BRANDON WONG – SG DAT 1

I. INTRODUCTION
II. APPROACH
III. CHALLENGES
IV. NEXT STEPS
V. 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

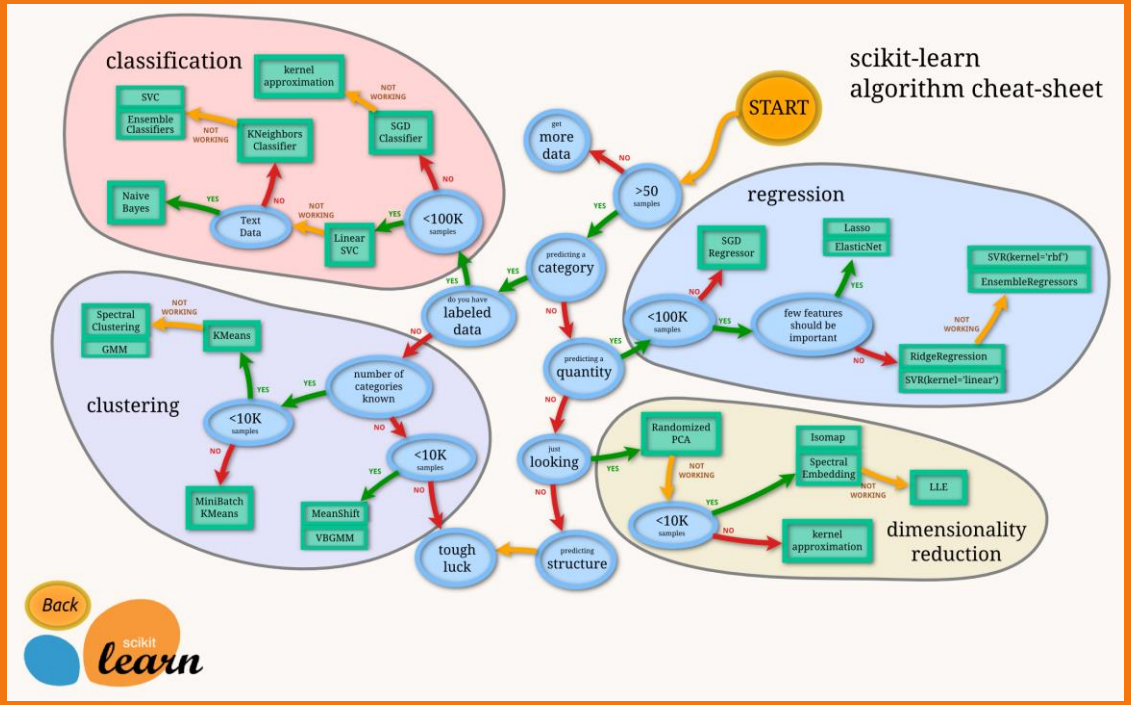
File Name	Available Formats
sample_submission.csv	.zip (226.76 kb)
train.csv	.zip (2.51 mb)
test.csv	.zip (4.74 mb)
product_descriptions.csv	.zip (34.77 mb)
attributes.csv	.zip (27.21 mb)
relevance_instructions	.docx (105.01 kb)

Train and Test have similar columns

BUT

relevance score not provided in Test.

APPROACH



Picking a Model

I have:

- Labeled data (Classification?)
- Not a huge volume, classifiers (Random Forest?)

I need:

- Root words / Stems (Snowball Stemmer / PyStemmer?)
- Rank variable importance

APPROACH (CONT)

id	product_uid	product_title	search_term	relevance
2	100001	Simpson Strong	angle bracket	3
3	100001	Simpson Strong	l bracket	2.5
9	100002	BEHR Premium	deck over	3
16	100005	Delta Vero 1-Ha	rain shower head	2.33
17	100005	Delta Vero 1-Ha	shower only fauce	2.67
18	100006	Whirlpool 1.9 cu	convection otr	3
20	100006	Whirlpool 1.9 cu	microwave over s	2.67
21	100006	Whirlpool 1.9 cu	microwaves	3
23	100007	Lithonia Lightin	emergency light	2.67
27	100009	House of Fara 3,	mdf 3/4	3

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

APPROACH (CONT)

Sample text: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

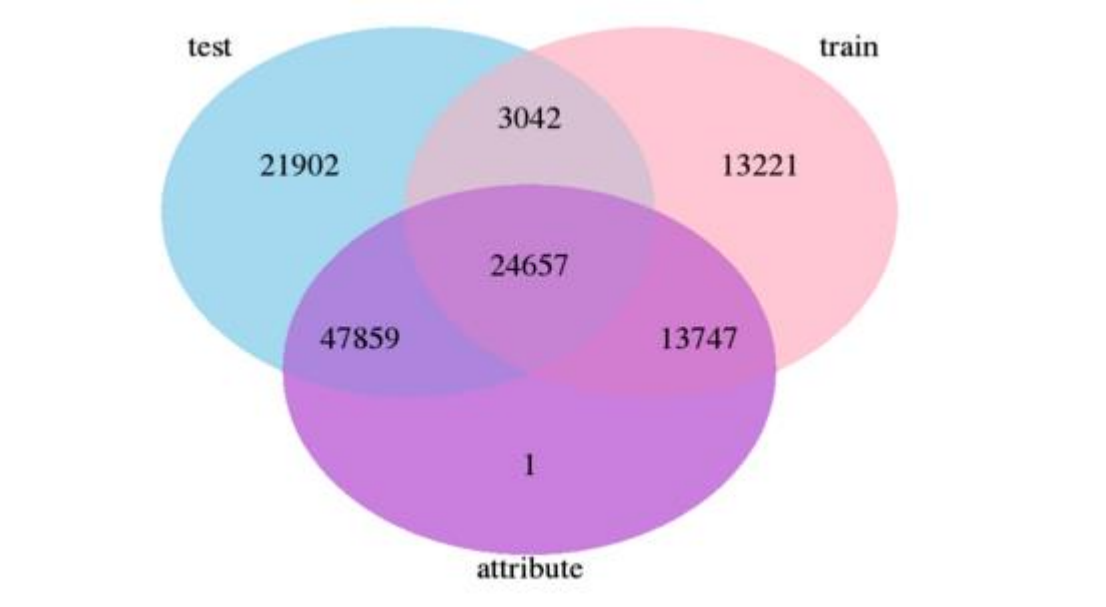
Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

Why Stemmer?

Messy!

Product Names
vs
Search Terms

APPROACH



There is one value in attributes.csv file that is not in either train or test files.

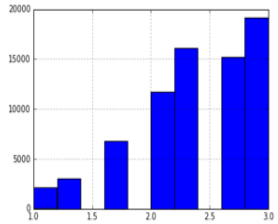
On further investigation there are 155 rows that do not have a *product_uid* value

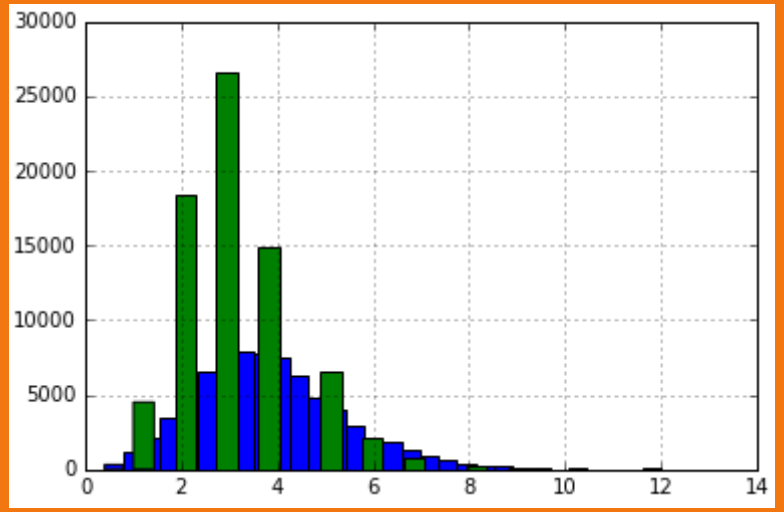
APPROACH (CONT)

Relevance Counts

```
In [7]: # Relevance counts
        training_data.relevance.hist()
        training_data.relevance.value_counts()

Out[7]: 3.00    19125
        2.33    16060
        2.67    15202
        2.00    11730
        1.67     6780
        1.33     3006
        1.00     2105
        2.50       19
        2.25       11
        2.75       11
        1.75        9
        1.50        5
        1.25        4
        Name: relevance, dtype: int64
```





Search Terms Dist (binned)

APPROACH (CONT)

Playing around, yay or nah?

	id	product_uid	product_title	search_term	relevance	on_point
0	2	100001	Simpson Strong-Tie 12-Gauge Angle	angle bracket	3.00	yay
1	3	100001	Simpson Strong-Tie 12-Gauge Angle	l bracket	2.50	yay
2	9	100002	BEHR Premium Textured DeckOver 1-gal. #SC-141 ...	deck over	3.00	yay
3	16	100005	Delta Vero 1-Handle Shower Only Faucet Trim Ki...	rain shower head	2.33	nah
4	17	100005	Delta Vero 1-Handle Shower Only Faucet Trim Ki...	shower only faucet	2.67	yay
5	18	100006	Whirlpool 1.9 cu. ft. Over the Range Convectio...	convection otr	3.00	yay
6	20	100006	Whirlpool 1.9 cu. ft. Over the Range Convectio...	microwave over stove	2.67	yay
7	21	100006	Whirlpool 1.9 cu. ft. Over the Range Convectio...	microwaves	3.00	yay
8	23	100007	Lithonia Lighting Quantum 2-Light Black LED Em...	emergency light	2.67	yay
9	27	100009	House of Fara 3/4 in. x 3 in. x 8 ft. MDF Flut...	mdf 3/4	3.00	yay

APPROACH (CONT)

```
MultinomialNB
Accuracy: 66.37%
Accuracy on training data: 0.76

BernoulliNB
Accuracy: 66.38%
Accuracy on training data: 0.76

Logistic Regression
Accuracy: 67.65%
Accuracy on training data: 0.77

"microwaves" is judged by clasifier to be...
... on point.
"what am I typing" is judged by clasifier to be...
... not on point.
"deck over" is judged by clasifier to be...
... on point.
```

APPROACH (CONT)

```
df_all['search_term'] = df_all['search_term'].map(lambda x:str_stemmer(x))
df_all['product_title'] = df_all['product_title'].map(lambda x:str_stemmer(x))
df_all['product_description'] = df_all['product_description'].map(lambda x:str_stemmer(x))

df_all['len_of_query'] = df_all['search_term'].map(lambda x:len(x.split())).astype(np.int64)

df_all['product_info'] = df_all['search_term']+"\t"+df_all['product_title']+"\t"+df_all['product_description']

df_all['word_in_title'] = df_all['product_info'].map(lambda x:str_common_word(x.split('\t')[0],x.split('\t')[1]))
df_all['word_in_description'] = df_all['product_info'].map(lambda x:str_common_word(x.split('\t')[0],x.split('\t')[2]))
```

```
rf = RandomForestRegressor(n_estimators=20, max_depth=7, random_state=0)
clf = BaggingRegressor(rf, n_estimators=50, max_samples=0.1, random_state=25)
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
```

LEADERBOARD

#	Δ1w	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	Turing test *	0.44014	113	Tue, 29 Mar 2016 16:06:47 (-4d)
2	—	. *	0.44222	69	Fri, 01 Apr 2016 07:00:18 (-29.9h)
3	↑11	Alex&Andreas&Nurlan *	0.44268	79	Thu, 31 Mar 2016 20:32:12
954	↓51	Ganapriya Kalavagunta	0.48720	1	Sun, 13 Mar 2016 16:38:06
955	new	**Brandon Wong**	**0.48721**	**2**	**Fri, 01 Apr 2016 19:15:11**
956	↓52	Iqbal Hossain	0.48721	7	Fri, 12 Feb 2016 17:01:41 (-6.1d)

Your Best Entry ↑
You improved on your best score by 0.00000.

You just moved up 87 positions on the leaderboard. Tweet this!

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. Check out winning strategies on Kaggle, improve on them
2. Keep competing in Kaggle competitions
3. Try out Natural Language Processing