Home Depot Product Search Relevance

Srivarsini Rangarajan

Chaitanya Krishna Sai Kosaraju

Sai Kolanupaka Meghana Bharadwaj

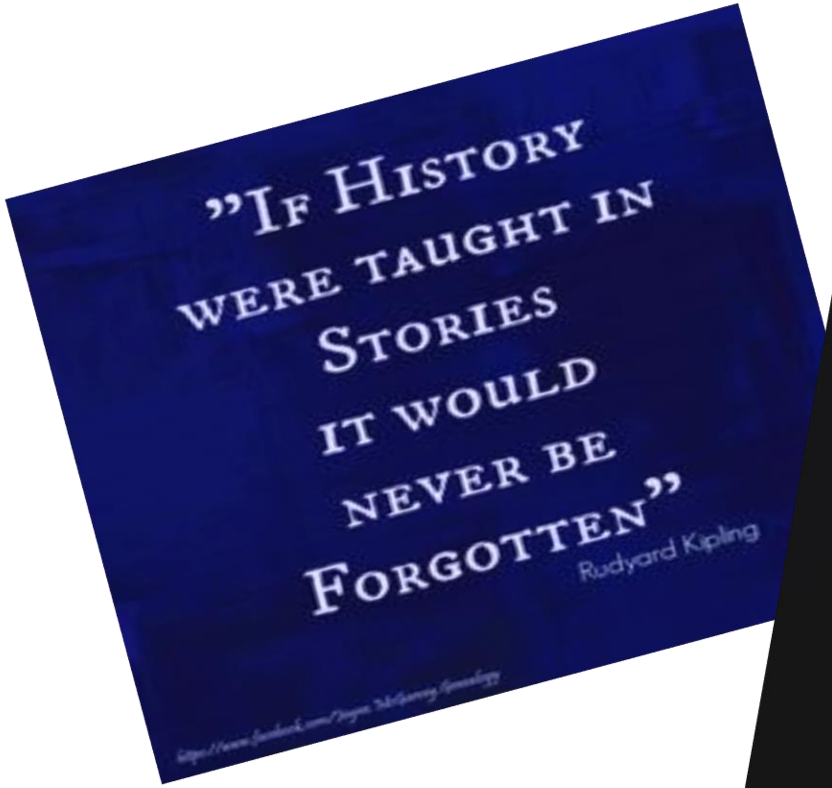
Online Shopping

- Great product at a great price
- Convenience
- Fewer traps
- Compulsive shopping
- Discreet shopping
- Crowds
- Variety

The Consumer Buying Process

www.shanepatrickjones.com

Information Retrieval



Identification of the Problem and Opportunity

- What is the problem you are solving?

  Prediction of relevance score for the given combination of search terms and products

- What are the benefits to solving this problem?
  - The relevance parameter enhances the search results.
  - Better Marketing for the organization.
  - Improving the customer's shopping experience.
- If you solve it, why is it important?
  - Its presence would help eliminate/ reduce human intervention in programming/ updating & maintaining the search engine.
  - It enhances the value of a product.
  - The vast amounts of data observed can be synthesized to seize and predict customer behavior patterns for further analysis.

Importance of this Problem

- Why is it important to solve this problem?
  - Without the use of information retrieval and ranking, the accuracy and efficiency for a match would be erroneous.
- What are you're *a priori* hypotheses?
  - To maximize the relevancy score prediction by reducing the Root Mean Square Error in our prediction model
- Who cares about it?
  - Industries involved – From Retail to travel
- If you were selling, who would buy your solution?
  - Kaggle and Home Depot

Prior Work

- What has been done in the past to solve this problem?
  - Nearest centroid classifier
  - Google - PageRank, Hilltop Algorithm, Topic-Sensitive PageRank
  - Text Scoring
  - Amazon – A9
- What has worked, what has not worked?
  - Information retrieval is striving to find the best fit but as the data grows the algorithm also need to be evolved.
  - Every search algorithm has its own advantages and disadvantages
- What did you like/dislike?
  - Text Scoring approach is the one that we liked and we started our analysis upon that
- What do you plan to leverage?
  - Take advantage of the existing tools like Python Machine Learning packages and other mining tools

Technical Approach

- What are the analytical steps you are going to use to build your result?

Classification and Supervised Learning

  - Pruning the data set – Python Stemming
  - Implement the search algorithm – TF-IDF
  - Create the frequency set for the terms –Python
  - Determine the relevancy factor
  - Result Visualization

Prediction using regression

- What are the questions you will answer?
  - Performance of the algorithm in terms of
    - Cross Validation Value
    - Root Mean Square Error
    - Execution time

Data

Files

train.csv

The training set, contains products, searches, and relevance scores.

test.csv

The test set, contains products and searches.

product_descriptions.csv

Contains a text description of each product.

attributes.csv

Provides extended information about a subset of the products (typically representing detailed technical specifications).

Snapshot of dataset
Train Data

id	product_uid	product_title	search_term	relevance
2	100001	Simpson Strong-Tie 12-Gauge Angle	angle bracket	3
3	100001	Simpson Strong-Tie 12-Gauge Angle	l bracket	2.5
9	100002	BEHR Premium Textured DeckOver 1-gal. #SC-141 Tugboat Wood and Concrete Coating	deck over	3
16	100005	Delta Vero 1-Handle Shower Only Faucet Trim Kit in Chrome (Valve Not Included)	rain shower head	2.33
17	100005	Delta Vero 1-Handle Shower Only Faucet Trim Kit in Chrome (Valve Not Included)	shower only faucet	2.67

Product Description

product_uid	product_description
100001	Not only do angles make joints stronger, they also provide more consistent, straight corners. Simpson Strong-Tie offers a wide variety of angles in various sizes and thicknesses to handle light-duty jobs or projects where a structural connection is needed. Some can be bent (skewed) to match the project. For outdoor projects or those where moisture is present, use our ZMAX zinc-coated connectors, which provide extra resistance against corrosion (look for a "Z" at the end of the model number).Versatile connector for various 90 connections and home repair projectsStronger than angled nailing or screw fastening aloneHelp ensure joints are consistently straight and strongDimensions: 3 in. x 3 in. x 1-1/2 in.Made from 12-Gauge steelGalvanized for extra corrosion resistanceInstall with 10d common nails or #9 x 1-1/2 in. Strong-Drive SD screws
100002	BEHR Premium Textured DECKOVER is an innovative solid color coating. It will bring your old, weathered wood or concrete back to life. The advanced 100% acrylic resin formula creates a durable coating for your tired and worn out deck, rejuvenating to a whole new look. For the best results, be sure to properly prepare the surface using other applicable BEHR products displayed above. California residents: see Proposition 65 informationRevives wood and composite decks, railings, porches and boat docks, also great for concrete pool decks, patios and sidewalks100% acrylic solid color coatingResists cracking and peeling and conceals splinters and cracks up to 1/4 in.Provides a durable, mildew resistant finishCovers up to 75 sq. ft. in 2 coats per gallonCreates a textured, slip-resistant finishFor best results, prepare with the appropriate BEHR product for your wood or concrete surfaceActual paint colors may vary from on-screen and printer representationsColors available to be tinted in most storesOnline Price includes Paint Care fee in the following states: CA, CO, CT, ME, MN, OR, RI, VT

Attributes

Product_id	Name	Value
100001	Product Height (in.)	3
100001	Product Weight (lb.)	0.26
100001	Product Width (in.)	3
100002	Application Method	Brush, Roller,Spray
100002	Assembled Depth (in.)	6.63 in
100002	Assembled Height (in.)	7.76 in

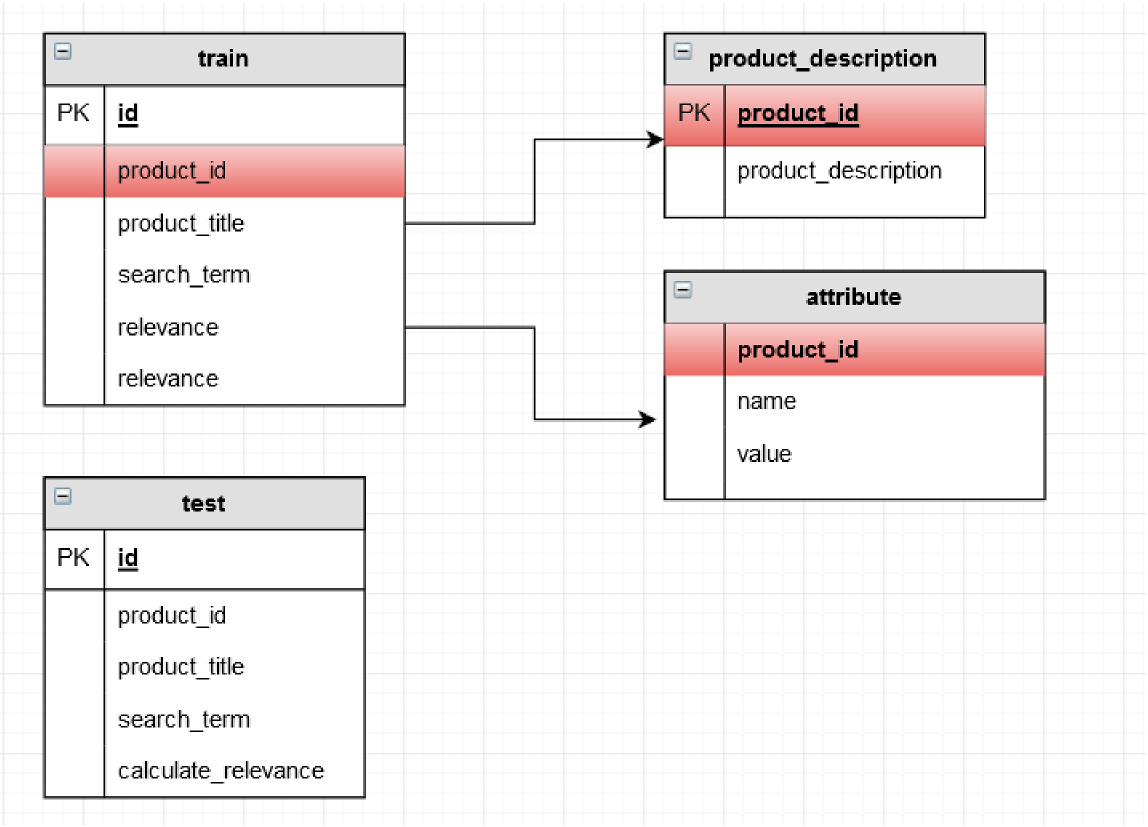
Test Data

id	product_uid	product_title	search_term	relevance
1	100001	Simpson Strong-Tie 12-Gauge Angle	90 degree bracket	?
4	100001	Simpson Strong-Tie 12-Gauge Angle	metal l brackets	?
5	100001	Simpson Strong-Tie 12-Gauge Angle	simpson sku able	?
6	100001	Simpson Strong-Tie 12-Gauge Angle	simpson strong ties	?
7	100001	Simpson Strong-Tie 12-Gauge Angle	simpson strong tie hcc668	?
8	100001	Simpson Strong-Tie 12-Gauge Angle	wood connectors	?
10	100003	STERLING Ensemble 33-1/4 in. x 60 in. x 75-1/4 in. Bath and Shower Kit with Right-Hand Drain in White	bath and shower kit	?
11	100003	STERLING Ensemble 33-1/4 in. x 60 in. x 75-1/4 in. Bath and Shower Kit with Right-Hand Drain in White	bath drain kit	?

Metrics

- What will you measure during your analysis?
  - Prediction accuracy of different algorithms
- How will you know when your are successful?
  - Cross Validation Score : 0 to 1
    - 0 being the best and 1 being the worst
  - Root Mean Square Error
    - 0 being the best and 1 being the worst
  - Execution time in seconds
    - Lower the better

Entity Relation



Algorithms used

- TF-IDF
- Support Vector Machine
- Decision Tree

TF-IDF

- Why TF-IDF?
  - Text search
  - Large text in the product description field
  - Familiar
- Consideration

  Calculating the relevance based on product description and product title alone

  Each word in search term as the word of interest and each record is considered as a document
- Approach
  1. Stemming of product description data
  2. Generated a frequency table for product description
  3. Create a dictionary containing search term for each search id
  4. Iterate over the frequency table to calculate the term frequency and Inverse document frequency

Challenges Faced…

- Generating the frequency for the huge dataset
- Once after creating the frequency table we were unable to interpret the algorithm for our data

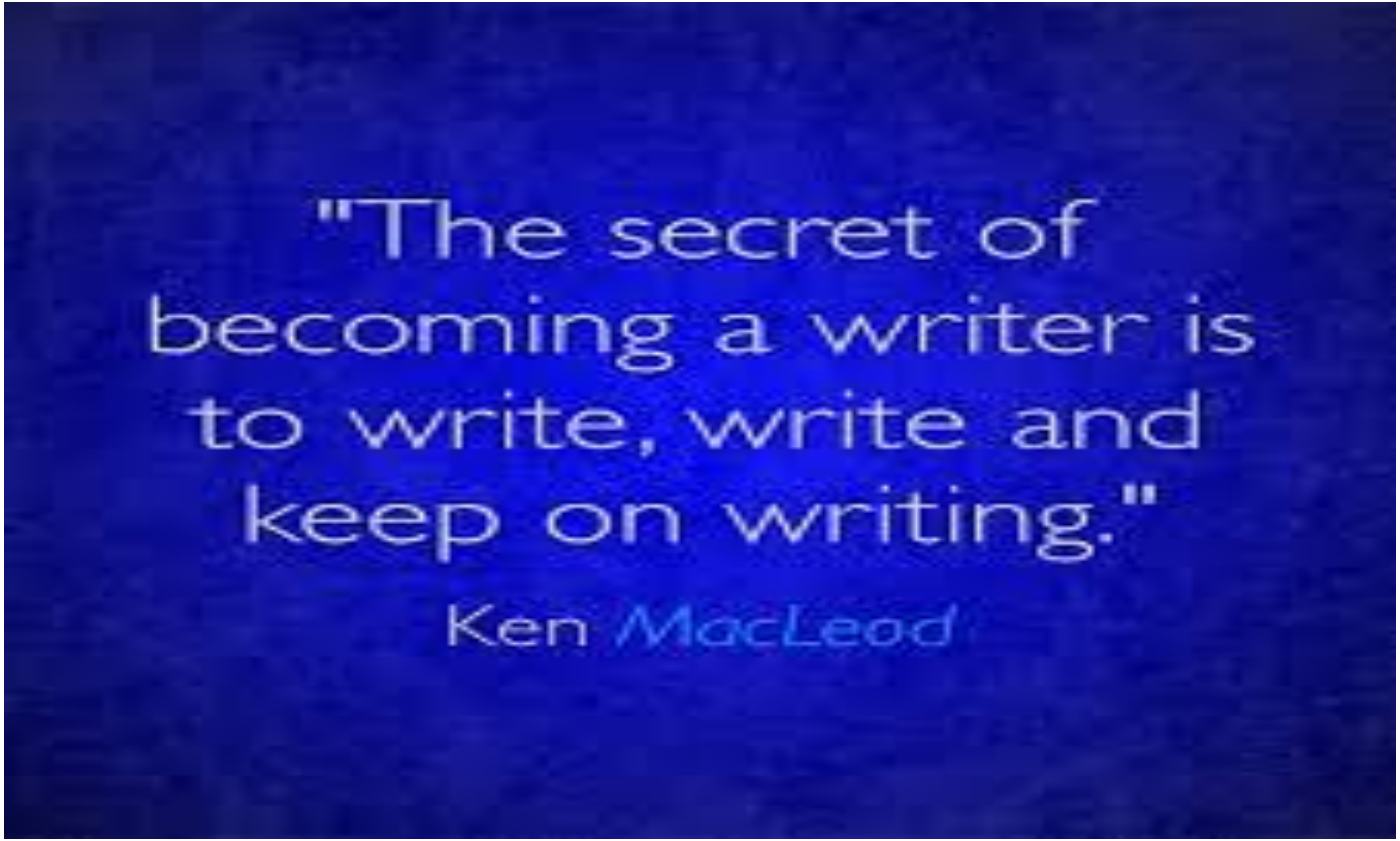
Solution:

Theoretical implementation of algorithm by taking a product and calculating the relevancy using the frequency table that we generated earlier to get an idea of how our results will look

Manual Example

- product prod_id = 1
- Search_id – 6(train) + 2(test)
- Search query – angle bracket
- Angle + bracket
- For each term in search query
  - Find tf-idf in product title
  - Find tf-idf in product description
- Finding tf-idf in product title
  - tf = count(angle) in record 1/total number of words in record 1
  - Idf = log[count(total_records)/1+No. of records containing angle]

After calculating it we got the relevance score of about 0.9 which is very low compared to the given relevancy score of 3.


"The secret of becoming a writer is to write, write and keep on writing."
Ken MacLeod

Support Vector Machine

Why SVC?

- Supervised learning algorithm
- Uses class labels and can be used for regression analysis as well
- Supported by sklearn in python

Consideration:

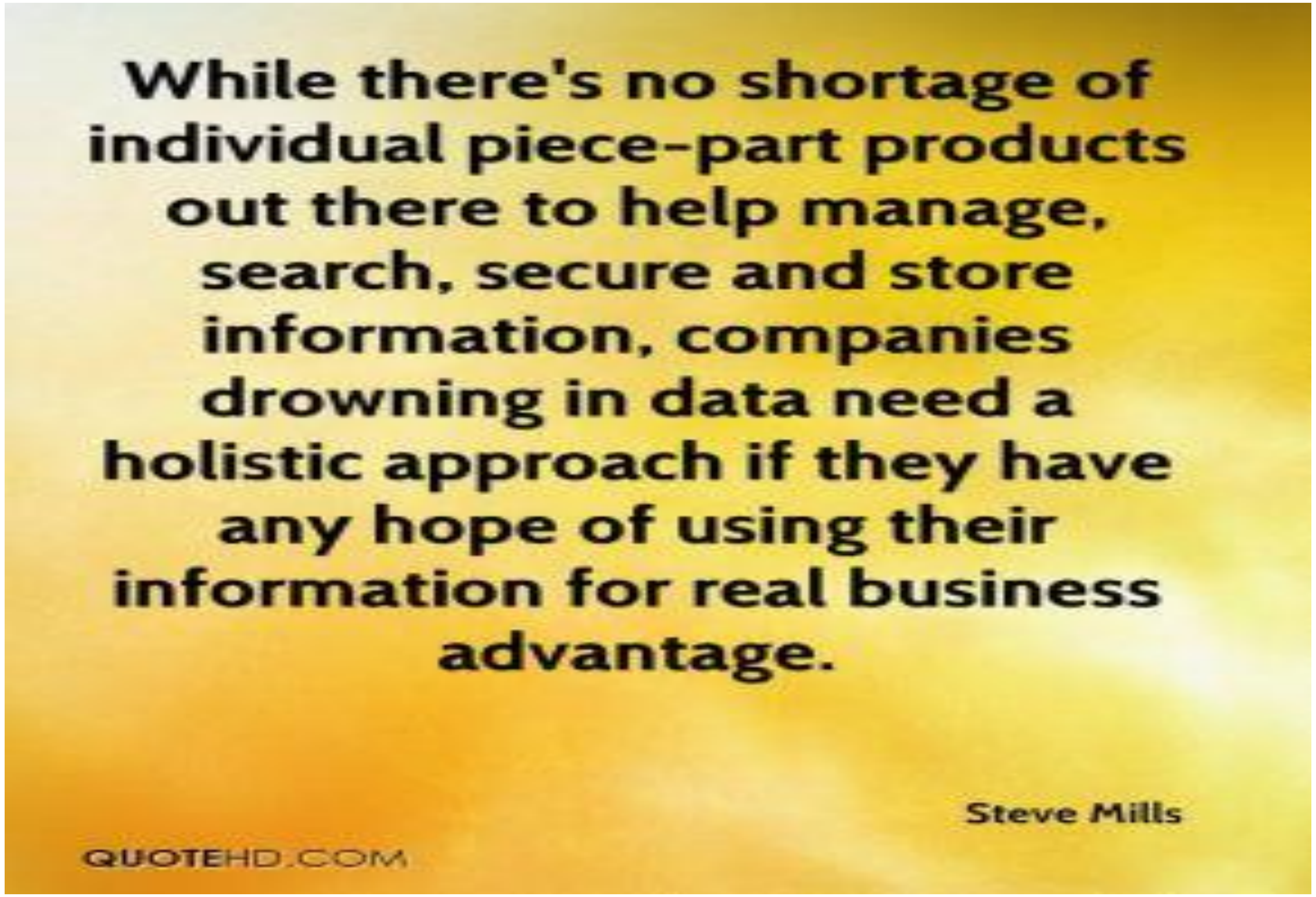
Executed only in a subset of around 1000 products

Approach

- Merge the all the available data into one data frame
- Try to figure out the search term frequency count in different labels such as title, description and attribute
- 5 numeric vectors – 4 frequency vector and 1 predictor
- Applied svm on training data and test data for 1000 products

Implemented SVM algorithm

- Import all the data into data frame
- Merge the data
  1. Eliminate all the missing values
  2. Combine the training and test data into one single frame
  3. Attach the product description column to it based on product id
  4. Apply stemming on all the columns containing string values – search_term, prod_title, prod_desc, prod_attr
  5. Find the length of the query
  6. Find term in title, description and attribute and count the frequency
  7. Merge the frequency columns to the existing data frame
- Create a data frame containing only the numeric labels – search-id, length of query, prod-id, term_description, term_attr, relevancy


While there's no shortage of individual piece-part products out there to help manage, search, secure and store information, companies drowning in data need a holistic approach if they have any hope of using their information for real business advantage.
Steve Mills
QUOTEHD.COM

Merged Data Frame

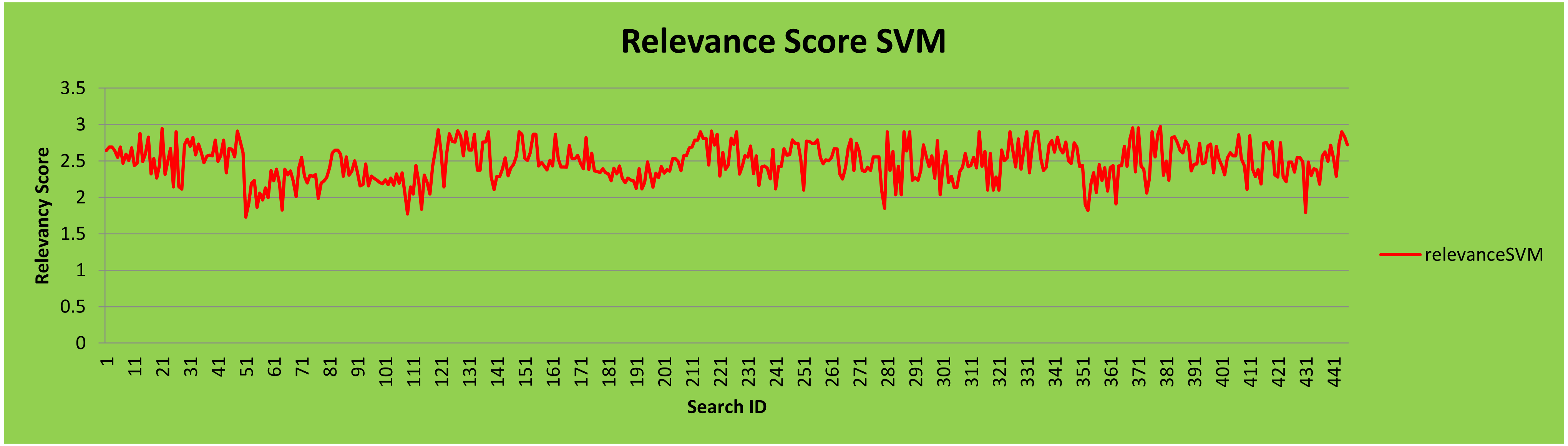
Search ID	Product ID	Relevancy	Query Length	Term Count Title	Term Count Description	Term Count Attribute
600	100095	3	3	3	3	2
605	100096	2	4	1	2	2
606	100096	2.67	3	1	1	1
611	100096	2.67	2	1	2	1
619	100097	2.33	6	4	5	4
626	100098	1.33	5	3	3	3
630	100100	3	2	2	2	2
1	100001	?	3	0	1	1
4	100001	?	3	1	1	1
5	100001	?	3	1	1	1
6	100001	?	3	2	2	2
7	100001	?	4	2	2	2

Algorithm continued…

- Using svm fit the relevance values based on the vectors search id, length of query, prod-id, term-title, term_description, term_attr
- Use the classifier to predict the relevancy score in the test dataset

Output

- Cross Validation Value – 0.7
- RMSE – 0.5771
- Execution Time – 31.622 seconds
- Graph



Decision Tree

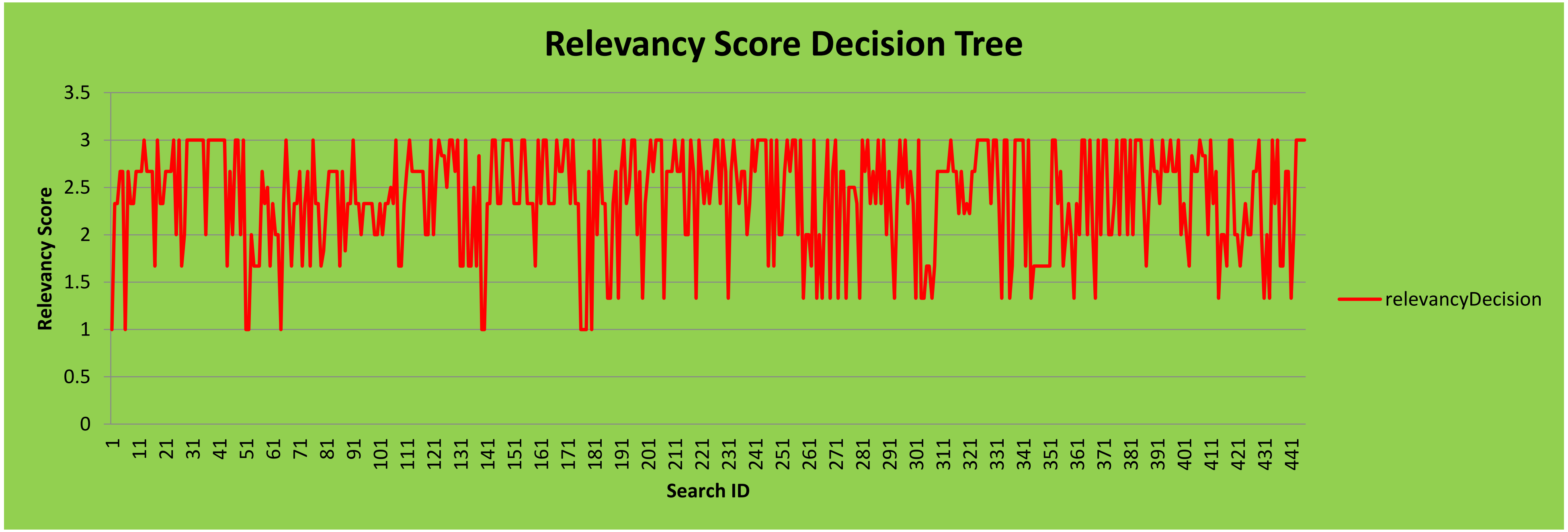
- Why Decision Tree?
  - Supervised learning algorithm
  - Familiar
  - Easy to interpret
  - It will work on our numeric vectors that we created for SVM
- Approach

  Apply decision tree classification using sklearn on the numeric vectors

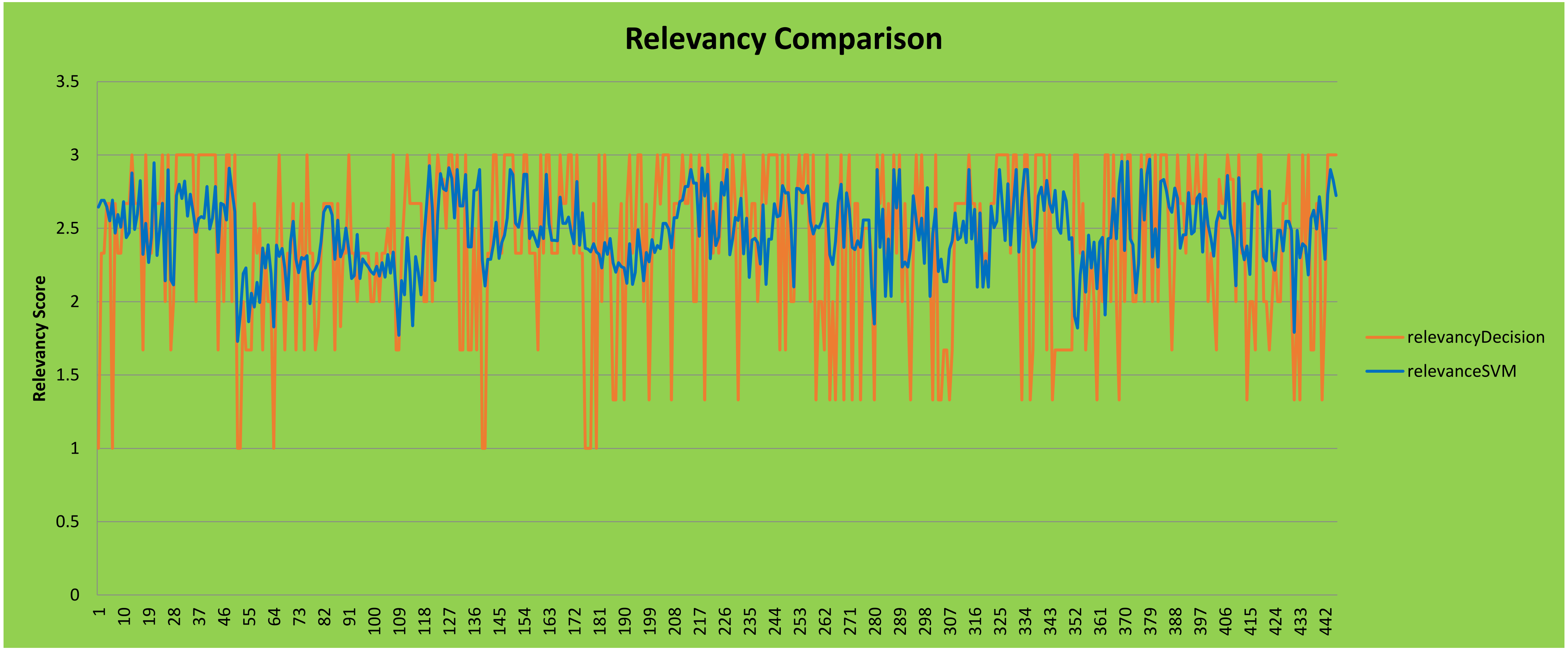
Output

- Cross Validation Value – 0.7
- RMSE – 0.8488
- Execution Time – 25.80 seconds

Graph:



Comparison of algorithms



Conclusions

- We found the SVM gives us better prediction than Decision tree
- Consideration of the entire data set may change the result
- Results could be improved by using features, brands, functionality of products while doing the classification
- Ensemble technique would bring even higher prediction accuracy.
- Relevancy Score generated will help the Home Depot Product Search team to maximize the customer satisfaction.

Questions and Suggestions…

Thank you!