Mid-semester Progress Report

Team Members:
HAITAO ZHOU (HAZ59), YINGZHI YANG(YIY50), XIN JIN(XIJ21)
Team Name:
DA-Boys

Background of Home Depot Product Search Relevance Competition

In the Home Depot Product Search Relevance competition, we are required to develop a model to predict the relevance of search results. Since the customers can't search for the products so accurate, so the more quickly they can find what they want, the more satisfied they will feel for the shopping experience.

Objectives and Significances

The objectives and significances of this project is that we should analysis the relevance between the product and the searching key words, and improve the searching system for the customers to find their product as quickly as possible.

Approach and Engineering Solution

The main approaches we are using is TF-IDF and the vector space model. As the project continues, we may use some more approaches to help us working better. The main tool we are using is R language, since R is a strong statistical computing ad graphics supported programming language.

Introduction to the final report

In the final report，we will describe our process of predicting Home Depot Product Search Relevance. We will find that feature engineering dominates the quality of classification. At least, our final report will contain the following parts:

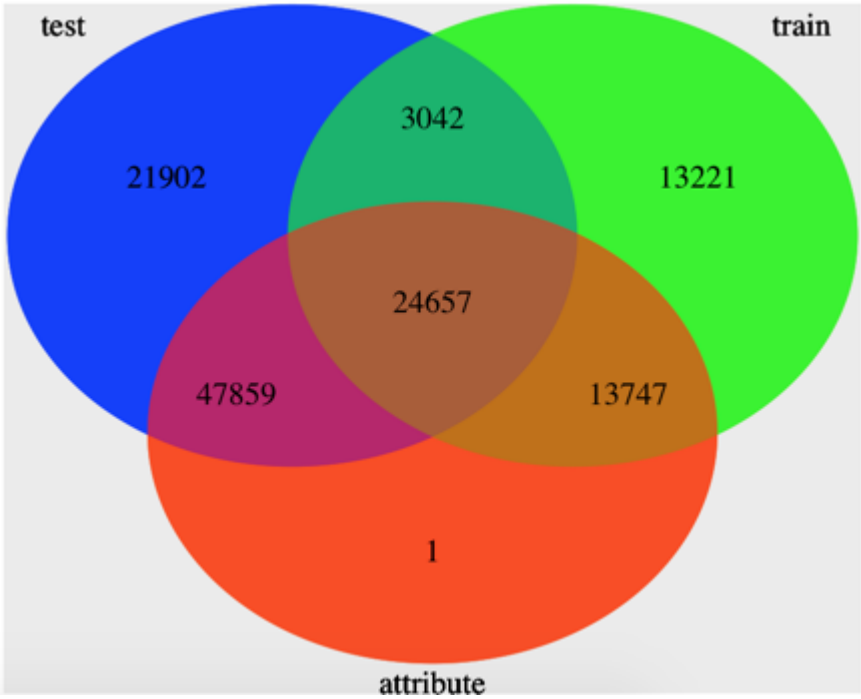
1.Abstract
2.Introduction of the project
3.Data Description
4.Model of solution
5. Diagrams and text description of data analytics
6. Source code
7. Test result
8.Conclusion

Basic Description of Data & Initial Results

In this portion, a description of the work that we have done and some initial results will be shown. These initial results may have some mistakes and flaws, we will improve and perfect them in the future research.

train.csv: In this data profile, product id, product title, search term and relevance are listed. We are going to find how the relevance is calculated using the relationship between product titles and search terms. This is the most important part of this project.
product_description.csv: This data profile lists the descriptions for all the products. We can find whether the descriptions will affect the relevance of search, combined with train.csv.
attribute.csv: After analyzing the value of products listed in this profile, we can find that if a customer searches a product by its value, how the relevance of result will be.
test.csv: We are going to use this profile to test our solution, a relevance value will be calculated for every items in the profile.

1. Overlap of the data

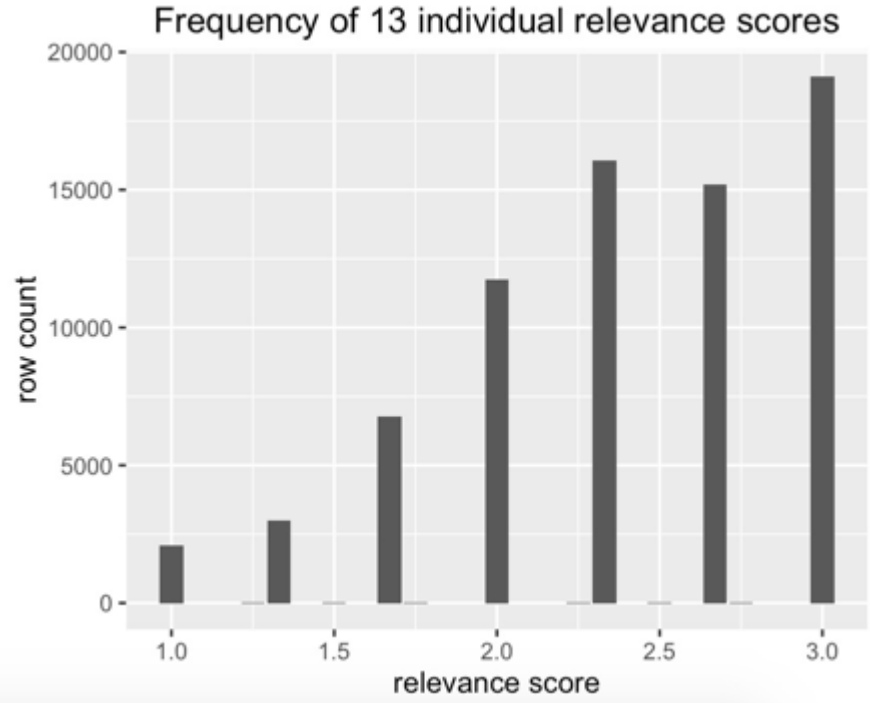


As the graph shown above, we can see the overlap of the data clearly:
1.There are 97460 unique product_uid in test, 54667 unique product_uid in the train, and 27699 of them are common.
2.There are 2044803 rows and 86264 unique product_uid in file attributes.cvs. We can see the intersection of product_uid acroos the three files.
3.There is 1 value in attributes.csv file that is not in either the Train or Test files. On investigating this there are 155 rows that do not have aproduct_uid value. These rows can be removed.

2.Word Cloud of Data

This graph is the word cloud of value in attribute.csv. Ignore some useless words such as 'Yes' and 'No', we get the general situation of the words' frequency in this file.
Actually, there are some flaws in this graph (some very frequent words are useless, many words are overlap and very unclear), we will improve it in the future.

3.Frequency of relevance score



As the graph shown above, we can see the frequency of 13 individual relevance scores in train.csv, which will help us to build the relevance model in the future works.

Plan for next steps

Week 10-11: Do some research on the internet for algorithms we will use and knowledge of data analysis, and do the analytics of data.

Week 12-13: Implement the algorithms for data analytics, and test the solution with test.csv. Improve the solution after testing.
Week 14-15: Submit the solution and finish the final report.