BRANDON WONG – SG DAT 1

I. INTRODUCTION

II. APPROACH

II. CHALLENGES

III. NEXT STEPS

III. Q&A

INTRODUCTION

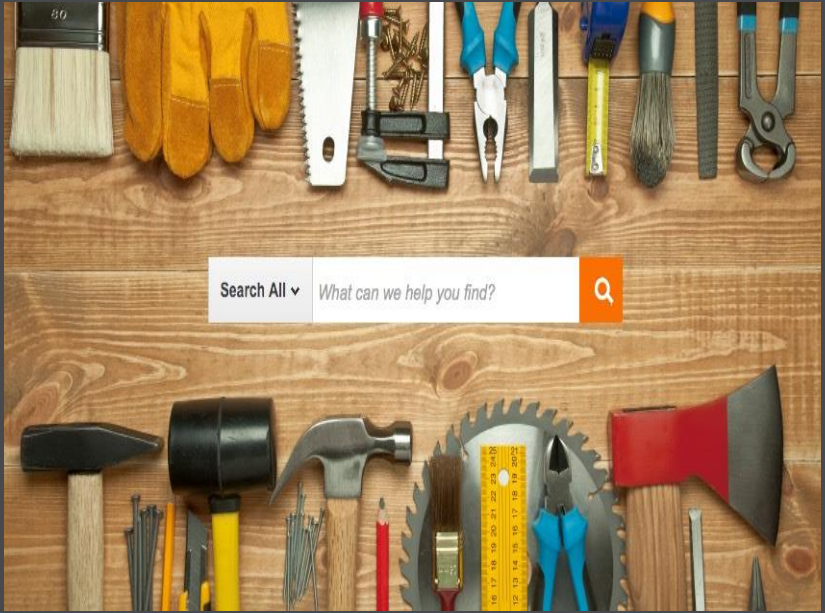
$40,000 *prize*

Timeline: 18/01/16 – 25/04/16 1,269 *teams*

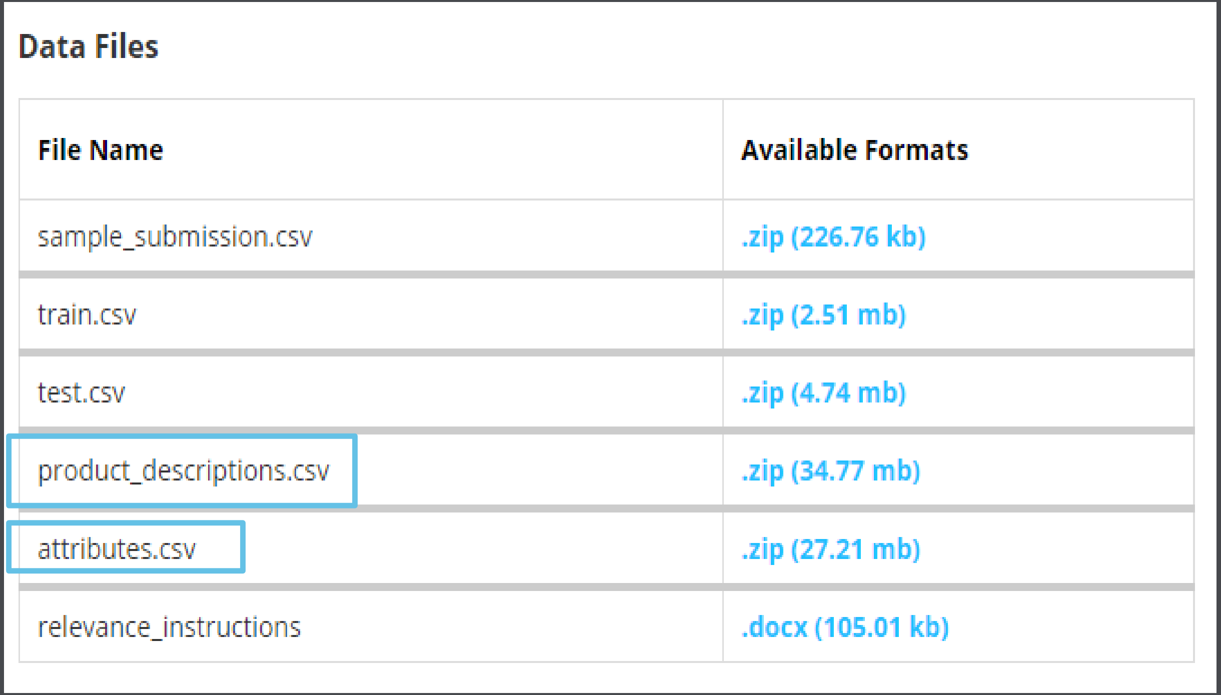
Predict Relevance of Search Results

Home Depot: Home Depot is an American retailer of home improvement and construction products and services

Competition Basics: Improve customer's shopping experience by developing a model that can accurately predict the relevance of search results



INTRODUCTION (CONT)



Data Files

File Name	Available Formats
sample_submission.csv	.zip (226.76 kb)
train.csv	.zip (2.51 mb)
test.csv	.zip (4.74 mb)
product_descriptions.csv	.zip (34.77 mb)
attributes.csv	.zip (27.21 mb)
relevance_instructions	.docx (105.01 kb)

Train and Test have similar columns BUT relevance score not provided in Test.

APPROACH

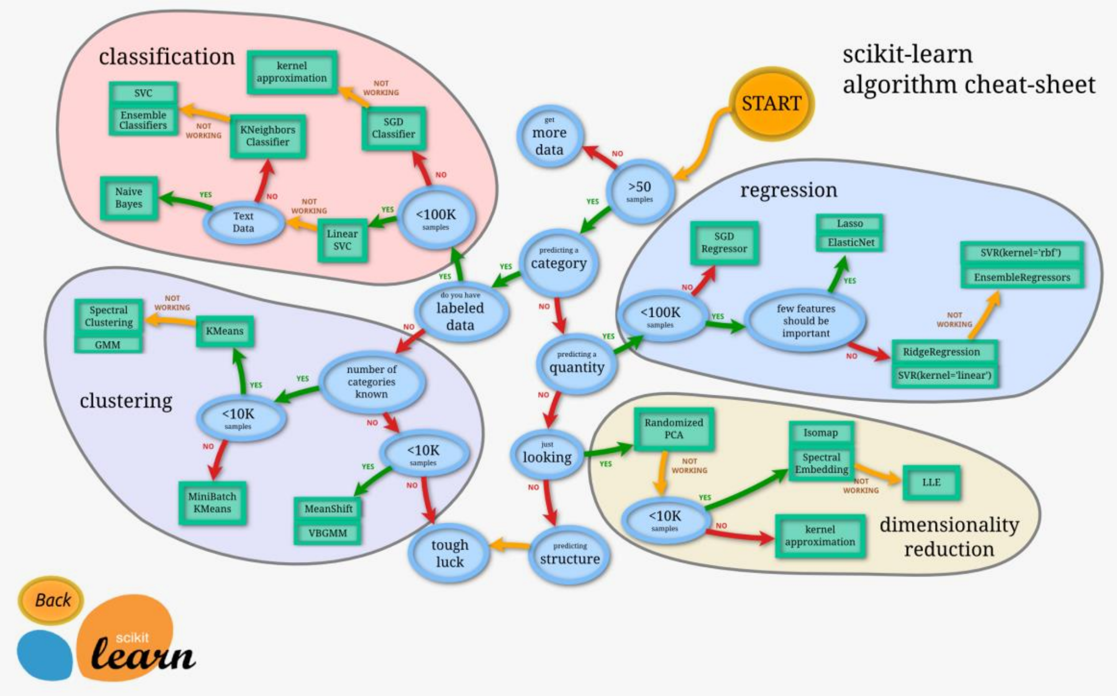
Picking a Model

I have:

- Labeled data (Classification?)
- Mid-scale volume, classifiers (Random Forest?)

I need:

- Root words / Stems (Snowball Stemmer / PyStemmer?)



APPROACH (CONT)

Relevance is a number between 1 (not relevant) and 3 (relevant)

E.g. Search for Steel Saw

Steel Saw (R = 3)
Steel Nails (R = 2)
Shovel (R = 1)

Each pair was *(search_term,product)* evaluated by at least 3 human raters.

The provided relevance scores are the average value of the ratings

CHALLENGES

1. Not trying to predict the true relevancy of the product as a response to a search query

2. Instead, build program to mimic human raters, assuming they are the most efficient method of assessing relevancy

3. Have to teach the models/machines to act like humans? Need to “create a search system auditor that can help measure the efficacy of changes in algorithms preferably in real time”

NEXT STEPS

1. Complete initial entry form and submit
2. Review the forums after competition is closed to see winning strategies
3. Keep competing in Kaggle competitions

Q&A