Home Depot Product Search Relevance

Dipak Kumar Singh

April 28, 2016

Abstract

In this Kaggle Search Result Relevance result, we were asked to find the relevance of a search in the Home Depot database given the search query, the product title and the product description. The model built is based on the TFIDF feature selection and model ensembling.

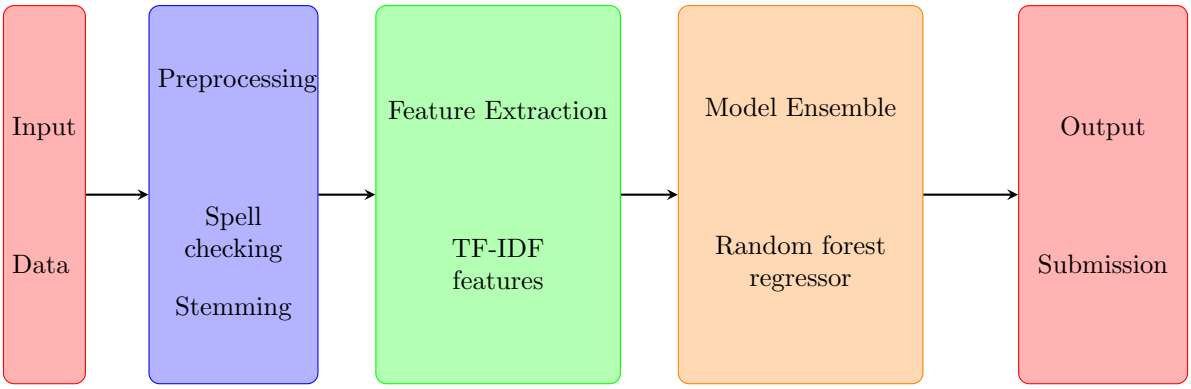
1 Introduction

The customers rely on the Home Depot products to meet their home improvement needs and the search engine of the Home Depot plays an important role to help customers to find out the right product they are looking for with ease. Therefore, it is very important to develop a model which can predict the product with higher relevance fulfilling the shopper's demand [3].

Our solution consists of two parts: feature engineering (TFIDF features) and model ensembling (random forest regressor).

Before we built our model, it was necessary to preprocess the data for spelling correction, stemming and finding some relevant information from the data.

The best single model we have obtained has a Private LB score of 0.48667.



2 Data Exploration

On investigation, we found that there are 74067 rows in the train.csv data and 166693 rows in the test.csv data.
Examining the column product_id there are 54667 unique values in the train.csv

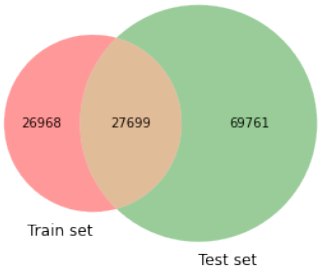


Figure 1: Training and Test

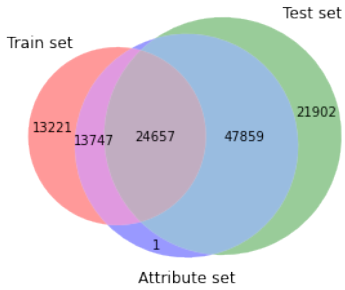


Figure 2: Training, Test and Attributes

data and and 97460 in the test.csv data. There are total of 27699 product_uid values that intersect. This is displayed below in the venn diagram3.
There is one value in attributes.csv file that is neither in Train or Test dataset and there are 155 rows that have missing product_uid value. These rows has been removed.

Furthermore, there are 13 unique relevant score, where the scores (1.25,1.5,1.75,2.25,2.5,2.75) can be neglected since they have relatively low frequencies. The average relevance score decreases as the number of product_uid increases as shown in fig 5 and 6.
On examining the attributes.csv file, we discovered that there are 2044648 rows and relate to 86263 unique products. The attribute name contains 63 color names and 10 brand name. The **mgr brand name** is considered to be the standard brand name.

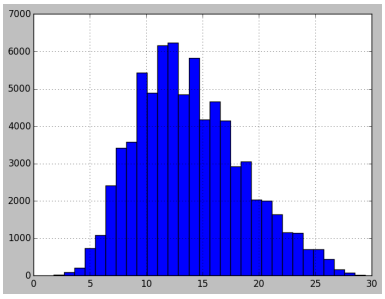


Figure 3: Training and Test

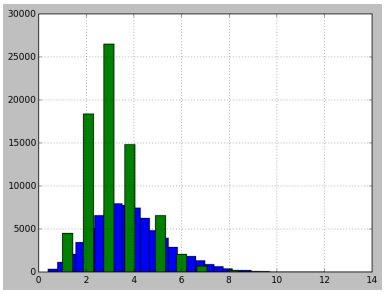


Figure 4: Training, Test and Attributes

3 Preprocessing

The following steps were performed in cleaning up the data set.

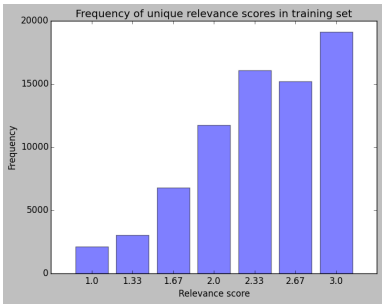


Figure 5: Relevance frequency

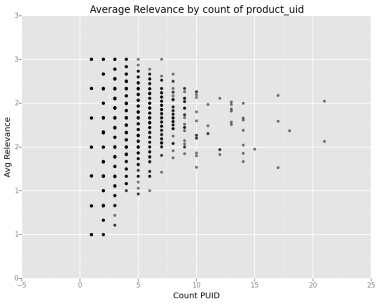


Figure 6: Avg relevance wrt product_uid

3.1 Spelling correction

We found a significant amount of misspellings in the search query (about 13%). Therefore, we used google search engine to fix the typos by sending request to the google engine. The sample of the misspellings we have identified are listed in the Table 1.

misspellings	corrections
steele stake	steel stake
gas mowe	gas mower
closetmade	closetmaid
3/4 vlve	3/4 valve
1x6 cedar boaed	1x6 cedar board
mosia grout sealer	mosaic grout sealer

Table 1: Spelling correction.

3.2 Stemming

We also performed stemming before generating features with Porter stemmer or Snowball stemmer from NLTK package.

4 Feature Selection

We extracted TF-IDF features for cooccurence terms between query unigram and product title, and query unigram and product descriptions. This feature selection is very helpful as it reduce the computation burden and provide better model by selecting some important features on which the the model is built on.

We performed SVD to the above TF-IDF features to obtain a dimension reduced feature vector. Such reduced version was mostly used together with non-linear models, e.g., random forest.

5 Model Ensemble

Given the form of the relevance score, it was more convincing to use regression for prediction. Earlier similar competition, Crowdflower Search Relevance solution also follows similar strategy. We therefore, applied the Random forest regressor to our input data [2].

6 Result

The submission thus generated achieved a private LB score of 0.48667(1228 position out of 2125 teams), where the winning score is 0.43192 [1].

7 Future Work

To enhance our result, we can generate more features based on distance feature, counting feature which can be applied on the combination of model ensembles such as XGBoost, ExtraTreeRegressor etc.

8 Appendix

8.1 Data Description

8.1.1 File Description

- train.csv - the training set, contains products, searches, and relevance scores
- test.csv - the test set, contains products and searches. You must predict the relevance for these pairs.
- product_descriptions.csv - contains a text description of each product. You may join this table to the training or test set via the product_uid.
- attributes.csv - provides extended information about a subset of the products (typically representing detailed technical specifications). Not every product will have attributes.
- sample_submission.csv - a file showing the correct submission format
- relevance_instructions.docx - the instructions provided to human raters

8.1.2 Data Fields

- id - a unique Id field which represents a (search_term, product_uid) pair
- product_uid - an id for the products
- product_title - the product title
- product_description - the text description of the product (may contain HTML content)

- search_term - the search query
- relevance - the average of the relevance ratings for a given id
- name - an attribute name
- value - the attribute's value

We hope you find writeLaTeX useful, and please let us know if you have any feedback using the help menu above.

References

[1] *Home Depot Product Search Relevance*, 2016 (accessed April 24, 2016). `https://www.kaggle.com/c/home-depot-product-search-relevance/leaderboard`.

[2] Chenglong Cheng. *Kaggle Crowd Flower*, 216 (accessed April 24, 2016). `https://github.com/ChenglongChen/Kaggle_CrowdFlower`.

[3] Kaggle Inc. *Home Depot Product Search Relevance*, 2016 (accessed April 24, 2016). `https://www.kaggle.com/c/home-depot-product-search-relevance`.