Information Retrieval Tool Brief

Overview and Objective

- With the recent pandemic taking place, there has been a boom in online shopping. This has created more of a need now than ever for ways to shop comfortably online. Many bad shopping experiences come from the difficulty in finding the right products. To solve this, a search engine must be created to match user search queries to products efficiently.

- In this presentation, we describe the design and implementation of a search engine capable of returning a list of products, given a query

Outline

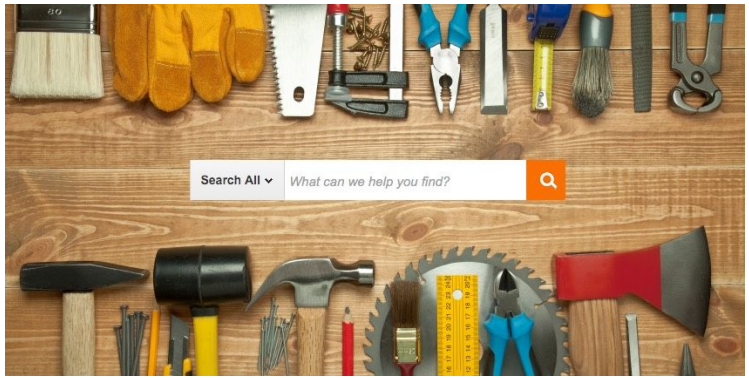
- Retrieval task and motivation
  - Problem Statement and Approach
  - Dataset & Preprocessing
- Search engine design
  - The Architecture
  - The Framework and Tools
  - BM25 & DFR Lucene API Implementation
- Evaluation
  - Preprocessing the relevance data
  - Evaluation method
  - Evaluation metrics
- Live demo
- Conclusion and recommendations
- Group Management
- Bibliography

Retrieval task and motivation

Task: Product search

Scenario: Home Depot website search

Motivation: Return the most relevant list of products given a query (or information need)



Problem Statement and Approach

- Home Depot Product Search Relevance dataset:
  - There are various datasets online that involve products, however the home depot product search relevance dataset has a large online catalogue where each query being given a relevance score between 1 (not relevant) and 3 (very relevant).
- Models analysed:
  - BM25 (default)
  - BM25 (k1 = 1.0, b = 0.0)
  - BM25 (k1 = 2.0, b = 1.0)
  - DFR
- Evaluation Criteria:
  - MAP
  - Precision
  - Recall
  - F1-Score

Dataset

In the dataset there are a number of products and real customer searches from Home Depot's website.

- **Format:** text-only (no images etc)
- **Number of documents (or products):** 74068
  - Product title
  - Product description
- **Number of query terms:** 7741

For each search term roughly anywhere between 5-10 documents (or products) have been given an average relevance rank by human annotators spanning from **1 (not relevant)** to **3 (highly relevant)**

Problem Statement and Approach

Problem statement: The home depot catalogue features tens of thousand of products. How can we return the most relevant (e.g. 10) products from this huge list given a query?

Approach:

1) Preprocess the product ID data
2) Index documents via elasticsearch
3) Select retrieval model(s)
4) Iterate through search query terms from the data (roughly 7.5k)
5) Compare the search engine results to the human ranked relevance data and compute evaluation metrics (MAP, precision, recall, etc)
6) Adjust model as needed to optimise performance

Preprocessing

- **Case folding** - lowercasing all text documents (avoid inconsistencies when using search queries)
- **Stopword removal** - filtering out words that do not positively contribute to relevancy, and also saving time/space during indexing of documents
- **Punctuation removal** - simplify information retrieval of text documents (Mahmud 2013).
- **Lemmatization** - gain better understanding of user queries and returning most relevant results. Can improve recall of searches - associating more documents with the same query word(s) i.e. root (Lane *et al.*, 2019)

Example search term **BEFORE** preprocessing

	id	product_uid	product_title	search_term
0	2	100001	Simpson Strong-Tie 12-Gauge Angle	angle bracket
1212	3784	100664	Simpson Strong-Tie 12-Gauge Angle	angle bracket
1229	3826	100672	Everbilt 1-1/2 in. Zinc-Plated Corner Brace (4...	angle bracket
1402	4319	100739	Everbilt 1 in. Zinc-Plated Corner Brace (20-Pack)	angle bracket
1953	6030	101036	Crown Bolt 1 in. x 72 in. Plain Steel Angle wi...	angle bracket
2624	8105	101370	Everbilt 3 in. Zinc-Plated Corner Brace (4-Pack)	angle bracket
2740	8474	101440	OWT Ornamental Wood Ties 45-Degree Flush Insid...	angle bracket
2895	8989	101534	Simpson Strong-Tie ZMAX 18-Gauge Galvanized St...	angle bracket
4108	12829	102239	Simpson Strong-Tie 16-Gauge Concrete Form Angle	angle bracket
5818	18280	103250	Simpson Strong-Tie Z-MAX 2 in. x 4 in. 12-Gaug...	angle bracket
5829	18323	103262	Superstrut 2-Hole 90è_ Angle Bracket - Silver ...	angle bracket

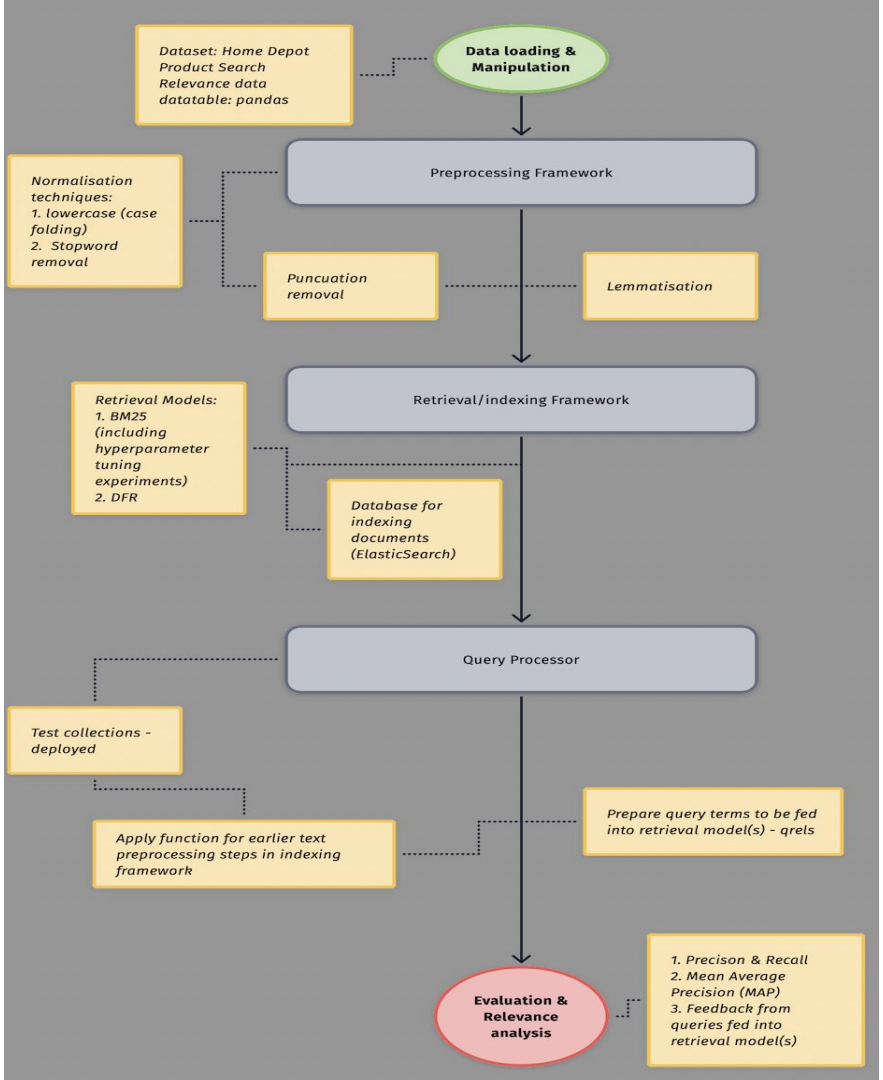
Example search term **AFTER** preprocessing

:

	id	product_uid	product_title	search_term
0	2	100001	simpson strongtie 12gauge angle	angle bracket
1212	3784	100664	simpson strongtie 12gauge angle	angle bracket
1229	3826	100672	everbilt 112 zincplated corner brace 4pack	angle bracket
1402	4319	100739	everbilt 1 zincplated corner brace 20pack	angle bracket
1953	6030	101036	crown bolt 1 x 72 plain steel angle 18 thick	angle bracket
2624	8105	101370	everbilt 3 zincplated corner brace 4pack	angle bracket
2740	8474	101440	owt ornamental wood tie 45degree flush inside ...	angle bracket
2895	8989	101534	simpson strongtie zmax 18gauge galvanized stee...	angle bracket
4108	12829	102239	simpson strongtie 16gauge concrete form angle	angle bracket
5818	18280	103250	simpson strongtie zmax 2 x 4 12gauge galvanize...	angle bracket
5829	18323	103262	superstrut 2hole 90è angle bracket silver galv...	angle bracket

Search Engine Design:

The Architecture



The Framework and Tools

Tool/API/Framework	Description
Jupyter/Colab Notebooks	IDE used for the search engine
Python	For the implementation of search engine components
Pandas	Stores the dataset
NLTK, re	Used for text preprocessing of the documents (text fields) and the search queries
Elasticsearch (python wrapper)	Deployment of the search engine i.e. database for indexing documents

Table 1: Tools and frameworks used for preprocessing and handling the Home Depot product text collections

BM25 & DFR Lucene API Implementation

- BM25 hyperparameters k1 and b were tuned:
  - K1 = 1.0, b = 0.0 (no document length normalization)
  - K1 = 1.2, b = 0.75 (default)
  - K1 = 2.0, b = 1.0 (full document length normalisation)
- DFR has components based on the Gianni Amati and Cornelis Joost Van Rijesbergen. 2002 paper
  - Geometric approximation of Bose-Einstein

Preprocessing the relevance data for evaluation

Conversion to binary score

- Scores < 2 become 0 i.e. non-relevant
- Scores ≥ 2 become 1 i.e. relevant

Qrel data tuple:

- (product_id, search term, relevance score)
- Examples:

  (183204, ‘angle bracket’, 1)

  (102402, ‘angle bracket’, 0)

  … *(total of 10)*

Evaluation

Search **scores** for extracted from the search results referencing the query and the product_id, which was a proxy for document ID. This was in the format: (query, product_id, **score**)

For the **query relevance** (qrel) ratings these were extracted in the format (query, produce_id, **qrel**

- Qrels (0-3) were then converted to a binary score where less than 2 is not relevant (0), and more than 2 is relevant (1)

To produce the evaluation, product_ids returned in k search results (predicted to be relevant, and with the highest score) were compared with their actual relevance (ground truth, qrel).

This allowed the number of false negatives, false positives, true positives,
and true negatives to be calculated, as well as **precision and recall**

Evaluation

Step 1:

Elasticsearch returns a ranked list of k documents.

Step 2 :

Compare the ranked list of documents to the (human assessed) relevant documents

Step 1

```
### RESULTS ####
124591 vigo 6jet shower panel system rain shower head handshower
120585 vigo 6jet shower panel system stainless steel
108854 melnor gentle rain shower head watering wand
128580 moen 16 overhead shower arm brushed nickel
110273 american standard ceiling mount 6 shower arm escutcheon s
112974 american standard ceiling mount 6 shower arm escutcheon p
100331 pulse showerspas kauai ii brushed nickel shower system br
113276 delta ara 1handle shower faucet trim kit chrome le shower
102822                                                    finish
102843                                                    ome fin
```

```
query rain shower head
True positive 0
False positive 0
False negative 10
```

Step 2

product_u			levance
1000			2.33
100331	pulse showerspas kauai ii brushed nickel showe...	rain shower head	2.67
100616	moen halo 3spray 9 rainshower showerhead chrome	rain shower head	3.00
101126	glacier bay 1spray 8 square showerhead chrome	rain shower head	3.00
101287	grohe powerampsoul cosmopolitan 4spray 712 sho...	rain shower head	3.00
102542	kohler forte singlefunction 1spray katalyst sh...	rain shower head	3.00
102654	moen ignite 5spray 9 rainshower showerhead chrome	rain shower head	3.00
102941	delta 1spray 8 overhead raincan shower head ch...	rain shower head	2.67

Evaluation verification

```
#output from df without duplicates removed for a search term
df.loc[df['search_term'] == 'rain shower head']
```

	id	product_uid	product_title	search_term	relevance	product_description
3	16	100005	delta vero 1handle shower faucet trim kit chro...	rain shower head	2.33	update bathroom delta vero singlehandle shower...
583	1870	100331	pulse showerspas kauai ii brushed nickel showe...	rain shower head	2.67	kauai rain shower system brilliantly simple de...
1124	3517	100616	moen halo 3spray 9 rainshower showerhead chrome	rain shower head	3.00	customize bathing experience moen halo 3spray ...
2120	6541	101126	glacier bay 1spray 8 square showerhead chrome	rain shower head	3.00	glacier bay 1spray 8 square showerhead chrome ...
2477	7615	101287	grohe powerampsoul cosmopolitan 4spray 712 sho...	rain shower head	3.00	grohe powersoul cosmopolitan shower created fu...
4652	14566	102542	kohler forte singlefunction 1spray katalyst sh...	rain shower head	3.00	forte singlefunction showerhead brings innovat...
4849	15205	102654	moen ignite 5spray 9 rainshower showerhead chrome	rain shower head	3.00	subtle elegant detail ignite collection create...
5325	16674	102941	delta 1spray 8 overhead raincan shower head ch...	rain shower head	2.67	create spalike sanctuary bathroom delta arzo 1...

```
query rain shower head
True positive 0
False positive 0
False negative 10
```

```
### RESULTS ####
124591 vigo 6jet shower panel system rain shower head handshower stainless steel
120585 vigo 6jet shower panel system stainless steel
108854 melnor gentle rain shower head watering wand
128580 moen 16 overhead shower arm brushed nickel
110273 american standard ceiling mount 6 shower arm escutcheon satin nickel
112974 american standard ceiling mount 6 shower arm escutcheon polished chrome
100331 pulse showerspas kauai ii brushed nickel shower system brushed nickel
113276 delta ara 1handle shower faucet trim kit chrome le shower head valve showerhead included
102822 pulse showerspas lanikai 3jet shower system chrome finish
102843 pulse showerspas kauai ii chrome shower system chrome finish
```

Evaluation metrics- queries 0-5

Query	TP	FP	FN	P	R	F1	TP	FP	FN	P	R	F1
angle bracket	0	0	10	0	0	0	0	0	10	0	0	0
deck over	0	0	6	0	0	0	0	0	6	0	0	0
rain shower head	0	0	8	0	0	0	0	0	7	0	0	0
convection otr	0	0	7	0	0	0	0	0	9	0	0	0
emergency light	0	0	9	0	0	0	0	0	0	0	0	0

Key: BM25- Default, DFR

TP: True Positive; FP: False Positive; FN: False Negative; P: Precision; R: Recall; F1: F1 Score

Evaluation metrics- queries 5-10

Query	TP	FP	FN	P	R	F1	TP	FP	FN	P	R	F1
mdf 34	0	0	0	0	0	0	0	0	0	0	0	0
steele stake	2	0	5	1.00	0.29	0.44	2	0	5	1.00	0.29	0.44
briggs and stratton lawn mower	1	0	2	1.00	0.33	0.5	0	0	3	0	0	0
hampton bay chestnut pull up shade	0	0	2	0	0	0	0	0	2	0	0	0
disposer	2	2	4	0.50	0.33	0.40	1	1	5	0.50	0.17	0.25

Key: **BM25- Default**, **DFR**

TP: True Positive; FP: False Positive; FN: False Negative; P: Precision; R: Recall; F1: F1 Score

Evaluation metrics

Model	MAP
BM25 (default)	0.36
BM25 (k1 = 1.0, b =0.0)	0.32
BM25 (k1 = 2.0, b = 1.0)	0.33
DFR	0.29

Table 2: MAP comparison of retrieval Model against text collections

Rounded to 2 decimal places

Evaluation strategy inspired in line with (Cormack and Lynam, 2006)

- Take the Mean Average Precision (MAP) @ 10 results set. Includes **185 qrels**
- Consider the best BM25 model compared with all various parameter tunings
- DFR vs. best BM25 model for baseline comparisons for a given set of 10 queries
  - Comparison using:
    - Precision
    - Recall
    - F1-score

Conclusion and recommendations

- Initially the retrieval models we intended to use were BM25 and BM25F, however we ended up using replacing BM25F with DFR
- It would have been helpful to include a meaningful way of evaluating the given relevance scores by the human annotators (as opposed to the binary)
- If given more time - we could have experimented with vector space models (TF-IDF) to see performance of search alternatives using alternative methods like cosine similarity

Bibliography

- Amati, G. and Van Rijsbergen, C. J. (2002) 'Probabilistic models of information retrieval based on measuring the divergence from randomness', *ACM Transactions on Information Systems*, 20(4), pp. 357–389. doi: 10.1145/582415.582416.
- Cormack, G. V. and Lynam, T. R. (2006) 'Statistical precision of information retrieval evaluation', in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06. the 29th annual international ACM SIGIR conference*, Seattle, Washington, USA: ACM Press, p. 533. doi: 10.1145/1148170.1148262.
- Lane, H., Howard, C. and Hapke, H. M. (2019) *Natural language processing in action: understanding, analyzing, and generating text with Python*. Shelter Island, NY: Manning Publications Co.
- Mahmud, S. R. (2013). 'A Simple Information Retrieval Technique'. *Int. J. on Recent Trends in Engineering and Technology*, 8(1), pp. 68-71
- Robertson, S. (2004) 'Understanding inverse document frequency: on theoretical arguments for IDF', *Journal of Documentation*, 60(5), pp. 503–520. doi: 10.1108/00220410410560582.
- Robertson, S. and Zaragoza, H. (2009) 'The Probabilistic Relevance Framework: BM25 and Beyond', *Foundations and Trends® in Information Retrieval*, 3(4), pp. 333–389. doi: 10.1561/1500000019.
- Yu, C.T. and Salton, G., 1976. Precision weighting—an effective automatic indexing method. Journal of the ACM (JACM), 23(1), pp.76-88.