



# SpecifiedBy

Joan Figuerola Hurtado

Co-founder & CTO

@joanfihu

joan@specifiedby.com

# Outline



- Crawling, structuring and storing data.
- Search engines: architectures, resources and evaluation.
- Recommendation engines: architectures, resources and evaluation.
- Q&A



# Getting Data

- Many ways to get data. Problem related.
- Scan books, manual collection, public Internet data, etc

## Web Crawlers

- Scrapy: open source, Python, Non-Blocking and efficient
- Identify Target Domain and Pages
- False Dead Ends & Infinite Loops
- CSS Selectors
- Post Processing & Storage

# CSS Selectors



0844 826 7766

enquiries@olsenuk.com

Visit



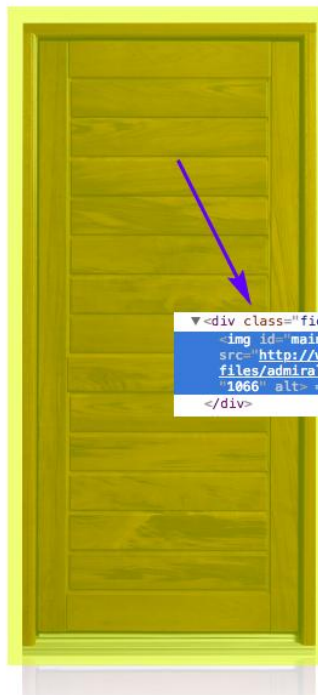
**h1#product\_title**

```
<div class="field-item even">
  <h1 id="product_title">
    <a href="/products/entrance-doors/je-
    trae">JeTrae</a>
    <span>&nbsp;&nbsp;&nbsp;Admiral</span>
  </h1>
</div>
```

## JeTrae | Admiral

The Admiral is a timeless design with simple, clean lines which create calm and harmonious entrance facades in both new and old developments. The door is supplied fully factory finished with a range of stains or paints. If stained the door is made from either Mahogany or Teak timber. When painted the door is finished with your choice of colour from the RAL Classic range (approx. 250 colours).

Timber available; Light Oak, dark Oak, Mahogany & Teak.



```
<div class="field-item even">
  
</div>
```

**img#main\_image**

```
<div class="field-items">
  <div class="field-item even">
    <p id="product_description">
      <p>Timber available; Light Oak, dark Oak,
      Mahogany & Teak</p>
      <p>&nbsp;&nbsp;&nbsp;</p>
    </p>
  </div>
</div>
```

**p#product\_description**



# Post Crawl Processing & Storage

## **Post Crawl Processing:**

- Remove HTML tags
- Resize images
- Classify products
- Convert PDF to text
- Extract properties
- Deduplication
- Etc.

## **Storage:**

Many built-in storage pipelines:

- MySQL
- CSV
- JSON
- S3
- XML

# Search Engines



Google Search

I'm Feeling Lucky

amazon.co.uk prime

TVs

prime Explore One-Day delivery

Deliver to Joan EDINBURGH EH39DR

Shop by Department

Joan's Amazon Today's Deals Gift Cards & Top Up Sell Help

Hello, Joan Your Account Your Prime Your Lists Basket

Electronics Best Sellers Deals Phones & Accessories TVs & Home Cinema Camera & Photo Audio & HiFi Computers & Accessories Wearable Technology Accessories

## TV Store

Explore the wide range of [Ultra HD 4K TVs](#), [Smart TVs](#), [Full HD TVs](#), [LCD TVs](#), [LED TVs](#), [OLED TVs](#), [3D TVs](#), and [A/V Accessories](#). Learn more about TV technology and more with our [TV Buying Guide](#).

Show results for

< Electronics & Photo  
< Home Cinema, TV & Video  
TVs

Refine by

### Screen Size

- ☐ Up to 29"
- ☐ 30 - 39"
- ☐ 40 - 49"
- ☐ 50 - 59"
- ☐ 60 - 69"
- ☐ 70" & up

### HD Format

- ☐ 1080p Full HD
- ☐ 4K Ultra HD
- ☐ 720p HD Ready

### Display Technology

- ☐ CRT
- ☐ LCD
- ☐ LED
- ☐ OLED
- ☐ Plasma
- ☐ QLED

### Featured Brands

- ☐ LG Electronics
- ☐ Toshiba
- ☐ Samsung
- ☐ Sony

## Which TV is right for you?



- Screen Size: All
- Screen Resolution: All
- Features: All
- Brands: All
- Freeview: All
- Delivery Option: All
- Price: All

3,643 results

Find Now

## Most gifted

LG 43UJ630V 43 inch 4K Ultra HD HDR...

LG Electronics

★★★★☆ 536

£429.99 £339.00

prime

August DA100D - 10" Portable TV with...

★★★★★ 338

£149.50 £139.95

prime

Samsung UE55MU6220 55" 4K Ultra HD...

★★★★★ 4

£595.00

prime

Sony Bravia KDL40RE453 Full HD HDR TV...

★★★★☆ 153

£460.00 £328.99

prime

LG 43UJ670V 43 inch 4K Ultra HD HDR...

LG Electronics

★★★★☆ 117

£499.99 £386.00

prime

## Best sellers

# ElasticSearch & SQL



Elasticsearch is a search engine based on Lucene. It provides a distributed, multitenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents.

- Document Indexing
- TFIDF Scoring by default
- Field Boosting
- Reranking
- Autocomplete
- Misspelling errors
- Related searches
- Stemming
- Geolocation searches
- Etc

ES search is good for full text searches but difficult to work with with complex queries -> ES for full text -> SQL for the rest.



# Quality Score

Data rating of the listing. Not a user review.

## Title:

- Malvern - Tilt & Turn (Bad)
- Malvern - Tilt & Turn Stainless Steel Security Window W: 1500mm (Good)

## Images:

- High Resolution (Good)
- > 3 images (Good)

## Files:

- Has Files (Good)
- > 1 files (Good)
- Has 3D model (Good)

$$QS(P) = St*Wt + Si*Wf + Sf*Wf + \dots + Sx*Wx$$

## Improvements for Interpon D1036 Textura QS: 73<sup>x</sup>

Correcting the following issues could increase **Interpon D1036 Textura** popularity by **66.156%**.

- Add an adjective to your product name.
- The product description is too short. We recommended you provide a description of at least 140 words.
- Add more images.
- You should add more files for download. Products with 10 or more files available perform best.





# Popularity Score & Reranking

Measures how much users have engaged with a product.

- User A **clicks** on Product AA
- User A **downloads** a file from Product AA
- User B **clicks** on Product AA
- User B **compares** Product AA with Product BB

$$Ps(A) = \#clicks * Wc + \#downloads * Wd + \#comparisons * Wc$$

## Reranking:

$$S(query, document) = tfidf * Wtfidf + quality\_score * Wqs + popularity\_score * Wps$$

# Metrics

Search Click-Through-Rate(CTR):  $(\text{clicks} / \text{searches}) * 100$

Product Click-Through-Rate(CTR):  $(\text{clicks} / \text{impressions}) * 100$

Average Click Position:  $\text{SUM}(\text{click\_positions}) / \text{clicks}$

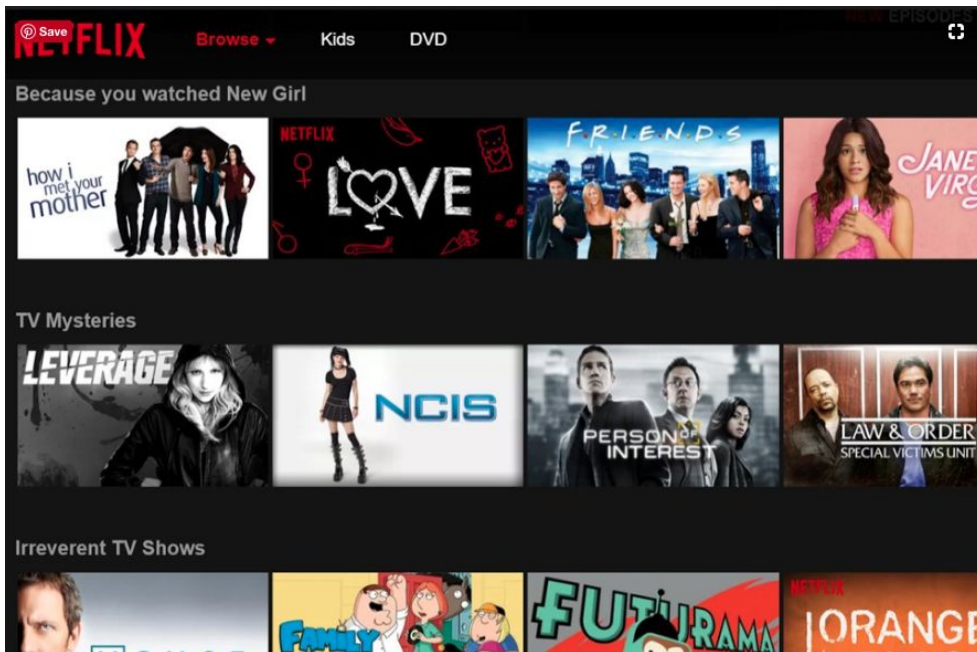
## Viewability

The screenshot shows the SpecifiedBy website interface. The top navigation bar includes links for 'Bathroom & Sanitary Fittings', 'Sanitary Waste Disposal', 'Sanitary Waste Bins', 'Sanitary Waste Macerator', and 'Chemical Sanitary Waste Disposal'. The main content area is titled 'Sanitary Waste Disposal' and features a list of products. The first product is 'WP133 - Sanitary Waste Bins' by Dolphin Solutions Ltd, described as a 'Prestige Waste Bin' with a '18/10 chrome nickel stainless steel bin - Satin finish, welded corners - no sharp edges - Lockable - Waste towels disposed through a self-closing, fire preventativ...'. The second product is '800.05.10090 HEWI Knee/Hand Operated Waste Bin' by HEWI UK, described as a 'HEWI Waste bin - cylindrical container with hinged lid - removable waste container for easy cleaning - lid opens by applying slight pressure to container with the knee or hand - with self-adhesive label with instructions for knee operation - integrated ...'. The third product is 'Nappy Disposal Unit' by Carson Hygiene Ltd, described as 'Our offensive waste units can be used for the disposal of nappies and incontinence waste. The waste is collected and disposed of in line with latest legislation and our unit exchange service provides a fresh, clean and hygienic unit with every service ...'. The left sidebar contains filters for 'Height (mm) (42)', 'Width (mm) (40)', 'Depth (mm) (40)', 'Finish(es)', 'Material(s)', 'Storage Volume (l) (29)', and 'Certifications'.

The screenshot shows a product catalog page with a list of sanitary waste disposal units. The units are listed in a grid format, each with a small image and a brief description. The units are arranged in a vertical list, with each unit having a small image and a brief description. The units are arranged in a vertical list, with each unit having a small image and a brief description. The units are arranged in a vertical list, with each unit having a small image and a brief description.



# Recommendations Engines



amazon.com

Recommended for You

Amazon.com has new recommendations for you based on [items](#) you purchased or told us you own.



[Google Apps Deciphered: Compute in the Cloud to Streamline Your Desktop](#)



[Google Apps Administrator Guide: A Private-Label Web Workspace](#)



[Googlepedia: The Ultimate Google Resource \(3rd Edition\)](#)



# Content Based Filtering

Features: text, category, images, sound, etc

Vectorize/Embeddings: Word2Vec, CNN, Binarize Labels, etc

Calculate Similarity: cosine, euclidean, etc

	Item 1	Item 2	Item 3	Item 4
Item 1	1	0.8	0.1	0.3
Item 2	0.8	1	0.6	0.5
Item 3	0.1	0.6	1	0.9
Item 4	0.3	0.5	0.9	1

# Collaborative Filtering



Sparks serendipity (Wow! factor)

Features: ratings from users to items. Explicit (i.e star reviews) and implicit (i.e clicks) ratings

	Item 1	Item 2	Item 3	Item 4
User 1	5	3	-	1
User 2	4	-	-	1
User 3	1	1	-	5
User 4	1	-	-	4
User 5	-	1	5	4

**Matrix  
Factorisation**

	Item 1	Item 2	Item 3	Item 4
User 1	4.97	2.98	<b>2.18</b>	0.98
User 2	3.97	<b>2.40</b>	<b>4.59</b>	0.99
User 3	1.02	0.93	<b>5.32</b>	4.93
User 4	1	<b>0.85</b>	<b>4.59</b>	3.93
User 5	1.36	1.07	4.89	4.12



# Resources


- <https://www.coursera.org/specializations/recommender-systems>
- <http://www.quuxlabs.com/blog/2010/09/matrix-factorization-a-simple-tutorial-and-implementation-in-python/>
- <http://lenskit.org/>
- <https://grouplens.org/>
- <http://www.mmids.org/>
- <https://spark.apache.org/docs/2.2.0/mllib-collaborative-filtering.html>
- <http://scikit-learn.org/>
- <https://radimrehurek.com/gensim/>
- <https://github.com/tensorflow/models>



# Attention Span For Personalisation

Time spend engaging with an item. More expressive than clicks/taps.

<https://arxiv.org/abs/1608.00147>

 Cornell University  
Library

arXiv.org > cs > arXiv:1608.00147

Search or Article  
(Help | Advanced search)

Computer Science > Information Retrieval

## Attention Span For Personalisation

[Joan Figuerola Hurtado](#)

*(Submitted on 30 Jul 2016)*

A click on an item is arguably the most widely used feature in recommender systems. However, a click is one out of 174 events a browser can trigger. This paper presents a framework to effectively collect and store data from event streams. A set of mining methods is provided to extract user engagement features such as: attention span, scrolling depth and visible impressions. In this work, we present an experiment where recommendations based on attention span drove 340% higher click-through-rate than clicks.

Subjects: **Information Retrieval** (cs.IR)

Cite as: **arXiv:1608.00147 [cs.IR]**  
(or **arXiv:1608.00147v1 [cs.IR]** for this version)

### Submission history

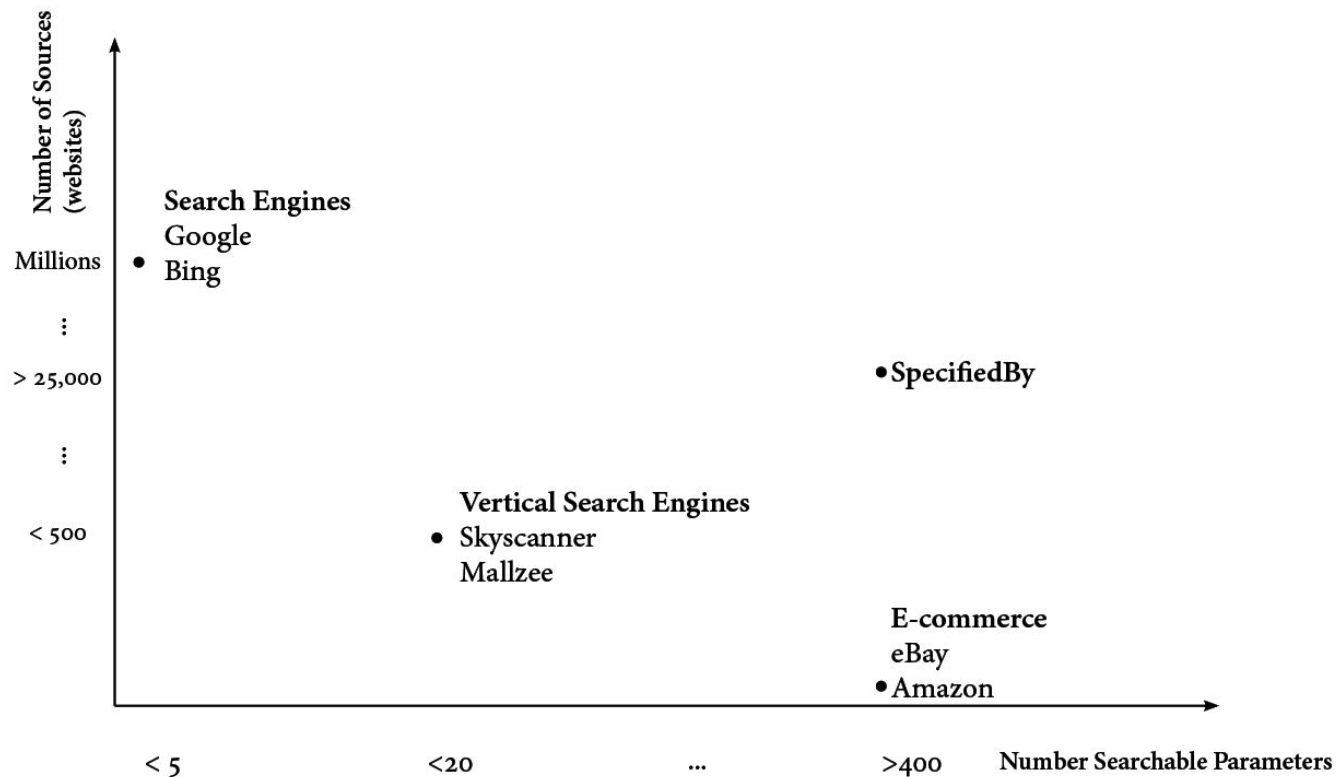
From: Joan Figuerola Hurtado [[view email](#)]  
[v1] Sat, 30 Jul 2016 16:53:05 GMT (470kb)



Thank you.



# Challenge



# Solution

