G9: Home Depot Product Search Relevance

Rohan Sadale, Akshay Kulkarni, Utkarsh Kajaria, Gautham Sunder

Motivation

To enhance customer experience and engagement on online platforms it is imperative that the search results are highly relevant to the search terms. Proliferation of products sold online add to the complexity of recommending products consistent with the search term. The objective of the project is to improve shopping experience of Home Depot customers by developing a model that can accurately predict the relevance of search results.

Dataset

The data for this problem is hosted on Kaggle. The two datasets of interest are, training data of approximately 74000 search terms and its relevance score to a certain product. Relevance score for each search term is the average rating of 3 experts ranging between 1.0 to 3.0. 1.0 indicating no relevance and 3.0 indicating high relevance. In addition to the above data, second dataset describes the products along with its attributes.

Project Proposal

The continuous target variables ranging between 1.0 to 3.0 will be discretized into **Relevent** *(>=2.0)* and **Not Relevant** *(<2.0)* . Predicting the relevance of a search terms with products can now be treated as a classification problem.

The pre-processing steps entail working with text data. Thus, pre processing and feature extraction steps would extensively involve text cleaning approaches such as spell correction, case-conversion, stop-word removal, lemmatizing, stemming, cosine similarity, intesect count, word2vec, etc. We will also look for other ways to improve our preprocessing tasks (feature extraction) as we work through the project.

Our approach to modeling is to use Naive Bayes classifier as a benchmark and compare the performance against other classification and ensemble techniques. Apart from predicting the relevance of search term we also plan to derive implicit relations between products given a search term. This will help find other similar products related to the search term which might not be apparent from the similarity measures of the product. For e.g. The relation between a ladder and a hammer might not be obvious from their product descriptions in spite of the fact that a person who buys a hammer probably would be needing a ladder to climb. In addition, we plan to analyze the effect that different representations of data (bag-of-words vs n-gram vs ad hoc feature generation) have on the performance of the model.

Evaluation Plan

We plan to use around 50k examples for training, 10k for cross-validation and remaining 10k for testing the model. We plan to evaluate performance of our classifier on two criteria viz. accuracy and reducing number of false positives (result classified as relevant when it was irrelevant).

References

- Data Source:- https://www.kaggle.com/c/home-depot-product-search-relevance/data
- Pattern Recognition and Machine Learning by Chris Bishop.