Title: Home Depot Product Search Relevance

Group member:

- Kewen Zhang, Uni: kz2246
- Pengfei Wang, Uni: pw2406
- Xiaoci Xing, Uni: xx2203
- Ziyue Wu, Uni: zw2338

Abstract:

In this search relevance project, our goal was to build a model to predict the relevance of search items and product on homedepot.com, given the searching items, resulting product titles and product descriptions. Our team's solution relies heavily on feature extraction/selection and model ensembling.

Our solution consists of three parts:

Text cleaning

Before generating features, we have realized that it's reasonable to process the data. So we cleaned the data with spelling correction, synonym replacement, removing dots and stop words. Then we selected the optimal solution of N for each N-grams feature.

Feature extraction

We had tried three types of feature:

- counting features(which is not used in the final result)
- distance features (1~8 grams)
- TF-IDF features (1~8 grams)

Selection and model ensembling.

Model ensembling consisted of two main steps. Firstly, we trained model library using different models, different parameter settings, and different subsets of the features. Secondly, we generated ensemble submission from the possible ensemble selections. Performance was estimated using cross validation within the training set. We tried both classification and regression to compare our results.

- Neural Net
- General Linear Model
- Machine Learning Methods