Project Requirement: can be found on the link https://www.kaggle.com/c/home-depot-product-search-relevance/data

Kaggle Score: 0.70580

I used python programming language.

Did the following to the training data.

1. Read the train, description, attributes and test files into pandas dataframe.
2. Calculate the term frequency of all the product titles and calculate the cosine similarity between the search terms and its corresponding product titles.
3. Calculate the term frequency of all the product description and calculate the cosine similarity between training search terms and its corresponding product descriptions.
4. Calculate the term frequency of all the product attributes by grouping the name and value fields of the attribute sheet on product_uid and calculate the cosine similarity between the search terms and its corresponding product attributes.

Calculate the cosine similarity between the test search terms and its corresponding product title, description and attributes (again grouping the name and value fields of attributes sheet on product_uid).

Group the training data based on the relevance it belongs to. According to the training data, we have 13 different relevance classes. For each group, calculate the average of the calculated cosine similarity of product titles, calculate the average of the calculated cosine similarity of product descriptions and calculate the average of the calculated cosine similarity of product attributes with the training search term. By doing this, we will have calculated average points of 13 different relevance classes.

For each test data pair of cosine similarity of search term and product title, product description and product attributes, using Euclidean distance, find the nearest neighbor to the calculated average points of different relevance. In our case its 13 different points or relevance classes.

Assign the relevance of the point which is closest to the test (cosine similarity pair of titles, description and attributes) point.

The relevance of all the test records are written into results.csv file. This program runs in less than 10mins and I get RMSE of 0.70580.