Mid-Semester Final Project Progress Report

Ja-Jan Hsu (jih247), Leilei Liu (lel74), Zhaoyan Ai (zha4)

I. Introduction

Every merchant is willing to know the exact goods shoppers want to buy by analyzing shoppers' basic description of the goods. So the Home Depot product search relevance model can be used for most of the online shopping websites.

The final goal for the competition is to create a model which can predict the relevance between a list of goods and the exact item in shopper's mind by giving each good a value from 1 to 3. If the value is 3, it means the item is highly relevant in this model.

According to the requirement of the competition, the prediction model should be speed, accuracy and delivering a frictionless customer experience. How to design the prediction model, how to choose the impact factors and how to set the value of each impact factor can be the most significant part of work in this project.
Data files

In this competition, we have five data files and one relevance_instruction word file.

Train.csv file: there are five data fields(id, product_title, product_uid,search_term and relevance).
Test.csv file: there are four data fields(id, product_uid, product_title and search_term).
Product_descriptions.csv file: there are two fields(product_uid and product_description).
Attributes.csv file: there are three field(product_uid, name and value).
Sample_submission.csv file: there are two data fields(id and relevance)

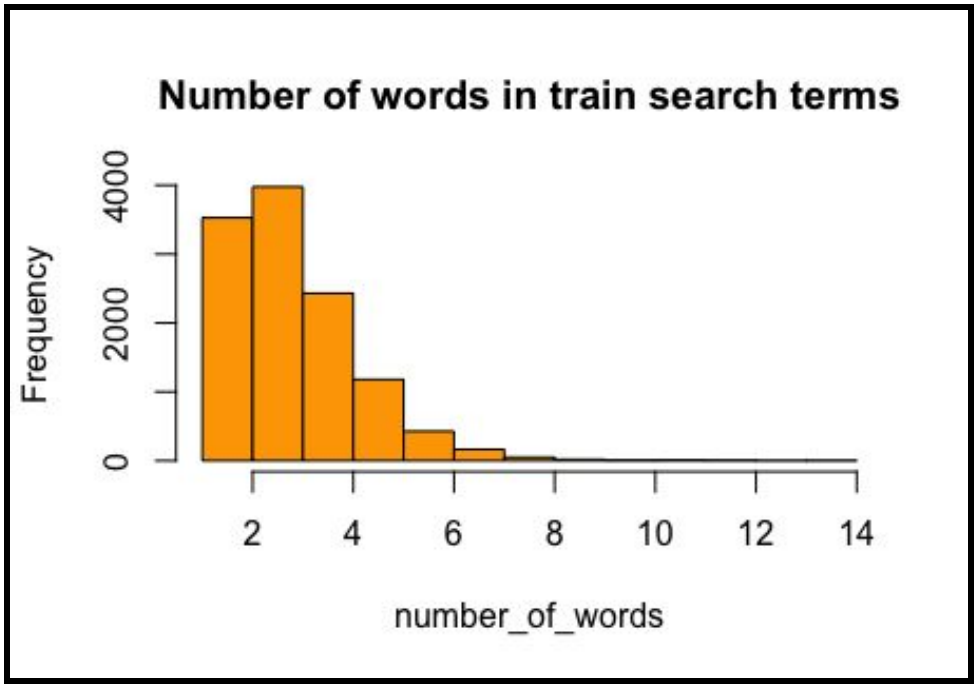
From the sample_submission.csv file, we can notice that we'll need to predict a relevance for each items. The resource data we have are train.csv,

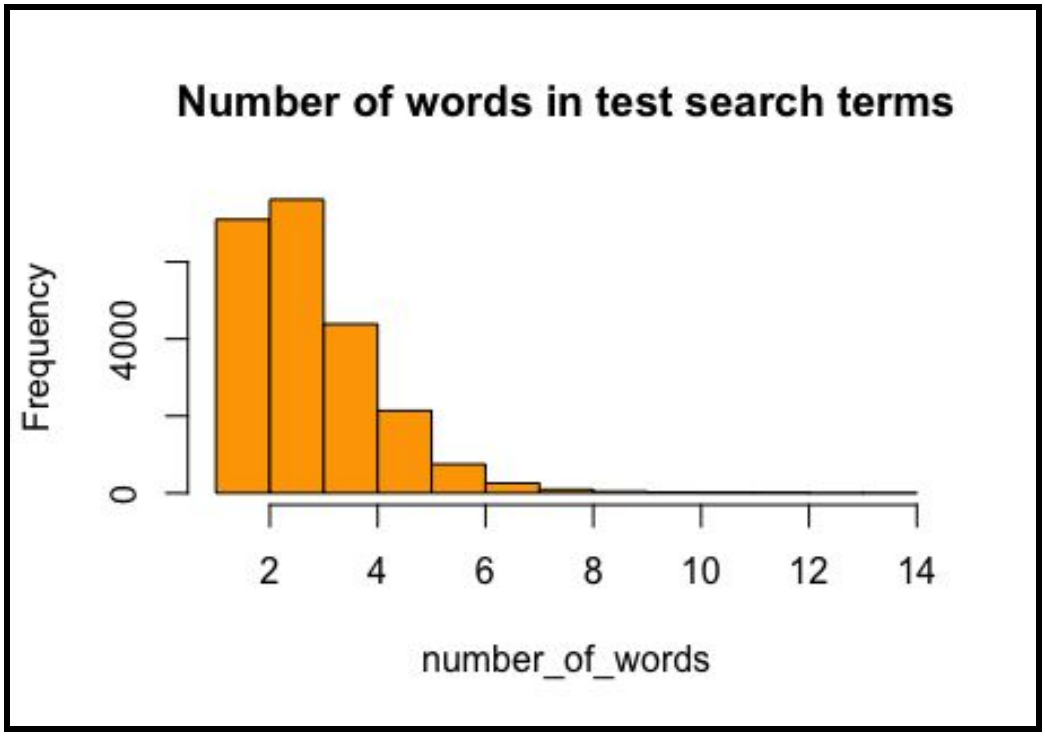
product_description.csv and attribute.csv files. We'll need to find out the relations between each data field and predict a most accurate relevance for each of them. From the relevance_instructions file, it specifies how relevances are being rated.

II. What we have done

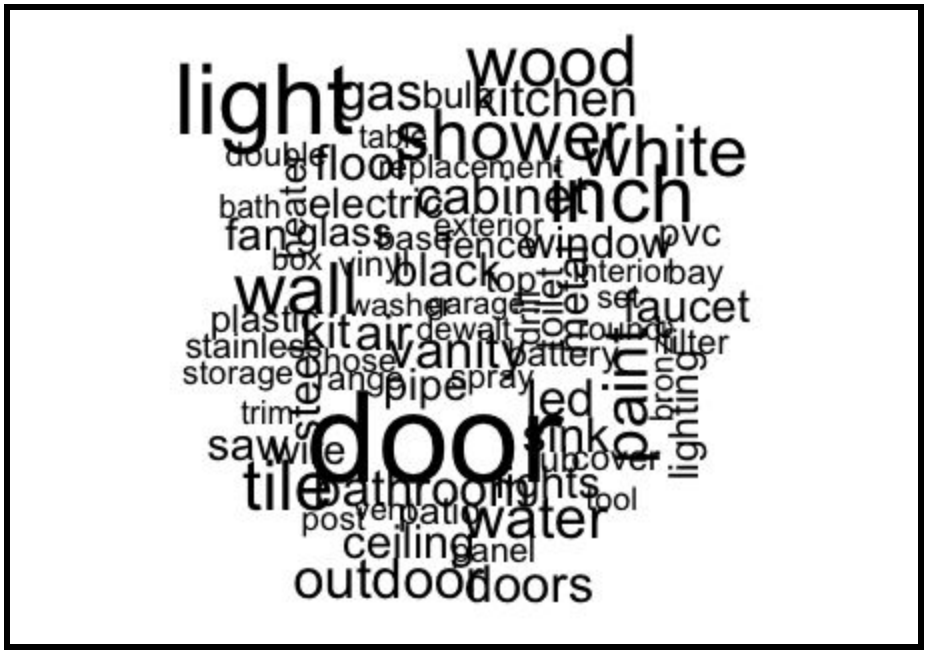
The data set is very large and we began with playing around with the data, like exploring the search terms and product data:

We references scripts on Kaggle and did a basic data analysis on our local machines.

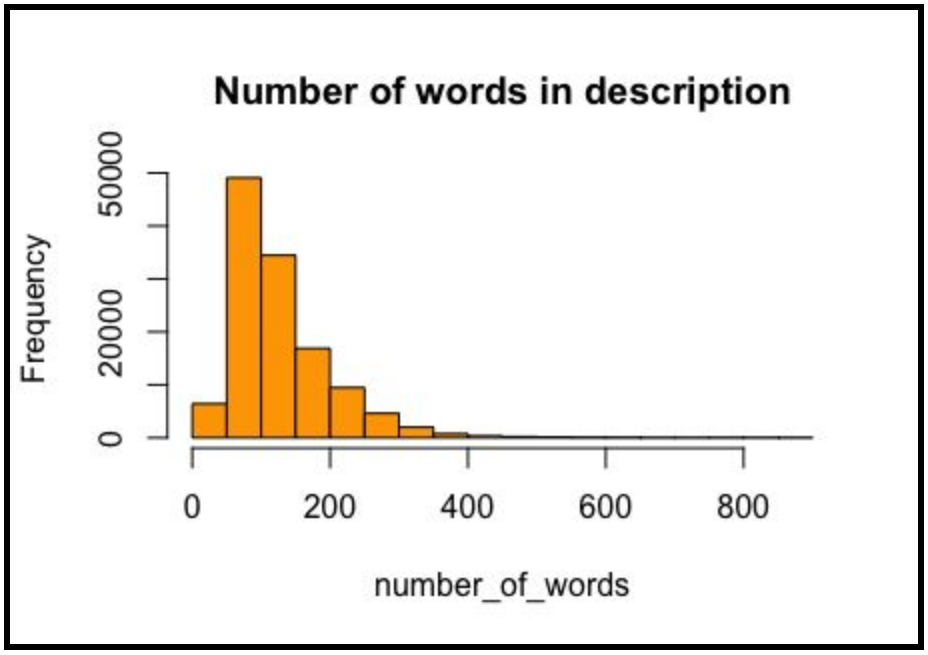




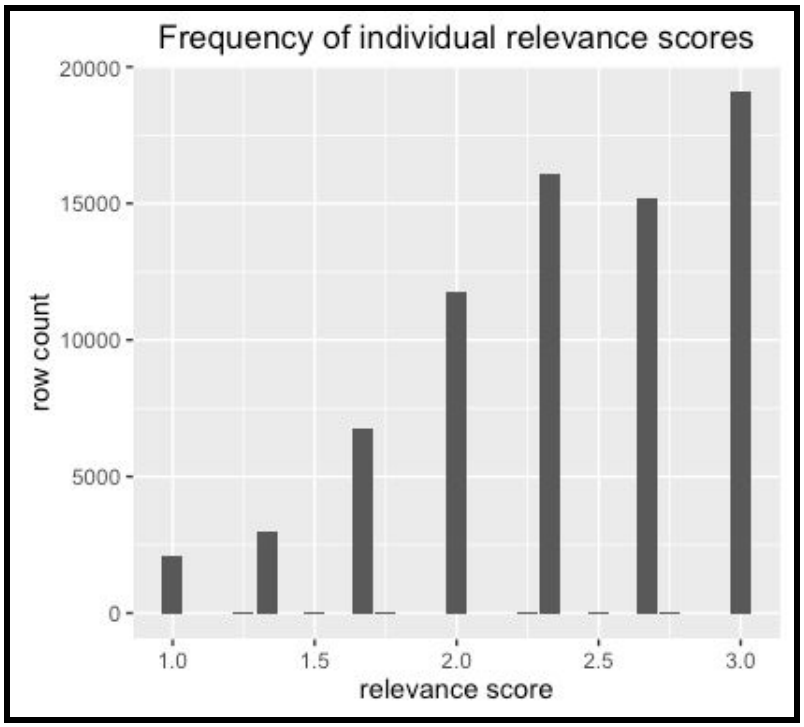
The frequency of words in train terms:



The frequency of Words in test terms:



The frequency of individual relevance scores in train data:



When we analyzed the search term data, we recognized some users misspelled words when they searching items on Home Depot website. This can affect the final relevance since these misspelled words cannot be matched. Then we replace the frequently appeared misspelled words with the correct ones.

III. How we will tackle this problem

As we have realized that some words are misspelled and due to the nature of the English language, stemming is inevitable for an effective text analysis. We will use the famous Snowball stemmer to detect roots of word and reduce noise. Then, we are going to build a term-document matrix and use term frequency - inverse document

frequency to address potential bias regarding longer search terms and some common word tokens like ‘of’ and ‘the’. In the next step, we are going to learn more about topic modelling like Latent Semantic Analysis (LSA), probabilistic latent semantic analysis (pLSA) and Latent Dirichlet Allocation (LDA) to apply on our project. After the model is built, we might try random forest and neural network to predict the relevance.

Reference:

1. Beginner Data Analysis. (n.d.). Retrieved March 14, 2016, from https://www.kaggle.com/dsoreo/home-depot-product-search-relevance/testing-r/notebook
2. Search Word Cloud. (n.d.). Retrieved March 14, 2016, from https://www.kaggle.com/kelexu/home-depot-product-search-relevance/word-cloud
3. Home Depot Data Exploration. (n.d.). Retrieved March 14, 2016, from https://www.kaggle.com/universome/home-depot-product-search-relevance/homedepot-data-exploration-forked/notebook