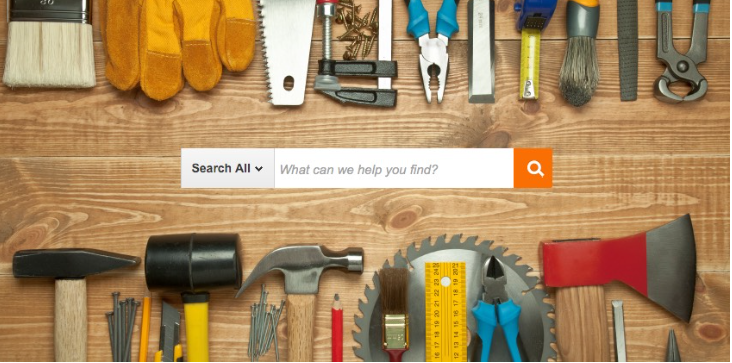
Home Depot Product Search Relevance



Author: Dingling Ge, Zefan Feng

Final Project: INFO 7374 Algorithmic Digital Marketing

Shoppers rely on Home Depot’s product authority to find and buy the latest products and to get timely solutions to their home improvement needs. From installing a new ceiling fan to remodeling an entire kitchen, with the click of a mouse or tap of the screen, customers expect the correct results to their queries – quickly. Speed, accuracy and delivering a frictionless customer experience are essential.

In our project, we would help Home Depot’s to improve their customers' shopping experience by developing a model that can accurately predict the relevance of search results.

Search relevancy is an implicit measure Home Depot uses to gauge how quickly they can get customers to the right products. Currently, human raters evaluate the impact of potential changes to their search algorithms, which is a slow and subjective process. By removing or minimizing human input in search relevance evaluation, Home Depot hopes to increase the number of iterations their team can perform on the current search algorithms.

1. Introduction

* The Goal of our project is to develop a model that can accurately predict the relevance of search results.
* This data set contains a number of products and real customer search terms from Home Depot's website. The challenge is to predict a relevance score for the provided combinations of search terms and products. To create the ground truth labels, Home Depot has crowdsourced the search/product pairs to multiple human raters.

1. Data

* The dataset is 74MB.

Data source: https://www.kaggle.com/c/home-depot-product-search-relevance/data

* File descriptions:

1. train.csv - the training set, contains products, searches, and relevance scores
2. test.csv - the test set, contains products and searches. You must predict the relevance for these pairs.
3. 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.
4. attributes.csv - provides extended information about a subset of the products (typically representing detailed technical specifications). Not every product will have attributes.
5. sample\_submission.csv - a file showing the correct submission format
6. relevance\_instructions.docx - the instructions provided to human raters

* Data field:

1. id - a unique Id field which represents a (search\_term, product\_uid) pair
2. product\_uid - an id for the products
3. product\_title - the product title
4. product\_description - the text description of the product (may contain HTML content)
5. search\_term - the search query
6. relevance - the average of the relevance ratings for a given id
7. name - an attribute name
8. value - the attribute's value
9. Project details

* Challenges and Goals for implementation

1. Text preprocessing

* lemmatization

1. Text Feature selection

* Levenshtein
* TF-iDF
* Word2Vec

1. Modelling

* Logistic Regression
* GBDT
* Deliverables

1. JupiterNotebook
2. Codelab report

* Timeline

|  |  |
| --- | --- |
| Time frame | deliverables |
| 12/6/2019 | Project proposal |
| 12/7 -- 12/9 | Data preprocessing |
| 12/9 --12/11 | Feature selection, modeling |
| 12/11 -- 12/13 | Testing, documentation |

1. References

Kaggle: https://www.kaggle.com/c/home-depot-product-search-relevance/data