



# Building a Multi- faceted Recommender System

THINKFUL DATASCIENCE : FINAL CAPSTONE

ROY SHABAYA

# Introduction



## Research Topic :

Can we build a recommendation engine to provide high quality and relevant recommendations to existing and new customers



## Purpose :

In the competitive online retail market being able to provide high quality recommendations can be an advantage that increases customer satisfaction and drives incremental sales



## Data:

One year of customer, transaction and product data for a UK Based online retailer selling gift items across Europe. Obtained from UCI's Machine Learning Repository.

# Recommender System Basics

- ▶ Content – based on information about an item (description, attributes, etc)
  - ▶ Recommend items on the similarity of their attributes
  - ▶ Doesn't require user input, and allows recommendations for new products
- ▶ Collaboration – based on explicit or implicit feedback from customer base
  - ▶ Explicit – Can ask customers for ratings, rankings, reviews, etc.
  - ▶ Implicit – Gather data about activity: purchases, views, cart adds, etc.
  - ▶ Central premise : Past similarity in taste can be used to predict future preferences
- ▶ Knowledge Based – Ask user questions, make recommendations based on answers
- ▶ Hybrids – Combine different methods

# Exploration Initial Insights

- ▶ 541,909 rows of data
- ▶ 1 year of Data – December 2010 – December 2011
- ▶ 4,338 customers
- ▶ 4,026 unique products in dataset
- ▶ Wide gaps in products and customers
- ▶ Company appears to serve both industry and direct

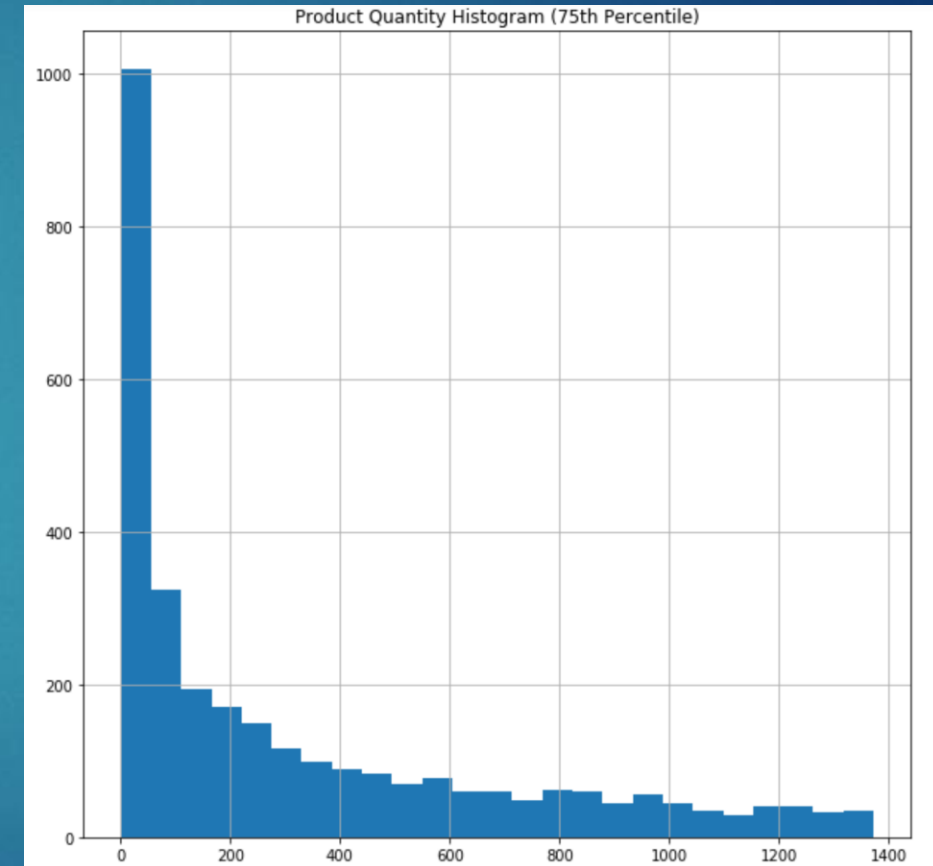


# Wide Gaps in Product Turn Frequency

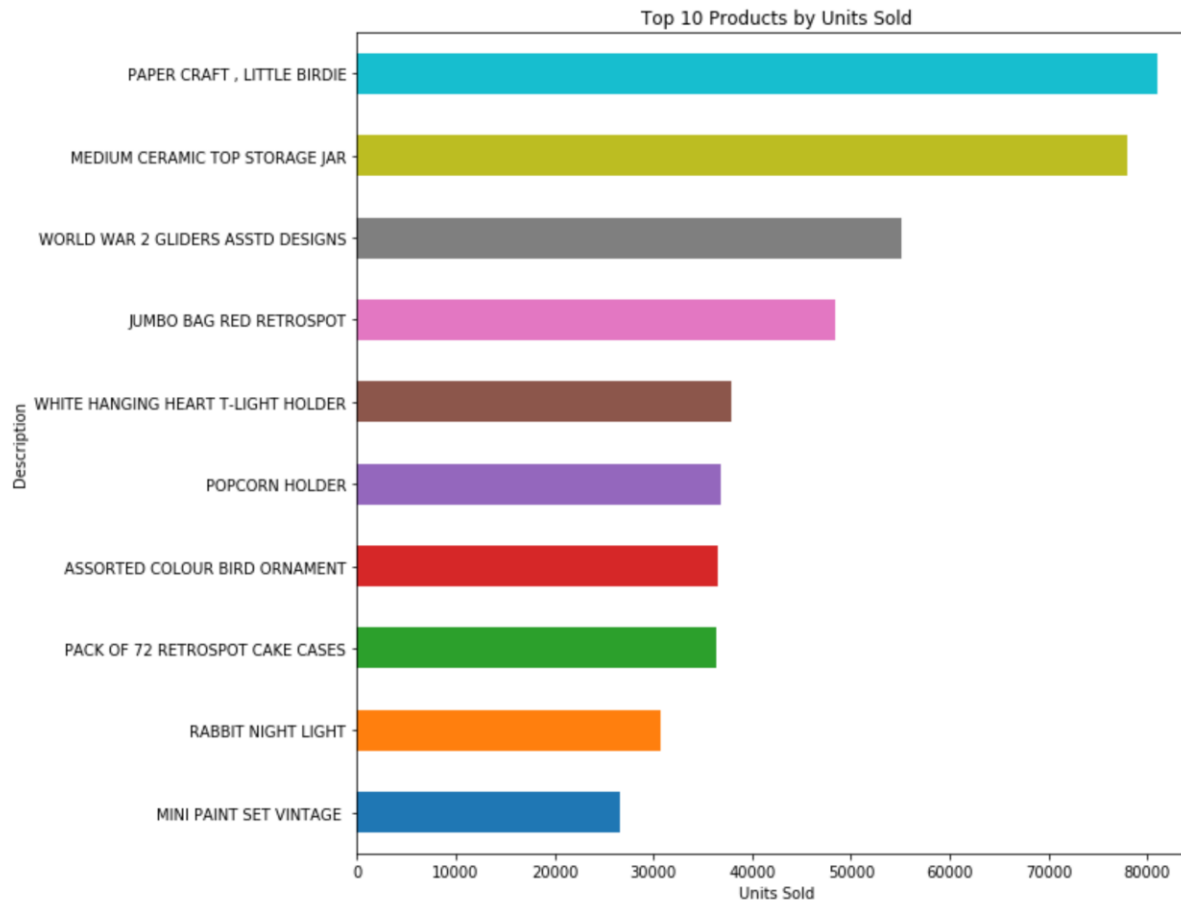
In [36]: `product_grouped.describe()`

Out[36]:

	Quantity	Revenue	ASP
<b>count</b>	4026.000000	4026.000000	4026.000000
<b>mean</b>	1388.071535	2649.449713	8.122115
<b>std</b>	3465.395806	7815.336619	205.624154
<b>min</b>	1.000000	0.003000	0.001000
<b>25%</b>	56.000000	126.757500	1.025689
<b>50%</b>	358.000000	666.900000	1.969831
<b>75%</b>	1372.750000	2190.140000	3.975347
<b>max</b>	80995.000000	206248.770000	11062.060000

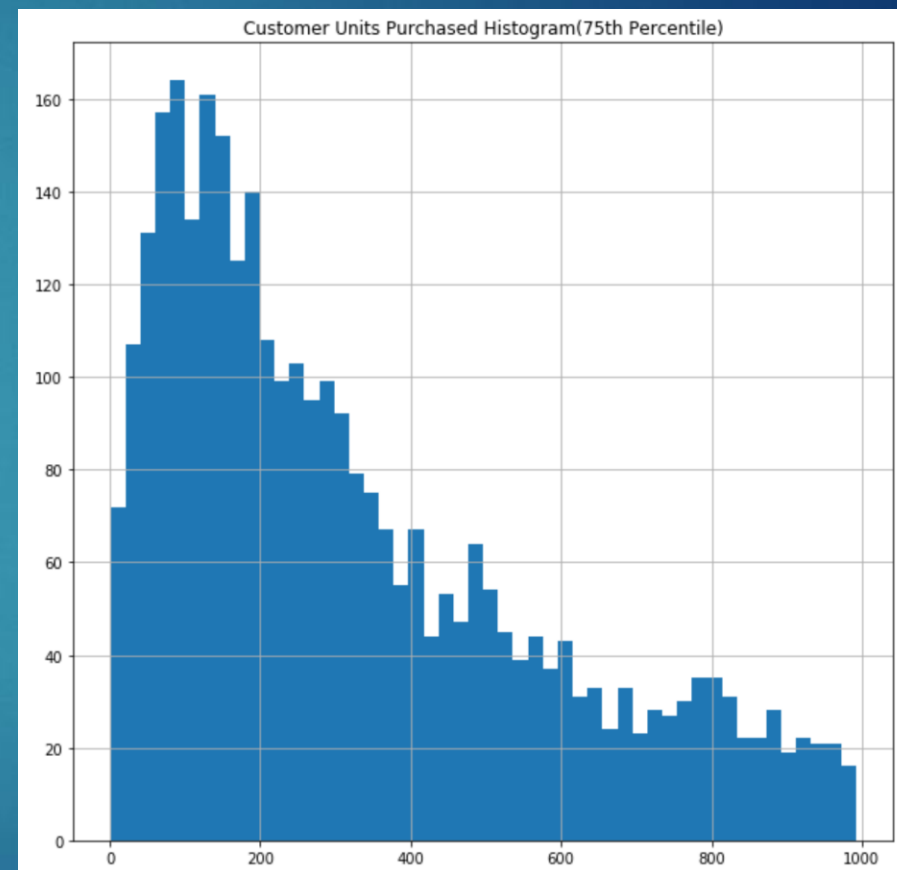


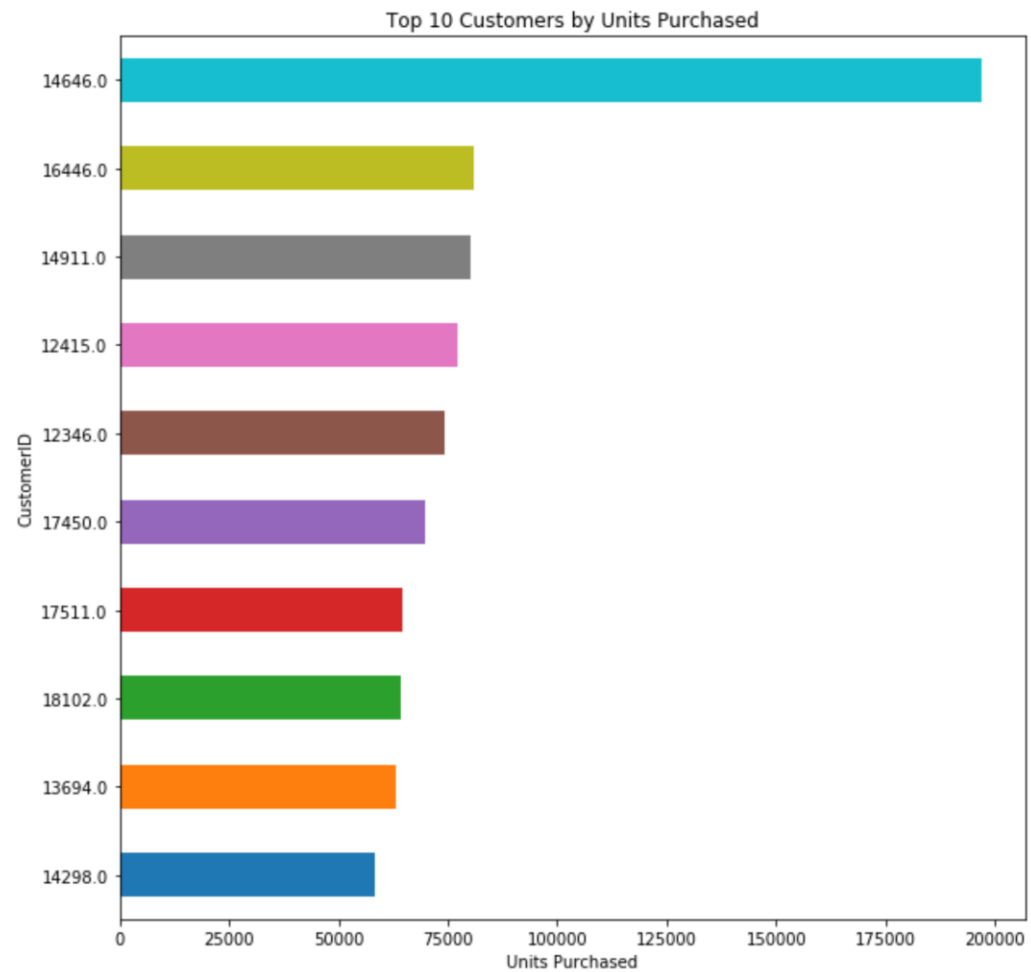
# Top Products



# Wide Gaps in Customers

	Quantity	Revenue	ASP
<b>count</b>	4338.000000	4338.000000	4338.000000
<b>mean</b>	1191.289073	2054.266460	3.013989
<b>std</b>	5046.081546	8989.230441	33.218307
<b>min</b>	1.000000	3.750000	0.085619
<b>25%</b>	160.000000	307.415000	1.408745
<b>50%</b>	379.000000	674.485000	1.808724
<b>75%</b>	992.750000	1661.740000	2.351760
<b>max</b>	196915.000000	280206.020000	2033.100000





# Top Customers



# Modeling : 2 Aspects



Similar Items –  
Recommend similar to  
what customer currently  
viewing



Market Basket –  
Recommendation of  
what other customers  
have purchased with a  
given item

# Natural Language Processing

- ▶ First key challenge is converting text to form algorithms can understand
- ▶ Cleaning is critical to ensure text is consistent
- ▶ Cleaned text is then converted to numerical representations - vectors

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Revenue
0	536365	85123A WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	15.30
1	536365	71053 WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
2	536365	84406B CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	22.00
3	536365	84029G KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
4	536365	84029E RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34

# Doc2vec

- ▶ Extension of Word2vec – a shallow neural network that vectorizes words
- ▶ Word2vec vectorizes single words, Doc2vec allows you to vectorize text that contains more than one word
- ▶ Works best on very large documents
- ▶ Results – Initial visual inspections suggests it is not recommending similar items
- ▶ Words that surround other words not necessarily helpful for product descriptions

Original (4025): «paper craft little birdie»

SIMILAR DOCS PER MODEL Doc2Vec(dm/m,d500,n5,w5,mc20,s0.001,t3):

First Match (1476, 0.8212401866912842): «pencils tall tube posy»

Second Match (1894, 0.8202415704727173): «pencils tube posy»

Third Match (294, 0.8173691034317017): «pencils tall tube skulls»

Fourth Match (319, 0.810957670211792): «ribbon reel making snowmen»

Fifth Match (3774, 0.8089277744293213): «embroidered ribbon reel sally»

Sixth Match (1138, 0.8070483803749084): «pencils tube skulls»

Seventh Match (3765, 0.8058570027351379): «embroidered ribbon reel ruby»

Eighth Match (3792, 0.8039880990982056): «embroidered ribbon reel claire»

Ninth Match (3775, 0.8029369115829468): «embroidered ribbon reel rachel»

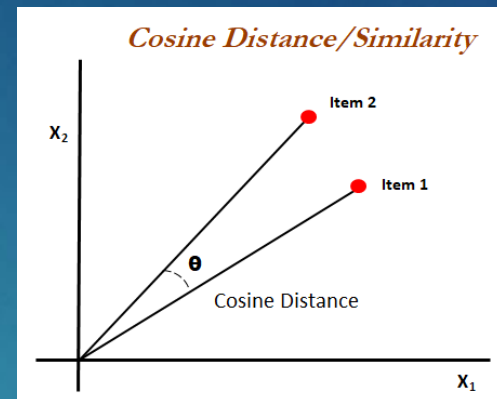
Tenth Match (3766, 0.8018028736114502): «embroidered ribbon reel emily»

# Term Frequency – Inverse Document Frequency (TFIDF)

- ▶ Numerical representation of text that reflects how important a word is in a document
- ▶ Term Frequency – How many times a word is used in a product description
- ▶ Inverse Document Frequency – How many product descriptions a word is used in
- ▶ Produces vectors which can be then used for analysis
- ▶ Singular-value Decomposition – Dimension reduction technique that reduces noise

$$\mathbf{tfidf}_{i,j} = \mathbf{tf}_{i,j} \times \log \left( \frac{N}{\mathbf{df}_i} \right)$$

# Cosine Similarity



- ▶ Similarity measure that takes cosine of the angle between two vectors
- ▶ Cosine of  $0^\circ$  is 1 – so more similar vectors will have a higher value
- ▶ Simple to implement with SKLearn package – also runs quickly
- ▶ Results – Initial visual inspections suggests it is recommending similar items

```
cosine_similarity('WHITE METAL LANTERN')
```

```
Original      : WHITE METAL LANTERN
First Match   : WHITE MOROCCAN METAL LANTERN
Second Match  : HANGING METAL HEART LANTERN
Third Match   : HANGING METAL STAR LANTERN
Fourth Match  : WHITE LOVEBIRD LANTERN
Fifth Match   : LANTERN CREAM GAZEBO
Sixth Match   : WHITE WITH METAL BAG CHARM
Sevent Match  : SMALL HANGING GLASS+ZINC LANTERN
Eight Match   : FRENCH CARRIAGE LANTERN
Ninth Match   : PAPER LANTERN 5 POINT STUDDER STAR
Tenth Match   : PAPER LANTERN 5 POINT SEQUIN STAR
```

# Similarity Models Review

- ▶ TFIDF with Cosine Similarity – Performing well, would continue to develop
- ▶ Doc2vec – Not performing well, potentially too few products for Word2vec, would not continue



# Market Basket

- ▶ Products “Frequently purchased together”
- ▶ Association Rules – Find products frequently on the same transaction
  - ▶ Support – Percent of orders that contain an itemset
  - ▶ Confidence – Percent of occurrences where item 1 is purchased, given that item 2 was also purchased
  - ▶ Lift – Measures whether there is a relationship, or co-occurrences are just due to random chance
- ▶ Utilized mlextends Apriori Algorithm

# Market Basket Results

	antecedents		consequents	antecedent support	consequent support	support	confidence	lift
13	(6 RIBBONS RUSTIC CHARM)	(SCANDINAVIAN REDS RIBBONS)		0.048029	0.022682	0.010913	0.227225	10.018059
1	(6 RIBBONS RUSTIC CHARM)	(JAM MAKING SET PRINTED)		0.048029	0.058439	0.011869	0.247120	4.228694
9	(6 RIBBONS RUSTIC CHARM)	(RECIPE BOX PANTRY YELLOW DESIGN)		0.048029	0.055874	0.010561	0.219895	3.935552
3	(6 RIBBONS RUSTIC CHARM)	(JAM MAKING SET WITH JARS)		0.048029	0.056930	0.010410	0.216754	3.807363
7	(6 RIBBONS RUSTIC CHARM)	(PACK OF 72 RETROSPOT CAKE CASES)		0.048029	0.066385	0.011969	0.249215	3.754079

Appears to be making reasonable recommendations...

	antecedents		consequents	antecedent support	consequent support	support	confidence	lift
1465	(REGENCY TEAPOT ROSES )	(REGENCY SUGAR BOWL GREEN)		0.020921	0.016194	0.011718	0.560096	34.586807
1463	(REGENCY TEAPOT ROSES )	(REGENCY MILK JUG PINK )		0.020921	0.016546	0.011366	0.543269	32.833937
1475	(REGENCY TEAPOT ROSES )	(REGENCY TEA PLATE ROSES )		0.020921	0.022380	0.010008	0.478365	21.374870
1478	(REGENCY TEAPOT ROSES )	(ROSES REGENCY TEACUP AND SAUCER )		0.020921	0.053561	0.011919	0.569712	10.636755
1443	(REGENCY TEAPOT ROSES )	(REGENCY CAKESTAND 3 TIER)		0.020921	0.099980	0.011366	0.543269	5.433785



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
<b>5234</b>	(REGENCY TEA PLATE GREEN , PINK REGENCY TEACUP ...	(REGENCY TEA PLATE PINK, GREEN REGENCY TEACUP ...	0.011549	0.012073	0.010499	0.909091	75.296443
<b>5235</b>	(REGENCY TEA PLATE PINK, GREEN REGENCY TEACUP ...	(REGENCY TEA PLATE GREEN , PINK REGENCY TEACUP ...	0.012073	0.011549	0.010499	0.869565	75.296443
<b>7379</b>	(SET/6 RED SPOTTY PAPER CUPS, PACK OF 6 SKULL ...	(PACK OF 20 SKULL PAPER NAPKINS, PACK OF 6 SKU...	0.012073	0.013648	0.011024	0.913043	66.897993
<b>7370</b>	(PACK OF 20 SKULL PAPER NAPKINS, PACK OF 6 SKU...	(SET/6 RED SPOTTY PAPER CUPS, PACK OF 6 SKULL ...	0.013648	0.012073	0.011024	0.807692	66.897993
<b>7351</b>	(PACK OF 20 SKULL PAPER NAPKINS, PACK OF 6 SKU...	(SET/6 RED SPOTTY PAPER CUPS, PACK OF 6 SKULL ...	0.011549	0.014698	0.011024	0.954545	64.943182
<b>7398</b>	(SET/6 RED SPOTTY PAPER CUPS, PACK OF 6 SKULL ...	(PACK OF 20 SKULL PAPER NAPKINS, PACK OF 6 SKU...	0.014698	0.011549	0.011024	0.750000	64.943182

However strongest lift (comparison to random chance – 1.0 = random) comes from set purchases – presumably would occur from resellers who buy many different variations of a product

Nature of dataset (unique products) and customers (direct vs industry) appears to be causing these results. Recommendations initially seem reasonable – can proceed with testing.

# Similarity Demo

- ▶ May have found company data is from
  - ▶ UK based retailer that ships throughout Europe
  - ▶ Many products and stock numbers match
- ▶ Currently recommendations appear to be based on “tags” – provides weak recommendations in many cases

# Live Now:



## Alarm Clock Red

£12.95


IN STOCK


★★★★★ Read 15 reviews

Quantity: 1

Add to Basket ►

 Sign up for free UK delivery over £20\*

 Speedy tracked delivery

 30 day returns (conditions apply)



### Description

### Returns

### Delivery

This red retro alarm clock is a great present for both men and women! All it requires to get started is 1xAA battery (not included)

**Material:** Metal, Plastic

**Dimensions:** Length: 11 cm Height: 9 cm Width: 6 cm

**Product Code:** 22727

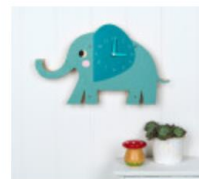
**Tags:** [Alarm Clocks and Wall Clocks](#) [Vintage and Retro Gifts](#) [Stocking Fillers For Him](#)

### Related Items

### Recently Viewed Items



Retro Blue Alarm Clock  
£12.95



Elvis The Elephant  
Wooden Wall Clock  
£24.95



Charlie The Lion Wooden  
Wall Clock  
£24.95



Milo The Penguin  
Wooden Wall Clock  
£24.95



Cookie The Cat Wooden  
Wall Clock  
£24.95



Hello Sunshine Wooden  
Wall Clock  
£24.95

# My Recommendations:

```
In [752]: similarity_tool('ALARM CLOCK BAKELIKE RED ')
```

```
Original      : ALARM CLOCK BAKELIKE RED  
First Match   : ALARM CLOCK BAKELIKE PINK  
Second Match  : ALARM CLOCK BAKELIKE GREEN  
Third Match   : ALARM CLOCK BAKELIKE IVORY  
Fourth Match  : ALARM CLOCK BAKELIKE CHOCOLATE  
Fifth Match   : ALARM CLOCK BAKELIKE ORANGE
```



Alarm Clock Pink

★★★★★ (12)

Price: £12.95    [buy now >>](#)



Retro Green Alarm Clock

★★★★★ (29)

Price: £12.95



Alarm Clock Ivory

★★★★★ (8)

Price: £12.95    [buy now >>](#)



Alarm Clock Chocolate

★★★★★ (3)

Price: £12.95    [buy now >>](#)



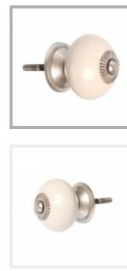
Alarm Clock Orange

★★★★★ (5)

Price: £12.95    [buy now >>](#)



Also currently  
live:



## Drawer Knob Ceramic Ivory

£3.95

IN STOCK



Quantity: 1

Add to Basket ►



Sign up for free UK delivery over £20\*



Speedy tracked delivery



30 day returns (conditions apply)



### Description

Returns

Delivery

Ivory round ceramic drawer knob with metal fixtures. Revitalise and old chest of drawers or your kitchen cupboards. (Screw length 4cm approx.)

**Material:** Metal, Ceramic

**Dimensions:** Diameter: 4.0cm

**Product Code:** 23035

**Tags:** [Furniture and Accessories](#) [Furniture and Accessories](#)

### Related Items

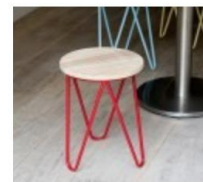
Recently Viewed Items



Ivory Fifties Style Wooden Stool  
£29.95



Blue Fifties Style Wooden Stool  
£29.95



Red Fifties Style Wooden Stool  
£29.95



Today's Special Blackboard Easel  
£69.95



French Iron Garden Table  
£125.00



Vintage First Aid Cabinet  
£69.95

# My Recommendations:

```
In [821]: similarity_tool('DRAWER KNOB CERAMIC IVORY')
```

```
Original      : DRAWER KNOB CERAMIC IVORY  
First Match   : DRAWER KNOB CERAMIC RED  
Second Match  : DRAWER KNOB CERAMIC BLACK  
Third Match   : BLUE SPOT CERAMIC DRAWER KNOB  
Fourth Match  : RED SPOT CERAMIC DRAWER KNOB  
Fifth Match   : BLUE STRIPE CERAMIC DRAWER KNOB
```



Drawer Knob Ceramic Red



Price: £3.95

[buy now >>](#)



Drawer Knob Ceramic Black



Price: £3.95

[buy now >>](#)



Blue Spot Ceramic Drawer Knob



Price: £3.95

[buy now >>](#)



Red Spot Ceramic Drawer Knob



Price: £3.95

[buy now >>](#)



Blue Stripe Ceramic Drawer Knob



Price: £3.95

[buy now >>](#)

# Let's look at a few more...



```
In [764]: similarity_tool('HAND WARMER OWL DESIGN')
```

```
Original      : HAND WARMER OWL DESIGN  
First Match   : HI TEC ALPINE HAND WARMER  
Second Match  : HAND WARMER BIRD DESIGN  
Third Match   : HAND WARMER SCOTTY DOG DESIGN  
Fourth Match  : HAND WARMER BABUSHKA DESIGN  
Fifth Match   : HAND WARMER RED RETROSPOT
```

```
In [765]: similarity_tool('DOORMAT FAIRY CAKE')
```

```
Original      : DOORMAT FAIRY CAKE  
First Match   : DOORMAT TOPIARY  
Second Match  : DOORMAT FRIENDSHIP  
Third Match   : DOORMAT AIRMAIL  
Fourth Match  : HANGING FAIRY CAKE DECORATION  
Fifth Match   : FAIRY CAKE DESIGN UMBRELLA
```

```
In [766]: similarity_tool('RABBIT NIGHT LIGHT')
```

```
Original      : RABBIT NIGHT LIGHT  
First Match   : SUNJAR LED NIGHT NIGHT LIGHT  
Second Match  : RED TOADSTOOL LED NIGHT LIGHT  
Third Match   : FAIRY TALE COTTAGE NIGHT LIGHT  
Fourth Match  : SET 10 LIGHTS NIGHT OWL  
Fifth Match   : SET 10 NIGHT OWL LIGHTS
```





# Conclusions

- ▶ TFIDF Cosine Similarity recommender performing well – would proceed with live testing
  - ▶ Modeling and recommendation production for 4,000 products takes a few minutes on my laptop
  - ▶ In enterprise environment could easily scale
- ▶ Doc2vec not performing well, would not continue to pursue
- ▶ Market Basket recommender shows some initial promise, can be tested
- ▶ Future Research to Explore:
  - ▶ Additional customer and product information could be beneficial – customer type (direct vs industry), product categories and attributes
  - ▶ Develop trending algorithm to complement other recommendations
  - ▶ Determine whether separate models are needed for direct vs industry



# Natural Language Processing

- ▶ First key challenge is converting text to form algorithms can understand
- ▶ Cleaning is critical to ensure text is consistent
- ▶ Cleaned text is then converted to numerical representations - vectors

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Revenue
0	536365	85123A WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	15.30
1	536365	71053 WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
2	536365	84406B CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	22.00
3	536365	84029G KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
4	536365	84029E RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34



THANK YOU – Questions?