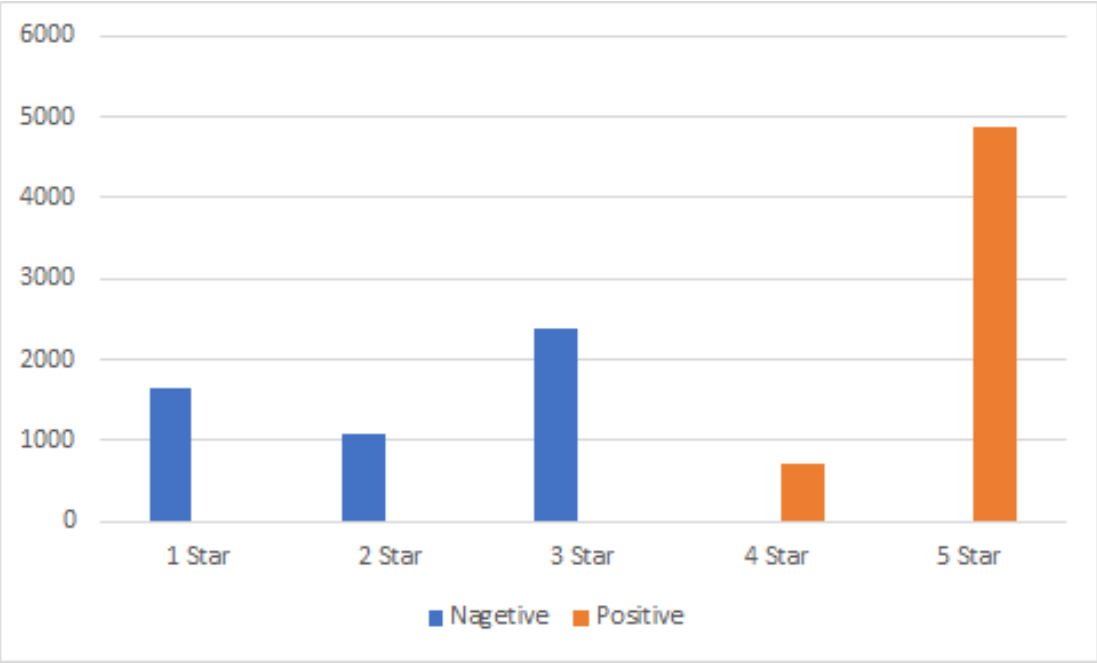
Positive correlation include:
(hardcover_price, star),
(kindle_price, star),
(hardcover_price, kindle_price)

Negative correlation include:
(book_rank, star),
(book_rank,review_amount)



Amazon.com Sentiment Labels

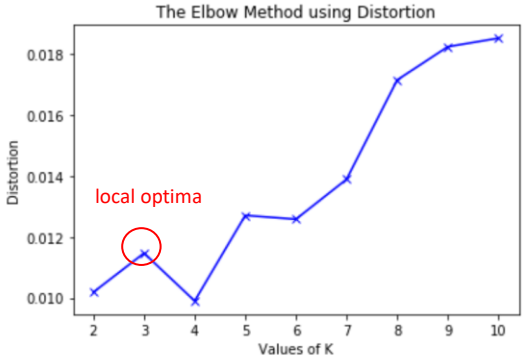
After reviewing 10,685 reviews, we find out Amazon labels the sentiment based on the ratings of star.

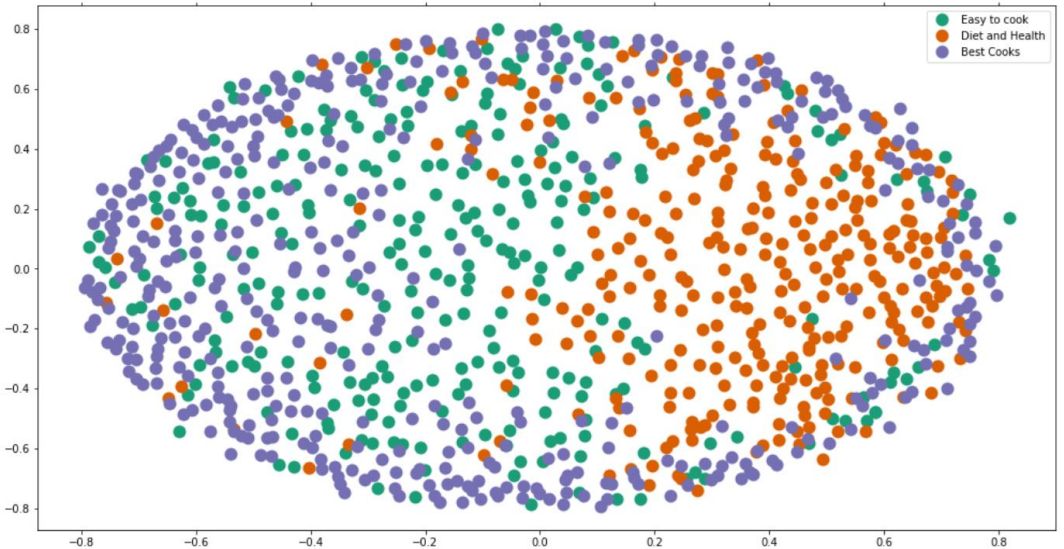


K-Means Clustering Based on Introduction + description

- Filter all meaningless words, like cook, food, recipes
- Find optimal number of clusters by using silhouette_score

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





- KMeansClusterer(cosine distance, repeat = 20)

K-Means Clustering

Cluster 0: Easy to cook

Cluster 1: Diet and Health

Cluster 2: Popular Chefs

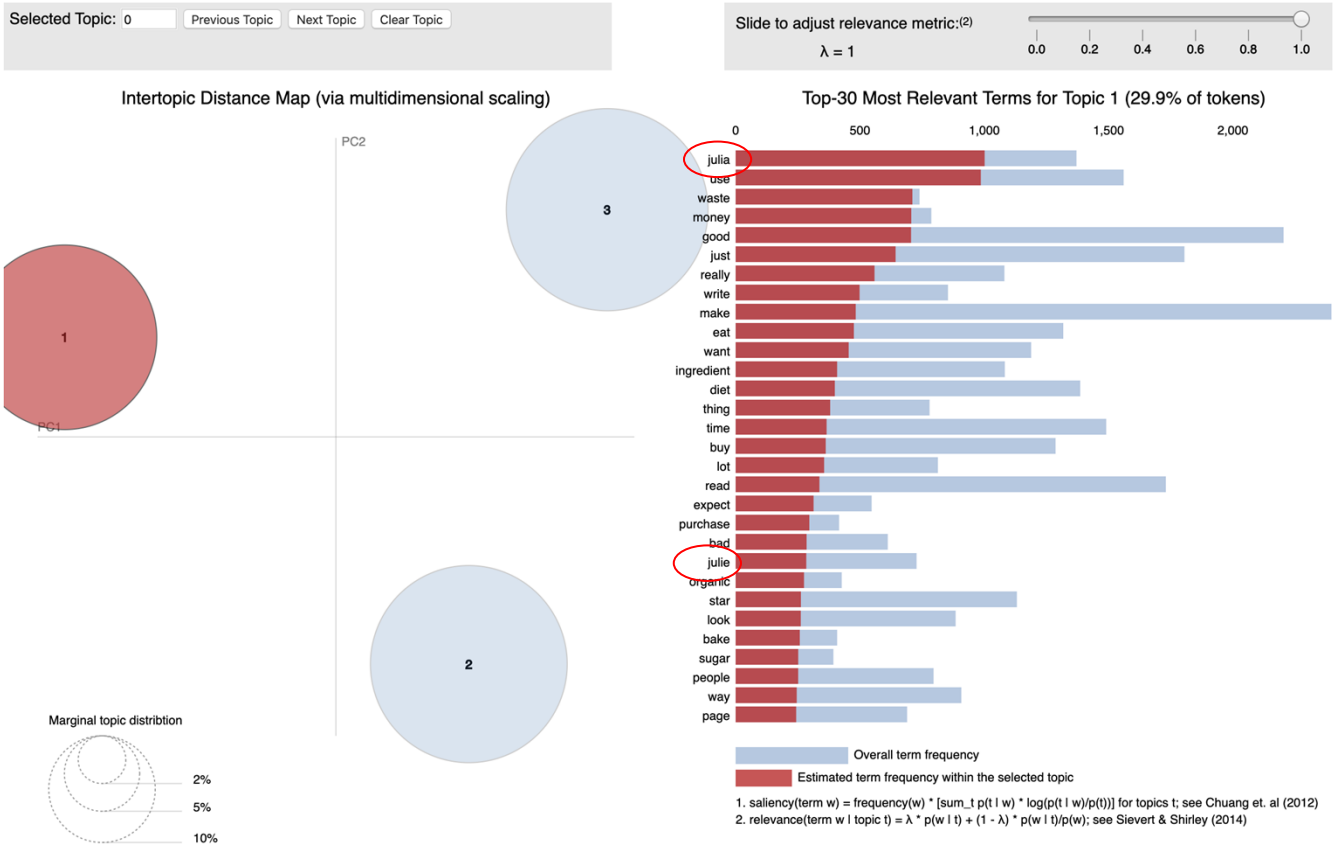
Topic Modeling on Book Reviews (10715 reviews)

- Used **`LatentDirichletAllocation & gensim.models.ldamodel.LdaModel`** to do Topic Modeling on 10715 reviews

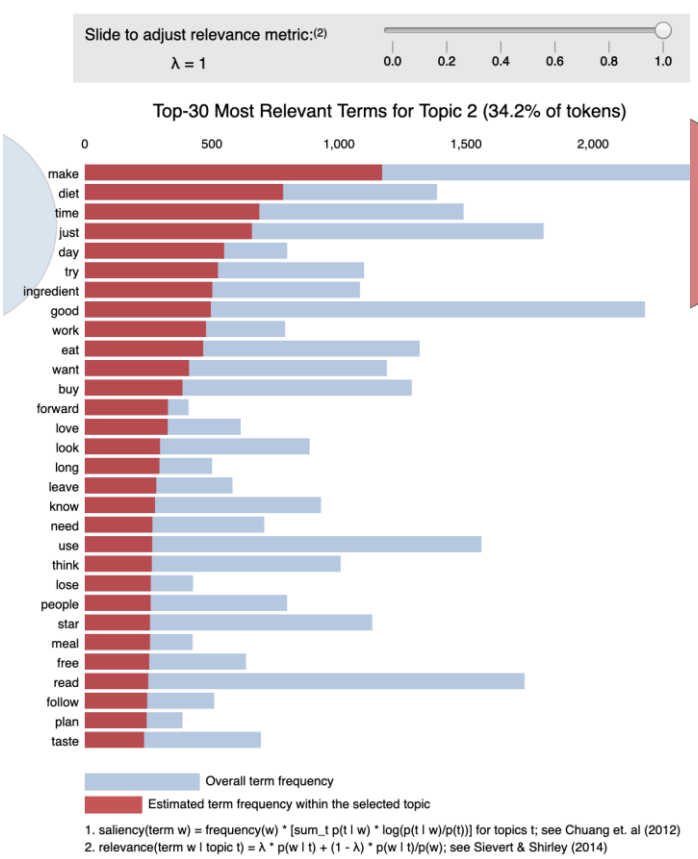
- Used **`GridSearchCV`** to search for the best parameter for **`n_component`**:
  - Although the result showed that 2 components result in the best Log Likelihood & Perplexity
  - We decide to use n = 3 for convenience and consistency with the K-Means Clustering

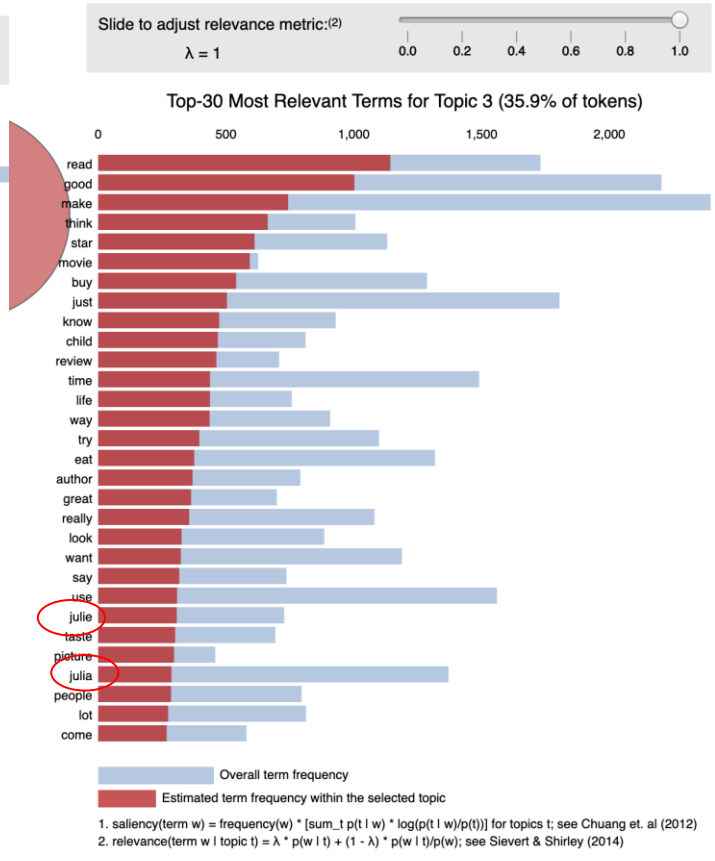
```
Best Model's Params:  {'n_components': 2}
Best Log Likelihood Score:  -211628.0623783079
# Log Likelyhood: Higher the better
Model Perplexity:  1102.8682382842942
# Perplexity: Lower the better. Perplexity = exp(-1. * log-likelihood per word
```

Topic Modeling on Book Reviews (10715 reviews)



Topic Modeling on Book Reviews (10715 reviews)





Conclusion:
Probably not a good idea to do Topic Modeling directly on all reviews,
as reviews on different books will be mixed together.

Sentiment Analysis

Based on 3 clusters - Naive

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

Based on the whole dataset - Naive

sentiment accuracy: 69%

Based on the whole dataset - CNN

Train:

Test : 85.58%

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]

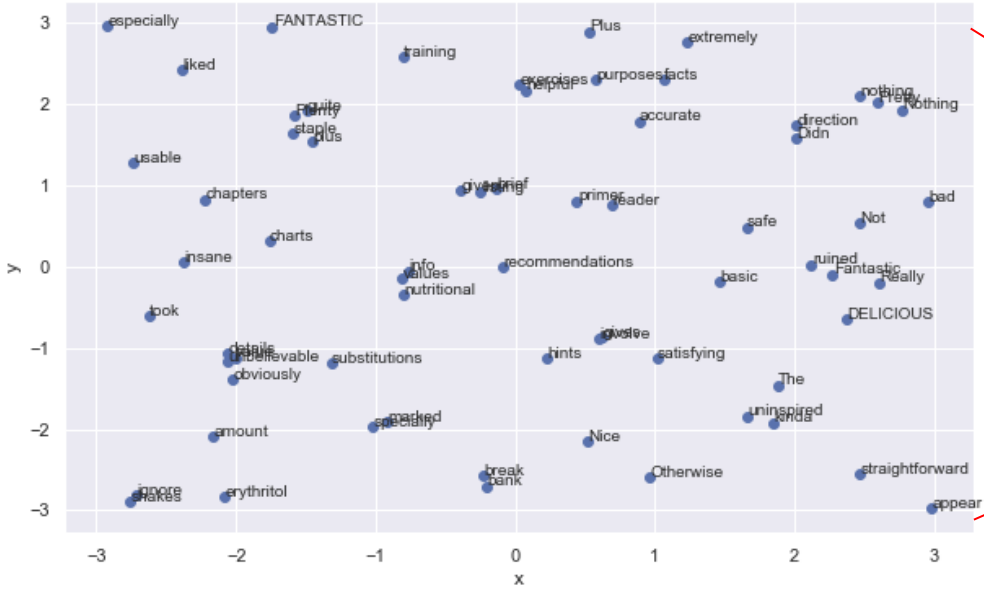
Total params: 1,075,941
Trainable params: 1,075,941
Non-trainable params: 0

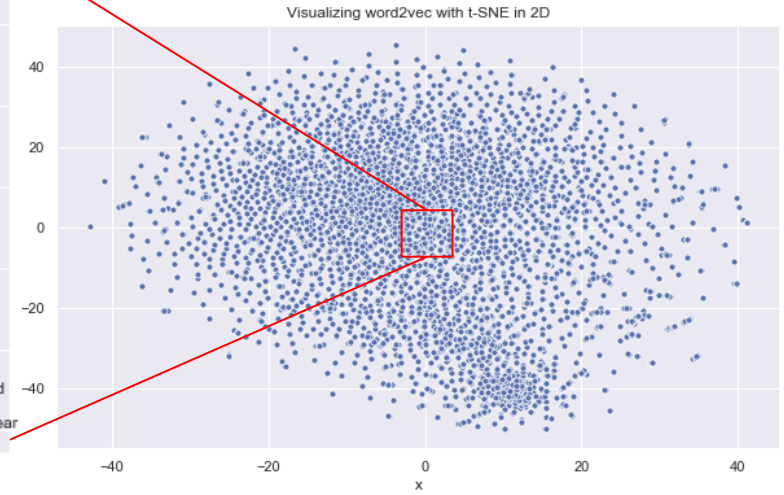
Word2Vec Modeling - ‘Easy to Cook’ Cluster

- **Aim**: build up relationships between words to derive insights
- Used all the review sentences from 3 clusters to build 3 word2vec model
- Visualized the word vectors using t-SNE in 2D plane

```
my_word_2_vec(0)
```

```
Raw Corpus contains 1,554,796 characters
The punkt tokenizer is loaded
We have 18,155 raw sentences
We have 18,155 clean sentences
The dataset corpus contains 288,586 tokens
The vocabulary is built
Training finished
Model saved
```



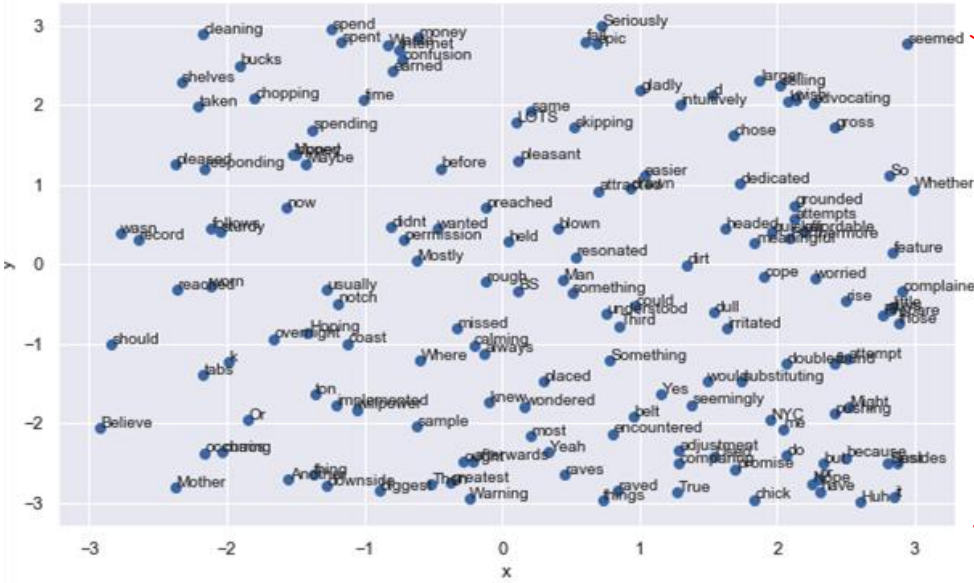


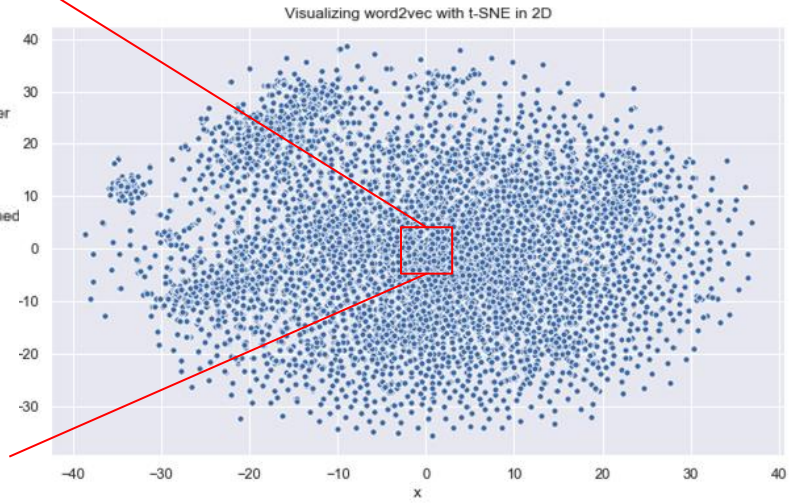
Word2Vec Modeling - ‘Diet & Health’ Cluster

- Used all the review sentences from 3 clusters to build 3 word2vec model
- Visualized the word vectors using t-SNE in 2D plane

```
my_word_2_vec(1)
```

```
Raw Corpus contains 1,955,684 characters
The punkt tokenizer is loaded
We have 22,640 raw sentences
We have 22,640 clean sentences
The dataset corpus contains 360,787 tokens
The vocabulary is built
Training finished
Model saved
```



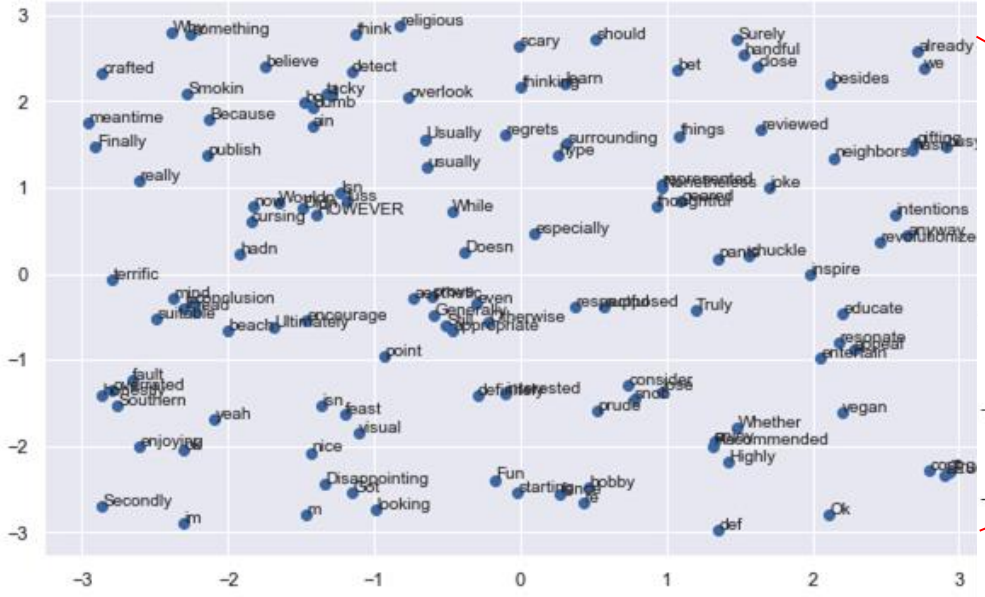


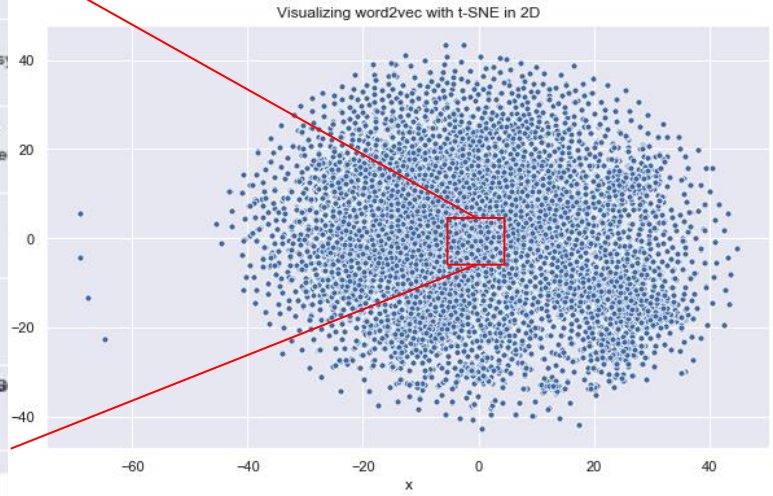
Word2Vec Modeling - ‘Popular Chef’ Cluster

- Used all the review sentences from 3 clusters to build 3 word2vec model
- Visualized the word vectors using t-SNE in 2D plane

```
my_word_2_vec(2)
```

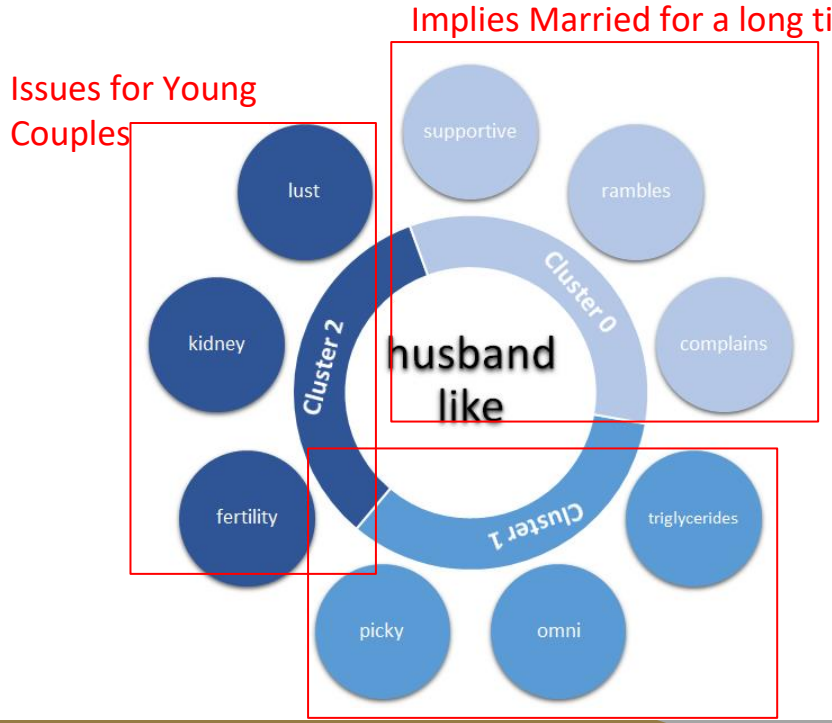
```
Raw Corpus contains 2,241,227 characters
The punkt tokenizer is loaded
We have 24,271 raw sentences
We have 24,271 clean sentences
The dataset corpus contains 414,546 tokens
The vocabulary is built
Training finished
Model saved
```





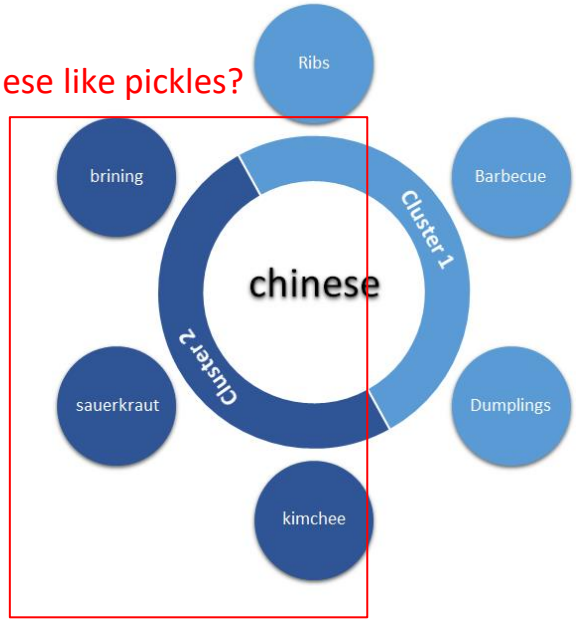
Based on Word2Vec Similarity Distance, We targeted several customer groups and searched for their similar words.

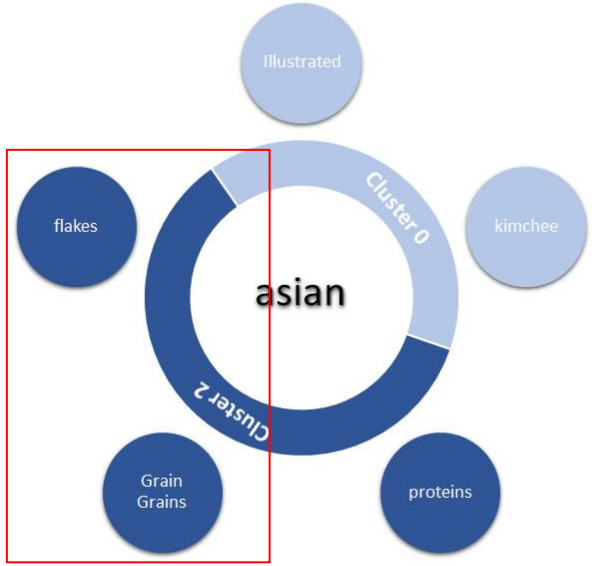
Implications on the Age of Book Audiences



Implications on Reader's Culture/BG

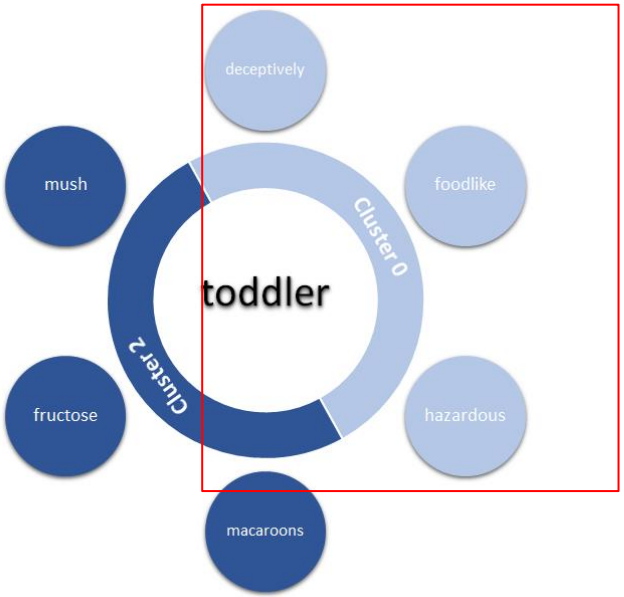
Implies Chinese like pickles?



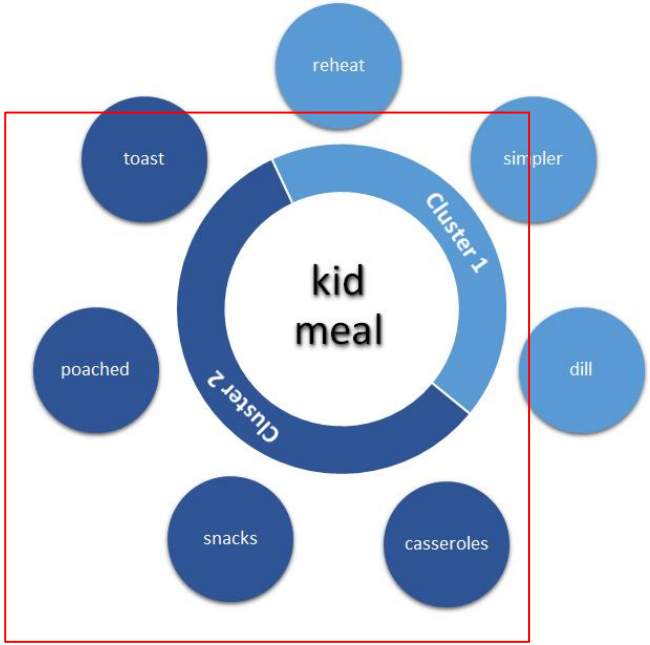


Staple food

Implications on Cookbooks for Kids

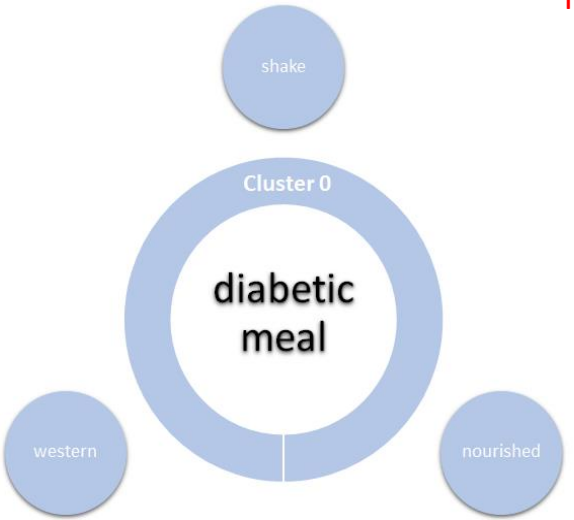


How make the food they don’t like (Solid Food)



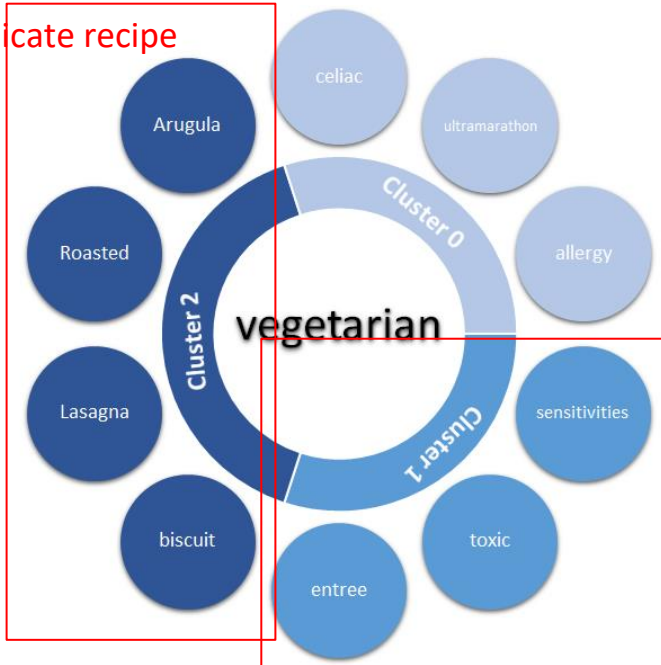
Let Moms easy to prepare lunch

Implications on Special Diets



Satiety and nutritious food, but without starch

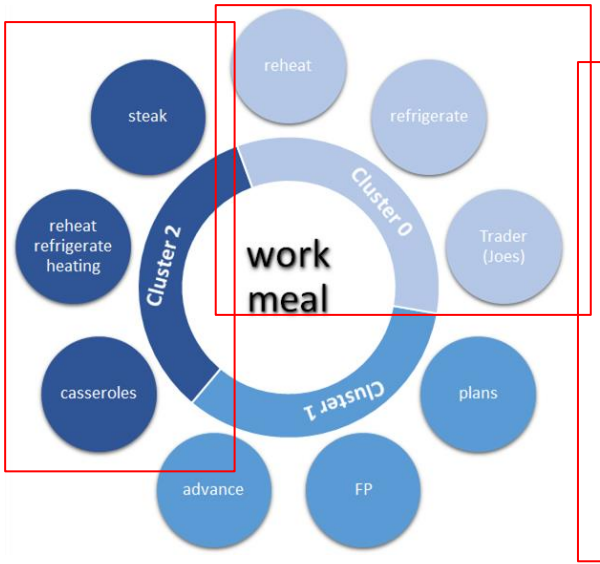
More complicate recipe

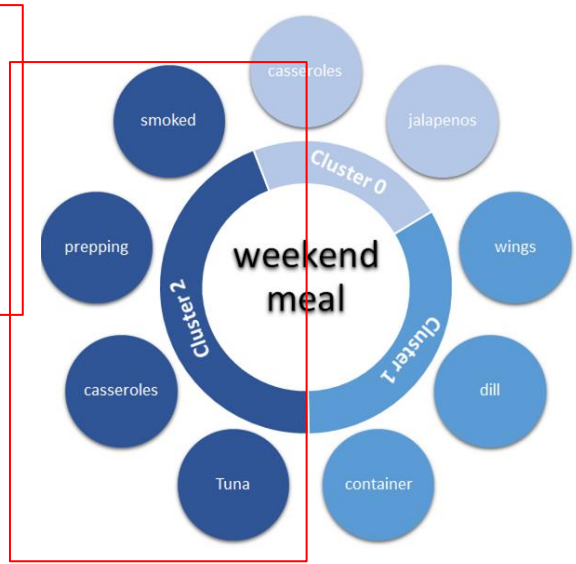


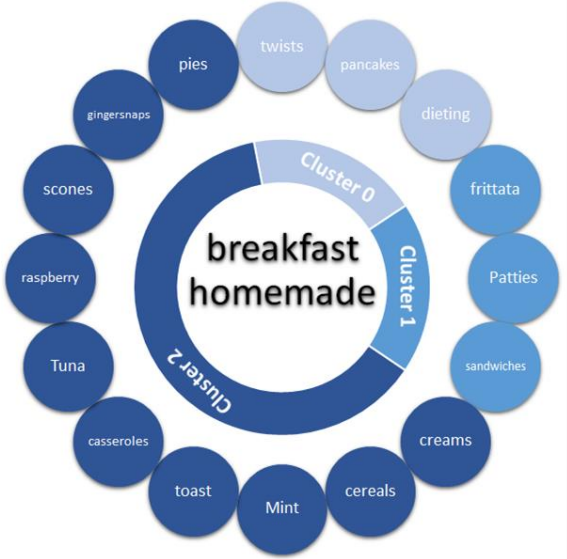
More care about health

Implications on Work vs Life

Work Class Preferences?







Long time to prepare and cook

Housewife and weekend

Consumer Behaviors & Business Insights

- **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.
- Understanding the **culture 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 a 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.

Other Conclusions from Analysis

- 3 Main Areas in Cookbooks:
  - **Easy-To-Cook**
  - **Diet & Health (Keto, Vegan, Gluten)**
  - **Famous Chefs**
- Negative reviews tend to focus on these aspects:
  - Authors, Price, Waste-of-time
- Novel Aspects:
  - Herbs, Detox, Ayurveda - Seeking Cure From Food?
  - Infused Water seemed to be a tough topic to write on

Q&A

Thank you!