Topic modeling

In this part of analyses, we build three topic models to get deeper insight from customer reviews and find further features that influence customer online rating for clothing fashion industry.

We applied 3 methods to conduct topic analysis including K-means clustering algorithm, LSI, and LDA. K-means is a basic clustering method. By using K-means cluster, each review was assigned into one group. LSI(Latent Semantic Indexing) and LDA(Latent Dirichlet Allocation) are different from K-means, they are probabilistic topic models that assumes documents are a mixture of topics. Normally, LSI is a transformation of LDA which has lower accuracy with higher speed.

The first step of topic modeling in this project is running K-means clustering analysis with different number of clusters from 3 to 6. By checking the Silhouette Coefficient values, we decided to generate 5 topics which came with the largest Silhouette Coefficient value.

Table: Silhouette Coefficient for Different Number of Clusters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number of Clusters | 3 | 4 | 5 | 6 |
| Silhouette Coefficient | 0.016 | 0.019 | 0.029 | 0.013 |

Below table shows the top 10 important features in each cluster generated by K-means method.

Table: K-means Clustering Result

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster 0: | Cluster 1: | Cluster 2: | Cluster 3: | Cluster 4: |
| top | easter | placket | 118lbs | dress |
| love | catalog | space | chest | love |
| size | hopes | section | impeccable | size |
| great | hoped | close | triangle | wear |
| fit | frumpy | mid | giant | fit |
| like | ok | lower | enormous | great |
| wear | gorgeous | button | accommodate | like |
| just | didn | ope | usually | beautiful |
| color | pretty | bubble | size | can |
| small | ordered | shut | six | flattering |

To keep consistency, we build 5-topic model with LSI and LDA. Below tables shows the result of LSI and LDA.

Table: Result of LSI

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
| '-1.000\*"nan" + -0.000\*"heavy" + 0.000\*"coat" + -0.000\*"slip" + 0.000\*"pockets" + -0.000\*"extremely" + 0.000\*"arms" + -0.000\*"leggings" + 0.000\*"amazing" + 0.000\*"navy"' | '0.234\*"dress" + 0.186\*"love" + 0.185\*"size" + 0.184\*"top" + 0.174\*"great" + 0.157\*"fit" + 0.148\*"wear" + 0.145\*"small" + 0.140\*"like" + 0.134\*"just"' | '0.703\*"dress" + -0.274\*"shirt" + -0.236\*"jeans" + -0.209\*"great" + -0.168\*"sweater" + -0.145\*"top" + -0.145\*"soft" + -0.140\*"pants" + -0.127\*"cute" + 0.107\*"beautiful"' | '0.385\*"dress" + -0.284\*"small" + 0.278\*"comfortable" + 0.275\*"great" + -0.237\*"top" + 0.193\*"love" + -0.168\*"size" + -0.163\*"large" + -0.161\*"ordered" + -0.155\*"xs"'), | '0.617\*"shirt" + -0.231\*"sweater" + -0.229\*"size" + 0.217\*"dress" + 0.200\*"cute" + -0.187\*"pants" + -0.141\*"petite" + -0.134\*"small" + 0.131\*"top" + -0.127\*"fit"')] |

Table: Result of LDA

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
| '0.030\*"nan" + 0.012\*"wash" + 0.011\*"like" + 0.010\*"dry" + 0.010\*"washed" + 0.008\*"top" + 0.006\*"romper" + 0.006\*"hand" + 0.006\*"look" + 0.005\*"retailer"' | '0.024\*"love" + 0.020\*"soft" + 0.020\*"sweater" + 0.018\*"wear" + 0.017\*"great" + 0.016\*"color" + 0.014\*"shirt" + 0.014\*"comfortable" + 0.014\*"like" + 0.011\*"super"' | '0.041\*"size" + 0.028\*"fit" + 0.015\*"like" + 0.013\*"wear" + 0.013\*"pants" + 0.013\*"petite" + 0.012\*"jeans" + 0.012\*"love" + 0.012\*"ordered" + 0.011\*"waist"' | '0.037\*"top" + 0.021\*"like" + 0.016\*"just" + 0.015\*"fabric" + 0.015\*"love" + 0.014\*"back" + 0.014\*"really" + 0.013\*"small" + 0.012\*"look" + 0.012\*"color"' | '0.060\*"dress" + 0.023\*"love" + 0.018\*"wear" + 0.017\*"great" + 0.015\*"size" + 0.013\*"fit" + 0.013\*"perfect" + 0.011\*"fits" + 0.011\*"can" + 0.010\*"small"' |

Comparing above results, we identify that the result of LDA is better to be interpreted and more make sense. In this case, we used LDA result for following projects analysis and defined names of each topic in the result of LDA (More details in below Table). There are 2 insights we can generated form topic modeling. First of all, how to clean clothes is a problem that customers considered when or after they are buying clothes. Secondly, customers considered style, color and comfortable for tops, while they consider more about size for trousers.

Table: Definition of Topics

|  |  |  |
| --- | --- | --- |
| Topic # | Topic Name | Details |
| Topic 1 | Laundry | This topic meanly discussed how to clean cloth |
| Topic 2 | Top waring | This topic discussed some kind of tops people ware and how people feel |
| Topic 3 | Trousers size | The topic identified for trousers, customers are take more attention on size. |
| Topic 4 | Overviews | The topic gave the overview of the cloth, including color, size, and materials. |
| Topic 5 | Good Dress | This topic interpreted how the clothes are well dressed for customers |

After that we created a table to find the relationship between different topics and ratings.

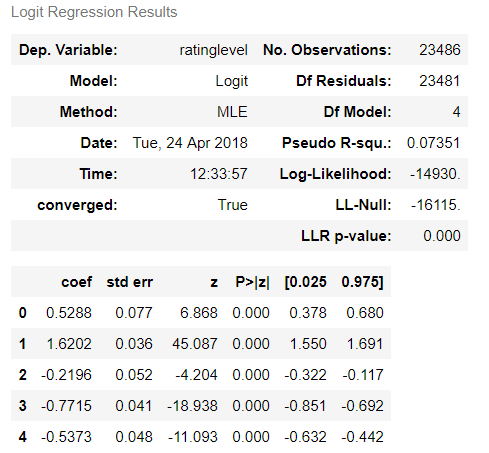
Table: Average probability of each topic appear in different ratings

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ratings** | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
| 1 | 0.41 | 0.24 | 0.08 | 0.19 | 0.08 |
| 2 | 0.44 | 0.24 | 0.08 | 0.18 | 0.06 |
| 3 | 0.41 | 0.22 | 0.09 | 0.18 | 0.08 |
| 4 | 0.26 | 0.21 | 0.15 | 0.22 | 0.15 |
| 5 | 0.17 | 0.15 | 0.20 | 0.23 | 0.24 |

From above table, we can find that low ratings always come with Topic 1 and Topic 2, while high ratings (4&5) has higher probabilities of Topic 3, Topic 4 and Topic 5.

Then we build the logistic regression model using the probability of each topic to predict the rating level. Below table show the result of logistic regression model.

Table: Result of Logistic regression model



The accuracy of the logistic regression model is about 64.93%.