# Recommending stocks

An La - 12th May, 2020

# Demo Script

Script:
https://colab.research.google.com/drive/1jWa0d05E4phx0QqT8JOBsI_p-JrS2xh9

Dataset: file ML-technicaltest-ecommerce.csv

https://drive.google.com/drive/folders/12aNzUg-zKHuJqa45Z3RkpzwD2uUqPq4s?usp=drive_link

# Outline

1. Problem Statement, assumptions and the process of modelling
2. Exploring data (link)
3. Methods Review & Planning (link)
4. Pre-processing and Feature engineering (link)
5. Recommending models (link)
    5.1. Content-based
    5.2. Collaborative Filtering
    5.3. The Hybrid
6. System design (link)
7. Summary (link)

# 1. Problem Statement, assumptions and the process of modelling

## 1.1 Problem Statement

Recommend list of 5 related stocks from the current stock.

Questions:

- Target from business?
  - Attracting more users -> target metric: number of users
  - Improve engagement of current users -> target metric: average number of purchases/week, months...
- Definitions of “related”?
  - Depending on the target

# 1. Problem Statement, assumptions and the process of modelling

## 1.2 Assumptions

Suppose the targets:

- Supporting users to explore products
- No business metrics.

Definitions of "related":

- Similar in *usage* (kitchen utensils, garden tools), *properties* (technical devices, decorating gadgets), *context* (Christmas, Summer)...
- "People buy x also buy y": expensive wall clock -> luxurious jewels (they're rich), a guitar -> paintings (they like arts), tree pots -> books (they're retired and enjoy life at home).

# 1. Problem Statement, assumptions and the process of modelling

## 1.3 The process of modelling

### Normal process of modelling

- Explore data: getting overview (1) and considering the target
- Analyze relationship between the target and data fields (2)
- Getting overview of current methods and Planning (3)
- Feature engineering (4)
- Modelling (5)
- Evaluation (6): offline metrics of models (such as precision) and business metrics.

With assumptions:

- No business target metrics -> No analysis (skip 2), the evaluation (6) is considered by just offline metrics

# Structure of The Slide

Following the process of modeling for this task.

- Section 2 and 3: explanation about data and how and why to select methods. These parts are preparation for section 4 and 5.
- Section 4 and 5: details of methods and challenges caused by particular cases of the dataset. Evaluation of methods is included.
- Section 6: a little bit about designing the appropriate pipeline.

# 2. Exploring data

- Separating train/test
- Getting overview
- Classifying stocks, users and invoices
- Relationship between users and stocks
- Relationship between users/stocks and unitprice/quantity
- Conclusion

This section mainly focuses on behaviour of users. Exploring more on stocks and their description is presented in section 4.

# Separating train/test

To make sure to keep the testing data private, the first step is splitting dataset. In the real context, we do not know will data of tomorrow is different from today, so that completely keep testing secret is necessary.

Splitting is based on InvoiceNo: training data has 90% invoices and testing has 10% invoice. Hope that removing 10% invoice does not bring too much difference from full data to training.

However, after splitting, some InvoiceNo are lost.

From now, just explore training data.

```
Number of purchases in full data:  541909
Number of purchases in training data:  315838
Number of purchases in testing data:  53911
Number of Invoices in full data:  25900
Number of Invoices in training data:  15411
Number of Invoices in testing data:  2474
```

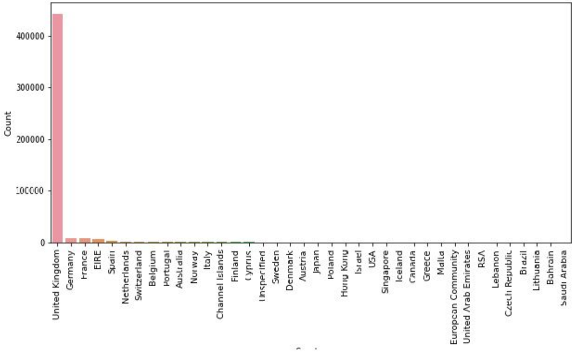
# Overview

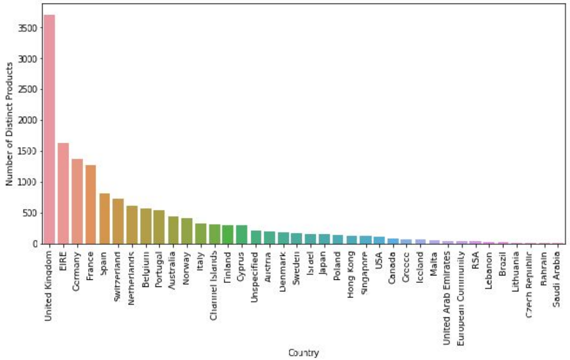
- 315838 lines in training data
- Each line is called a purchase, identified by Key: *InvoiceNo* and *StockCode*
- All columns have low missing rate, except CustomerID is 24%
- There are 38 countries, UK is dominant in lines and number of stocks.
- In most of countries, purchases have stable prices and quantity.

```
The number of unique invoice:  15411
Special cases of invoice:
 + None or multiple countries:  0
 + Negative Quantity:  3129
 + Negative Unit Price:  1

The number of unique items by StockCode:  3923
The number of unique items by Description:  4036

The number of unique users/customeres:  3669
```

Summary basic info of Data

| | InvoiceNo | StockCode | Description | Quantity | UnitPrice | CustomerID | Country |
|---|---|---|---|---|---|---|---|
| 0 | 563709 | 22152 | PLACE SETTING WHITE STAR | 3 | 0.42 | 15472 | United Kingdom |
| 1 | 574076 | 23485 | BOTANICAL GARDENS WALL CLOCK | 1 | 49.96 | <NA> | United Kingdom |
| 2 | 546417 | 20996 | JAZZ HEARTS ADDRESS BOOK | 24 | 0.42 | 14800 | United Kingdom |
| 3 | 549586 | 22666 | RECIPE BOX PANTRY YELLOW DESIGN | 1 | 6.63 | <NA> | United Kingdom |
| 4 | 568197 | 20713 | JUMBO BAG OWLS | 10 | 2.08 | 16746 | United Kingdom |



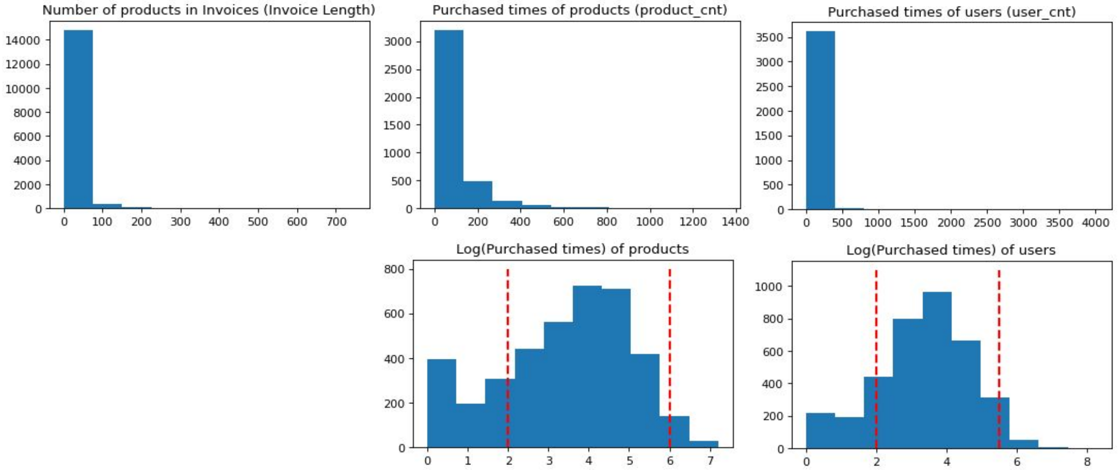

The number of lines (left) and stocks (right) in countries

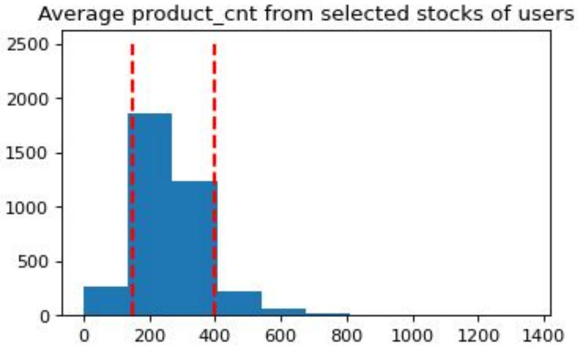
# Classifying

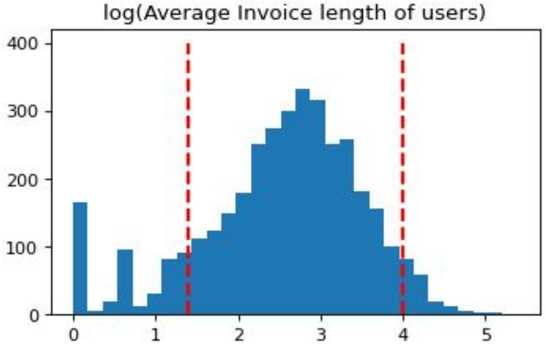
Extract some information

- Number of stocks in invoices (invoice length). 91% are less than 50 items.
- Product aspects: number of purchased times (product_cnt)
- User aspects: number of purchasing times (user_cnt), average of product_cnt from selected stocks of users, average length of invoices of users

These values can be helpful in classifying users, products.



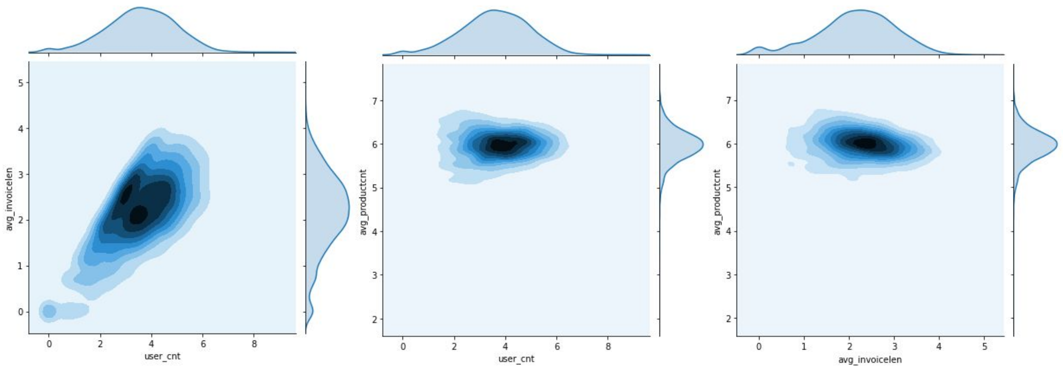
# Users and Stocks

Selecting user features:

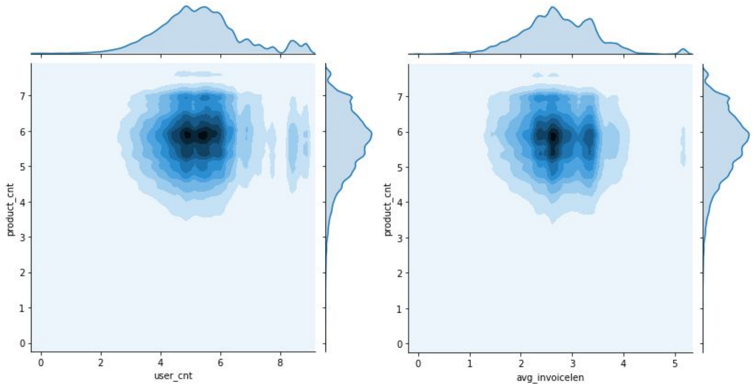
- user_cnt and avg_invoicelen have high correlation -> no need to use avg_invoicelen
- user_cnt and avg product_cnt seem to be independent to each other

Relationship with stock feature (product_cnt)

- No special pattern




Visualizing relationship between user features



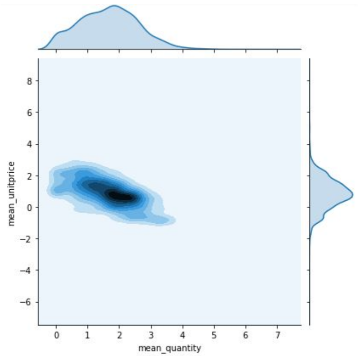

Visualizing relationship between user_cnt and product_cnt

# Quantity and UnitPrice

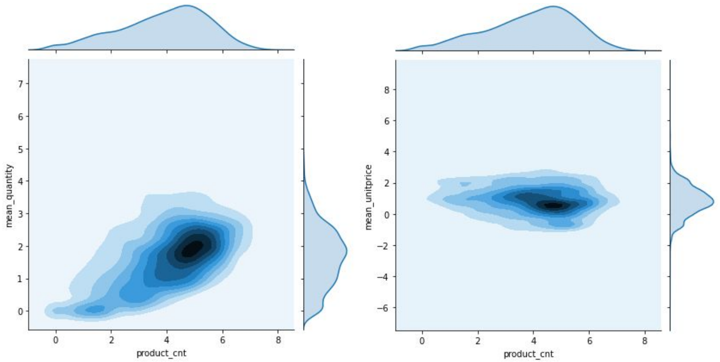
In general, in all stocks and purchases, higher unitprice, lower quantity.

Stocks which have higher product_cnt have higher quantity. This is nature, thus this relationship doesn’t bring insights.

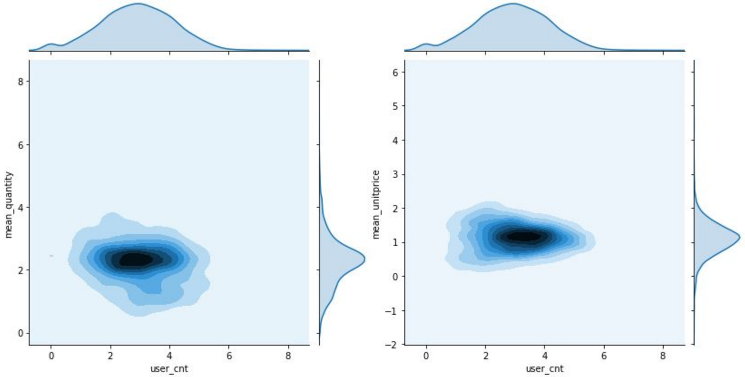
There is no special pattern between features of users and unitprice or quantity.




Relationship between average quantity and unitprice of all stocks




Relationship between average quantity/unitprice and purchased times of stocks (product_cnt)




Relationship between average quantity/unitprice and purchased times of user (user_cnt)

# Conclusion

- We have near 4000 products, but most of invoice have less than 50 items. This is natural in real-life, thus an obvious challenge of recommendation: **it's difficult to predict** the next choice of users, connection between users and items are sparse, and small rate of selection (**imbalanced output classes**).
- **No need to make recommendation for each country**, since data is strongly dominant by UK. Other countries may not have enough data for training
- All patterns in data are nature in services, therefore we don't have specific cases, **no need to customize with specific cases** of users or products.
- Because of no special patterns, **random subsets of data with large enough size** can reflect properties of the large data. This is helpful for training and testing methods.
  - E.g: using set of high-length invoices only can reduce sparsity but not lose too much information.
- Because of no special pattern, we lose some advantages of applying popular traditional methods, such as recommending only trending products is widely applied, especially when data distribution is matching with the Pareto principle (e.g Top 20% products contributes 80% purchases). **Carefully selecting strategies for methods is necessary.**

# 3. Methods review and Planning

1. Main models: Content-based and collaborative Filtering
2. Hybrid models
3. Post-processing
4. Offline evaluation
5. Selecting appropriate methods

# 3. Main models: Content-based and Collaborative Filtering

**Content-based**: recommending similar items from current items. Similarity is calculated by comparing description or metadata of items.

Performance of this method depends on

- Quality of metadata, description
- Pre-processing metadata, description
- Calculating similarity between items

Advantages: do not depend on density of connection.

**Collaborative Filtering**: find similar items or users based on the history:

- Item-item techniques: find similarity between items to make recommendation
- User-item techniques: find similarity between users then recommending unselected items from considering selected lists of similar users.

Both 2 types of techniques of Collaborative Filtering have 2 approaches:

- **Memory-based**: discover underlying patterns between users and items, have high accuracy.
  Traditional methods of this approach (e.g. *Matrix Factorization*) has low scalability due to high complexity. Recently, *graph-based methods* overcome these limitation.

- **Model based**: approximate memory-based methods to have better scalability but low accuracy (e.g ALS). This approach is just suitable for small system.

These techniques do not depend on content or explicit information.
Performance of these methods depend on

- Density of history

Advantages:

- do not depend on explicit information such as description

# The hybrid

Even a model can be tuned with different parameters with different strategies to deal with particular cases of data. Sometimes, we also need consider score from particular properties. E.g: we want to introduce more items on topic 16 to users, but they has low product_cnt, leads to rarely be recommended. In this case, a simple solution is using a score which imposes high priority on them.

Step by step to hybrid scores:

- Scale scores into the same range
- Manually tuning
- Automatic hybrid
  - Using ensemble methods: bagging or boosting (AdaBoost, XGBoost, GradBoost)...
  - Training a neural network

# Post-processing

Post-processing is applied to:

- Deal with cases which do not have enough information as input to feed to models, such as the current items are new, don't have enough data
- Smooth the results of models

Trending recommending is usually added to this part as default results for all cases of users and products.

# Offline evaluation

- Accuracy is reflected by precision.
- For long-term, some other aspect of recommendation system should be considered. Because recommending is an intervention to data in system. After applying recommending, the data distribution may be affected, such as:
  - parameters of methods maybe no longer appropriate.
  - Recommending too much on small set of products
  - Low personalization: No difference between recommending times of the same users or between users.

| Metric | | Meaning |
|---|---|---|
| **precision** | higher is better | The rate reflects how correct of recommendation |
| **popularity** | | How popular the products are. Calculating popularity of products, then averaging all products that are recommended. |
| **diversity** | higher is better | Personalization ability of recommendation: difference on recommending results between users |
| **coverage** | higher is better | rate of products recommended at least to 1 time |
| **congestion** | lower is better | difference on number of recommending times between products |

# Selecting appropriate methods

1. **Content-based:**

- Processing description to create vectors representing properties of products.
  - Define a list of properties: places (office, kitchen, garden), season (Christmas, summer, winter), time (modern or traditional), objectives (students, elder, pet)... With each property, define a set of decision rules and represent items as binary vector from description.
  - Run topic model algorithms to represent items as numeric vector of topics.
  - Use word2vec to represent words in description as numeric vector.
  - Classify items by their quantity and unitprice
- Calculating distance between vectors to show how related between items
  - Euclid distance
  - Cosine distance
  - ...

# Selecting appropriate methods

**2. Collaborative Filtering**

- Should not use model-based
  - Approximating memory-based -> less accuracy
  - The data size is large while connection between users and products is sparse.
- Traditional memory-based methods
  - can be used to validate graph-based methods and pre-calculated in offline to assist real-time functions.
- Use graph memory-based methods
  - Dealing with the sparse connection between users and products
  - Easy to compute and explain results
  - Considering personalization

# Selecting appropriate methods

**3. The hybrid**

- Since the data is extremely sparse, which leads to imbalance output classes, we should use simple model first for easier explaining the results, then apply more advanced techniques later.
- The other important problem is defining input, output format, selecting appropriate loss function for the hybrid model:
  - Because we don’t have time of each purchases, so that we can not group all purchases by users and split them to historical data as input and current choices as output.
  - Instead, I use Invoice. The input is set of each item in the invoice, and the output is the rest of items in the invoice. Thus, each invoice with n items bring the input n lines. Detail in section 5.
  - When users selected items, there is no information reflects how they like products. Therefore our labels are binary. Further processing to generate term of “rating” can be considered, such as from normalizing all quantity of all users with the same products.

# Conclusion

Selected methods

- Content-based: preprocessing and feature extraction from Description, then calculating similarity between items
- Collaborative Filtering: calculate implicit relationships of items and users by graph memory-based models. Considering personalization
- The hybrid: use Invoices to generate input and output, train a hybrid model simply by neural network.

# 4. Pre-processing and Feature Extraction

Fortunately, there is a marginal difference between full data and training data on StockCode and Description.

- Some StockCode have both lower and upper character -> fix them
- Some StockCode have multiple description -> keep the first description only

```
Number of unique StockCode and Description
 + Data full:
         field  Unique Count
0    StockCode          4070
1  Description          4223

 + Training data:
         field  Unique Count
0    StockCode          3923
1  Description          4036
```

| StockCode | Description |
|---|---|
| 84558A | 3D DOG PICTURE PLAYING CARDS |
| 84558a | 3D DOG PICTURE PLAYING CARDS |
| 85184C | S/4 VALENTINE DECOUPAGE HEART BOX |
| 85184C | SET 4 VALENTINE DECOUPAGE HEART BOX |

# 4. Pre-processing and Feature Extraction

- Applying topic model algorithm (LDA):
  - Grouping stock by Invoices
  - Considering invoices as document, products as words
  - Experiments with number of topics = [4, 8, 12, 16] … Select number of topics=16
  - Products are assigned to topics from 0 to 15. Products with no topics are assign as 16.

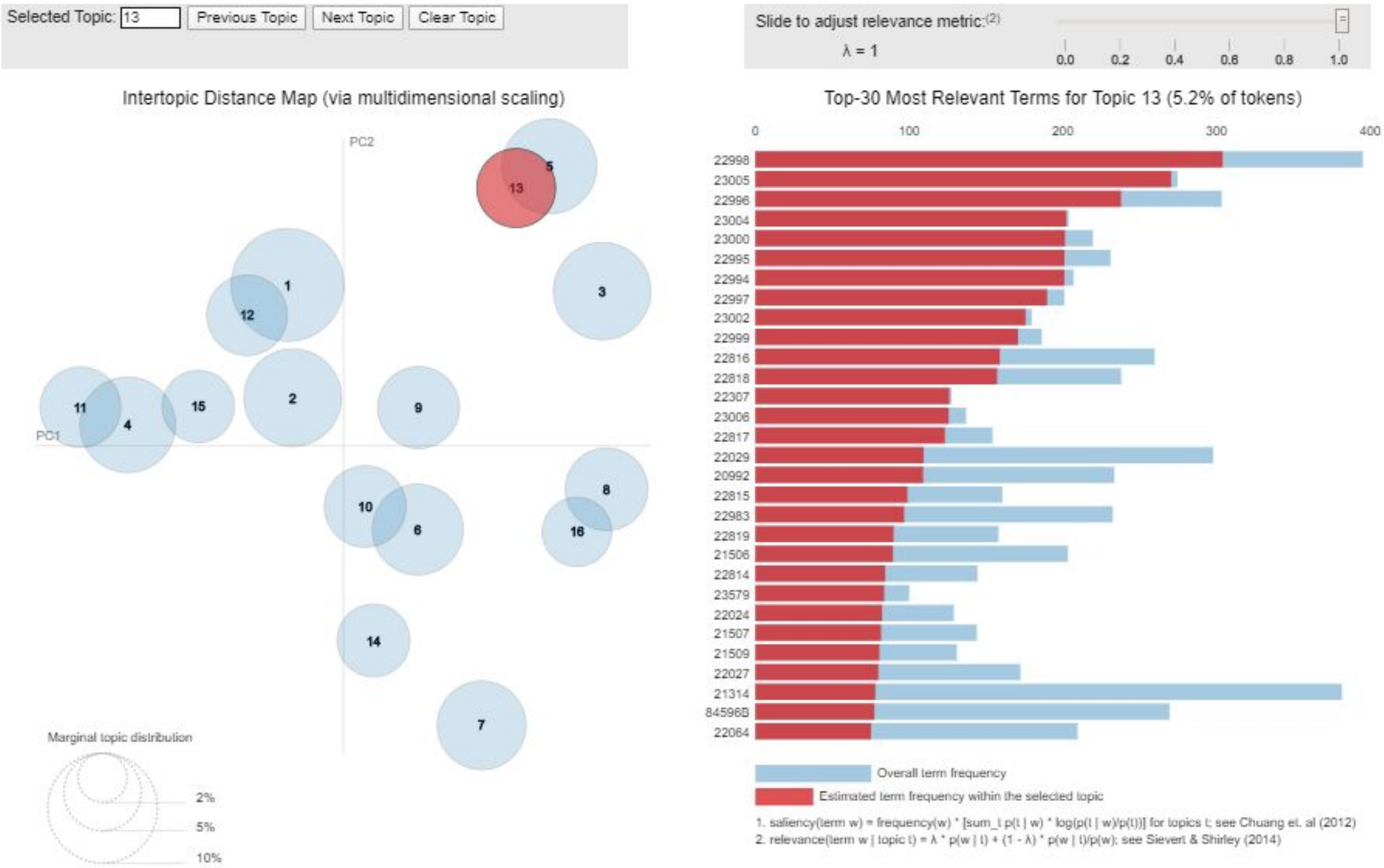
Results of clustering Products in the next slides.

| | InvoiceNo | Products |
|---|---|---|
| 0 | 536366 | [22633, 22632] |
| 1 | 536367 | [84879, 22745, 22748, 22749, 22310, 84969, 226... |
| 2 | 536370 | [22728, 22727, 22726, 21724, 21883, 10002, 217... |
| 3 | 536374 | [21258] |
| 4 | 536375 | [85123A, 71053, 84406B, 20679, 37370, 21871, 2... |

```
Number of products in topics 0: 885
Number of products in topics 1: 736
Number of products in topics 2: 929
Number of products in topics 3: 952
Number of products in topics 4: 721
Number of products in topics 5: 711
Number of products in topics 6: 679
Number of products in topics 7: 972
Number of products in topics 8: 666
Number of products in topics 9: 616
Number of products in topics 10: 754
Number of products in topics 11: 824
Number of products in topics 12: 1147
Number of products in topics 13: 657
Number of products in topics 14: 893
Number of products in topics 15: 713
Number of products with no topics:  631
```

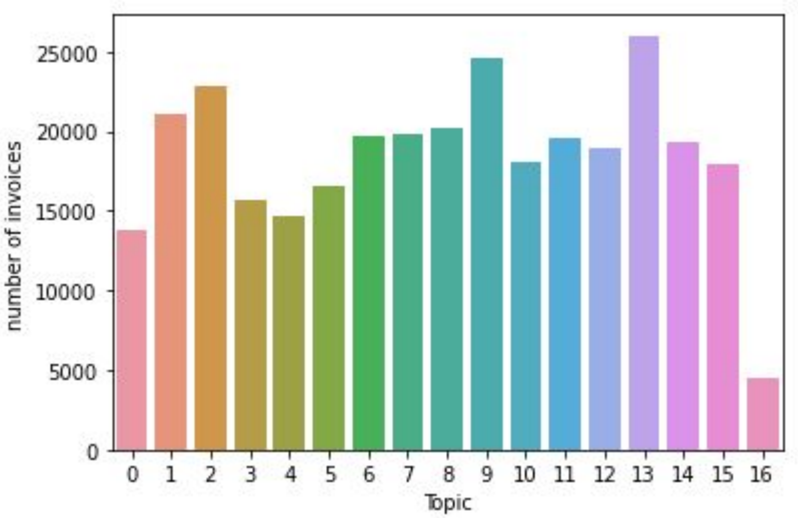
| | Topic #01 | Topic #02 | Topic #03 | Topic #04 | Topic #05 | Topic #06 | Topic #07 | Topic #08 | Topic #09 | Topic #10 | Topic #11 | Topic #12 | Topic #13 | Topic #14 | Topic #15 | Topic #16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | KEY FOB , BACK DOOR | BAKING SET 9 PIECE RETROSPOT | WOODEN STAR CHRISTMAS SCANDINAVIAN | TRAVEL CARD WALLET KEEP CALM | ANTIQUE SILVER TEA GLASS ETCHED | ALARM CLOCK BAKELIKE GREEN | JUMBO BAG RED RETROSPOT | PINK REGENCY TEACUP AND SAUCER | SET OF 3 CAKE TINS PANTRY DESIGN | LUNCH BAG RED RETROSPOT | POSTAGE | PACK OF 72 RETROSPOT CAKE CASES | CHRISTMAS PUDDING TRINKET POT | ASSORTED COLOUR BIRD ORNAMENT | DOTCOM POSTAGE | HAND WARMER OWL DESIGN |
| 1 | KEY FOB , SHED | VINTAGE SNAP CARDS | PAPER CHAIN KIT 50'S CHRISTMAS | TRAVEL CARD WALLET I LOVE LONDON | BLUE STRIPE CERAMIC DRAWER KNOB | ALARM CLOCK BAKELIKE RED | JUMBO BAG PINK POLKADOT | ROSES REGENCY TEACUP AND SAUCER | JAM MAKING SET WITH JARS | LUNCH BAG BLACK SKULL. | RABBIT NIGHT LIGHT | PACK OF 60 DINOSAUR CAKE CASES | 4 VANILLA BOTANICAL CANDLES | REGENCY CAKESTAND 3 TIER | WRAP CHRISTMAS VILLAGE | DOORMAT KEEP CALM AND COME IN |
| 2 | MAGNETS PACK OF 4 SWALLOWS | TRADITIONAL KNITTING NANCY | SET OF 20 VINTAGE CHRISTMAS NAPKINS | TRAVEL CARD WALLET VINTAGE TICKET | RED STRIPE CERAMIC DRAWER KNOB | ALARM CLOCK BAKELIKE PINK | JUMBO BAG BAROQUE BLACK WHITE | CHOCOLATE HOT WATER BOTTLE | RECIPE BOX PANTRY YELLOW DESIGN | ZINC FOLKART SLEIGH BELLS | SET OF 4 KNICK KNACK TINS DOILEY | 72 SWEETHEART FAIRY CAKE CASES | FLORAL FOLK STATIONERY SET | WHITE HANGING HEART T-LIGHT HOLDER | JUMBO BAG WOODLAND ANIMALS | HAND WARMER UNION JACK |
| 3 | KEY FOB , GARAGE DESIGN | CHRISTMAS CRAFT LITTLE FRIENDS | SET OF 3 WOODEN HEART DECORATIONS | TRAVEL CARD WALLET PANTRY | RED SPOT CERAMIC DRAWER KNOB | SOLDIERS EGG CUP | JUMBO SHOPPER VINTAGE RED PAISLEY | GIN + TONIC DIET METAL SIGN | SET OF 4 PANTRY JELLY MOULDS | LUNCH BAG CARS BLUE | BOX OF 6 MINI 50'S CRACKERS | PACK OF 72 SKULL CAKE CASES | DARK BIRD HOUSE TREE DECORATION | ANTIQUE SILVER TEA GLASS ETCHED | RECYCLING BAG RETROSPOT | HOT WATER BOTTLE KEEP CALM |
| 4 | 36 PENCILS TUBE RED RETROSPOT | PAPER CHAIN KIT 50'S CHRISTMAS | JUMBO BAG 50'S CHRISTMAS | TRAVEL CARD WALLET TRANSPORT | BLUE SPOT CERAMIC DRAWER KNOB | BICYCLE PUNCTURE REPAIR KIT | JUMBO STORAGE BAG SUKI | GREEN REGENCY TEACUP AND SAUCER | SET OF 6 SPICE TINS PANTRY DESIGN | LUNCH BAG SUKI DESIGN | CHRISTMAS LIGHTS 10 REINDEER | PACK OF 60 SPACEBOY CAKE CASES | SET OF 4 ROSE BOTANICAL CANDLES | HEART OF WICKER LARGE | LARGE CIRCULAR MIRROR MOBILE | HAND WARMER RED LOVE HEART |

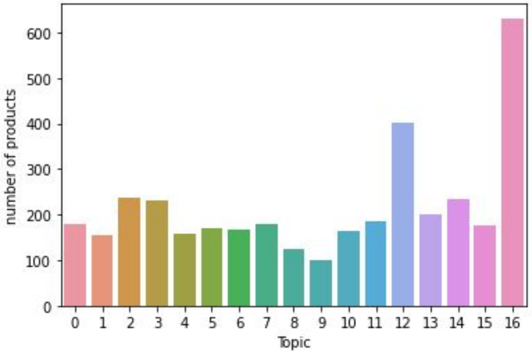

Topic 0
36 PENCILS TUBE SKULLS
36 PENCILS TUBE RED RETROSPOT
PLASTERS IN TIN SKULLS
MAGNETS PACK OF 4 HOME SWEET HOME
MAGNETS PACK OF 4 SWALLOWS
KEY FOB , SHED
KEY FOB , FRONT DOOR
KEY FOB , GARAGE DESIGN
PLASTERS IN TIN VINTAGE PAISLEY
KEY FOB , BACK DOOR
Topic 1
BAKING SET 9 PIECE RETROSPOT
PAPER CHAIN KIT 50'S CHRISTMAS
TRADITIONAL KNITTING NANCY
FELTCRAFT PRINCESS CHARLOTTE DOLL
CHRISTMAS CRAFT TREE TOP ANGEL
FELTCRAFT PRINCESS LOLA DOLL
PINK CREAM FELT CRAFT TRINKET BOX
VINTAGE SNAP CARDS
CHRISTMAS CRAFT LITTLE FRIENDS
FELTCRAFT CUSHION OWL
Topic 2
JUMBO BAG VINTAGE CHRISTMAS
SET OF 20 VINTAGE CHRISTMAS NAPKINS
SET OF 3 WOODEN HEART DECORATIONS
SET OF 3 WOODEN STOCKING DECORATION
JUMBO BAG 50'S CHRISTMAS
PAPER CHAIN KIT 50'S CHRISTMAS
PAPER CHAIN KIT VINTAGE CHRISTMAS
SET OF 3 WOODEN TREE DECORATIONS
60 CAKE CASES VINTAGE CHRISTMAS
WOODEN STAR CHRISTMAS SCANDINAVIAN
Topic 3
TRAVEL CARD WALLET TRANSPORT
TRAVEL CARD WALLET SKULLS
TRAVEL CARD WALLET PANTRY
TRAVEL CARD WALLET RETROSPOT
TRAVEL CARD WALLET I LOVE LONDON
TRAVEL CARD WALLET VINTAGE TICKET
TRAVEL CARD WALLET RETRO PETALS
TRAVEL CARD WALLET KEEP CALM
TRAVEL CARD WALLET UNION JACK
TRAVEL CARD WALLET SUKI
Topic 4
CLEAR DRAWER KNOB ACRYLIC EDWARDIAN
ANTIQUE SILVER TEA GLASS ENGRAVED
BLUE SPOT CERAMIC DRAWER KNOB
WHITE SPOT BLUE CERAMIC DRAWER KNOB
BLUE STRIPE CERAMIC DRAWER KNOB
RED STRIPE CERAMIC DRAWER KNOB
ANTIQUE SILVER TEA GLASS ETCHED
RED SPOT CERAMIC DRAWER KNOB
WHITE SPOT RED CERAMIC DRAWER KNOB
MULTI COLOUR SILVER T-LIGHT HOLDER
Topic 5
BICYCLE PUNCTURE REPAIR KIT
SOLDIERS EGG CUP
CLASSIC BICYCLE CLIPS
ALARM CLOCK BAKELIKE GREEN
ALARM CLOCK BAKELIKE CHOCOLATE
ALARM CLOCK BAKELIKE PINK
ALARM CLOCK BAKELIKE RED
ALARM CLOCK BAKELIKE IVORY
LONDON BUS COFFEE MUG
LUNCH BOX I LOVE LONDON
Topic 6
JUMBO BAG BAROQUE BLACK WHITE
RED RETROSPOT CAKE STAND
LANTERN CREAM GAZEBO
WHITE WOOD GARDEN PLANT LADDER
JUMBO BAG PINK POLKADOT
JUMBO STORAGE BAG SUKI
JUMBO BAG PINK VINTAGE PAISLEY
JUMBO SHOPPER VINTAGE RED PAISLEY
JUMBO BAG RED RETROSPOT
CREAM SWEETHEART MINI CHEST
Topic 7
HOT WATER BOTTLE I AM SO POORLY
PLEASE ONE PERSON METAL SIGN
COOK WITH WINE METAL SIGN
GREEN REGENCY TEACUP AND SAUCER
ROSES REGENCY TEACUP AND SAUCER
GIN + TONIC DIET METAL SIGN
HOT WATER BOTTLE TEA AND SYMPATHY
PINK REGENCY TEACUP AND SAUCER
HOT WATER BOTTLE KEEP CALM
CHOCOLATE HOT WATER BOTTLE
Topic 8
RECIPE BOX PANTRY YELLOW DESIGN
SAMPLES
JAM MAKING SET PRINTED
SET OF TEA COFFEE SUGAR TINS PANTRY
REGENCY CAKESTAND 3 TIER
SET OF 4 PANTRY JELLY MOULDS
SET OF 6 SPICE TINS PANTRY DESIGN
JAM MAKING SET WITH JARS
STRAWBERRY CERAMIC TRINKET BOX
SET OF 3 CAKE TINS PANTRY DESIGN
Topic 9
ZINC FOLKART SLEIGH BELLS
LUNCH BAG SUKI DESIGN
LUNCH BAG CARS BLUE
LUNCH BAG SPACEBOY DESIGN
JUMBO BAG RED RETROSPOT
RED RETROSPOT CHARLOTTE BAG
LUNCH BAG WOODLAND
LUNCH BAG RED RETROSPOT
LUNCH BAG BLACK SKULL.
LUNCH BAG PINK POLKADOT
Topic 10
PLASTERS IN TIN CIRCUS PARADE
BOX OF 6 MINI 50'S CRACKERS
SET OF 4 KNICK KNACK TINS DOILEY
POSTAGE
SET OF 5 LUCKY CAT MAGNETS
CHRISTMAS LIGHTS 10 REINDEER
BLUE HARMONICA IN BOX
RABBIT NIGHT LIGHT
PLASTERS IN TIN SPACEBOY
ROUND SNACK BOXES SET OF4 WOODLAND
Topic 11
72 SWEETHEART FAIRY CAKE CASES
PACK OF 60 SPACEBOY CAKE CASES
PACK OF 60 DINOSAUR CAKE CASES
60 TEATIME FAIRY CAKE CASES
SET/20 RED RETROSPOT PAPER NAPKINS
PACK OF 72 SKULL CAKE CASES
SMALL POPCORN HOLDER
60 CAKE CASES DOLLY GIRL DESIGN
PACK OF 60 PINK PAISLEY CAKE CASES
PACK OF 72 RETROSPOT CAKE CASES
Topic 12
FLORAL FOLK STATIONERY SET
DARK BIRD HOUSE TREE DECORATION
WOODEN HEART CHRISTMAS SCANDINAVIAN
CHRISTMAS PUDDING TRINKET POT
4 PEAR BOTANICAL DINNER CANDLES
FOLKART CLIP ON STARS
4 LAVENDER BOTANICAL DINNER CANDLES
4 VANILLA BOTANICAL CANDLES
MODERN FLORAL STATIONERY SET
SET OF 4 ROSE BOTANICAL CANDLES
Topic 13
WHITE HANGING HEART T-LIGHT HOLDER
ASSORTED COLOUR BIRD ORNAMENT
ANTIQUE SILVER TEA GLASS ETCHED
VICTORIAN GLASS HANGING T-LIGHT
NATURAL SLATE HEART CHALKBOARD
IVORY DINER WALL CLOCK
REGENCY CAKESTAND 3 TIER
HEART OF WICKER LARGE
SMALL WHITE HEART OF WICKER
SILVER HANGING T-LIGHT HOLDER
Topic 14
RED TOADSTOOL LED NIGHT LIGHT
SUKI SHOULDER BAG
DOTCOM POSTAGE
PIZZA PLATE IN BOX
RECYCLING BAG RETROSPOT
JUMBO STORAGE BAG SUKI
PINK REGENCY TEACUP AND SAUCER
JUMBO BAG WOODLAND ANIMALS
LARGE CIRCULAR MIRROR MOBILE
WRAP CHRISTMAS VILLAGE
Topic 15
HOT WATER BOTTLE KEEP CALM
HAND WARMER RED POLKA DOT
HAND WARMER UNION JACK
HAND WARMER RED LOVE HEART
HAND WARMER SCOTTY DOG DESIGN
DOORMAT VINTAGE LEAVES DESIGN
DOORMAT RED RETROSPOT
DOORMAT UNION FLAG
HAND WARMER OWL DESIGN
DOORMAT KEEP CALM AND COME IN

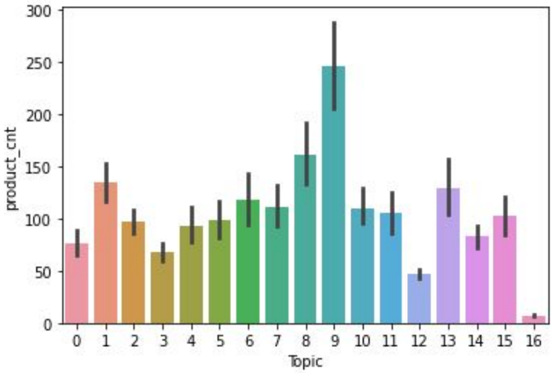

Selected Topic: 13
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 13 (5.2% of tokens)
0
100
200
300
400
22998
23005
22996
23004
23000
22995
22994
22997
23002
22999
22816
22818
22307
23006
22817
22029
20992
22815
22983
22819
21506
22814
23579
22024
21507
21509
22027
21314
84596B
22064
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

# Check topics on purchases and products

- Visualize the number of invoices on topics: topic 16 has the lowest number of invoices
- Visualize the number of products on topics and product_cnt of products on topics: topic 16 has the most number of products, all of them have very low product_cnt.



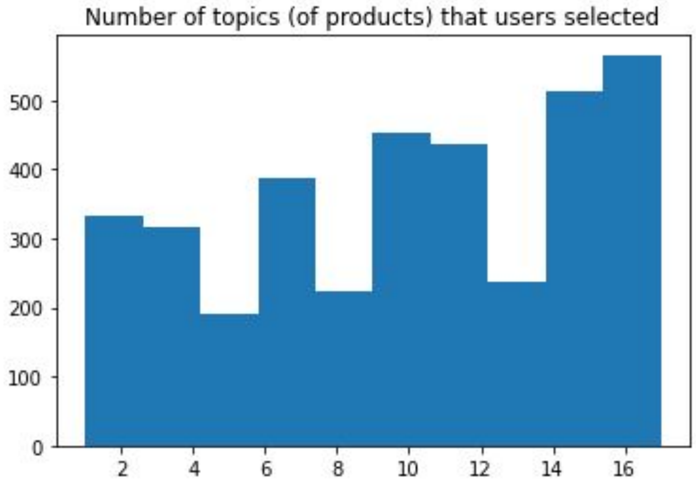



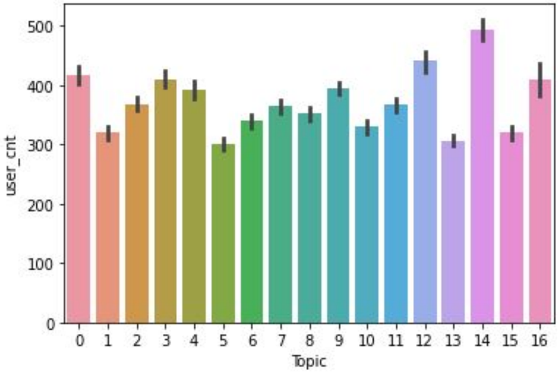



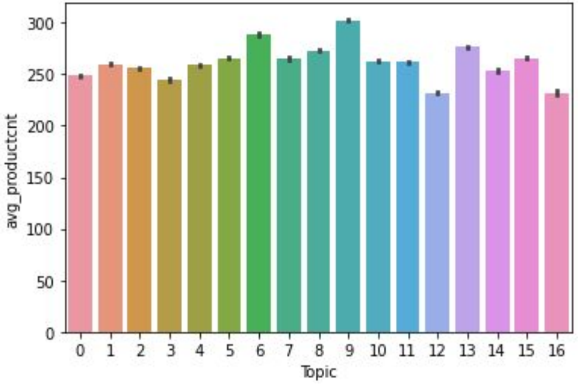
# Check topics on users

- Users selected products from many topics. Topics don’t have too much difference on user features (user_cnt and avg_productcnt).

=> Purchases, products, user activities are divided widely on all topics, no dominant topic or no too small topic.



## Grouping products by topics

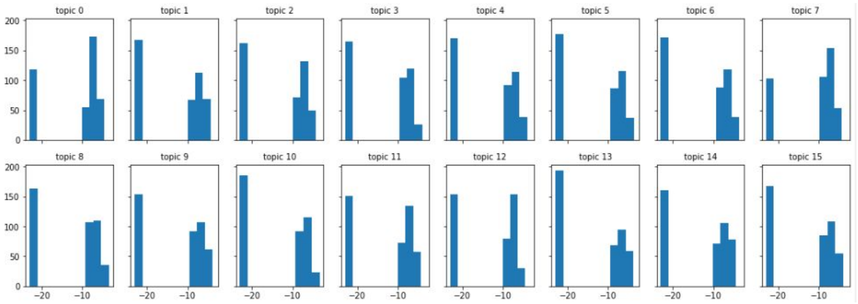
Consider products with at least 1 topic, dividing products by range of SumProb:

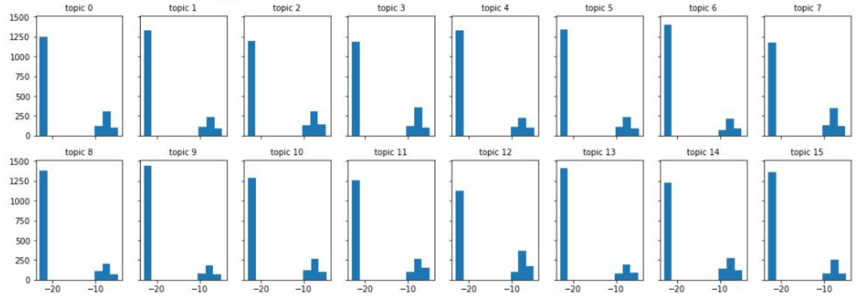
- High-prob products: SumProb>=1e-2
- Middle-prob products: 1e-2>SumProb>=1e-3
- Low-prob products: 1e-2>SumProb>=1e-3

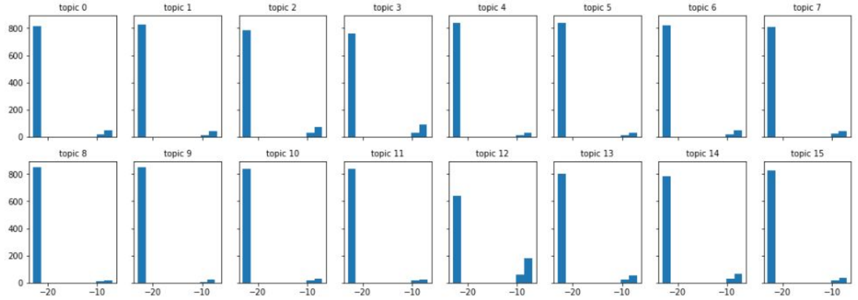
Visualizing log transformation of probabilities of High/Middle/Low-prob products on topics:

- 415 high-prob products have both high and low probabilities
- 1771 middle-prob products have both high and low probabilities, the number of low probabilities is higher
- 879 low-prob products mainly have low probabilities.

=> Values of probabilities have the same range on all topics.



# Conclusion

- Can generate clusters, corresponding to topics from topic model algorithms. This case uses LDA.
- Purchases, users and products spread widely on topics. No dominant topic, no too small topic.
- Values of probabilities on topics have the same range on all topics -> No need to add weights on topics.

# 5. Recommending

- Content-based recommending from topics
- Graph Memory-based Collaborative Filtering
- Recommending with user history
- The hybrid

# Content-based Recommending from Topics

Mechanism: base on similarity on description information of products:

- Each items are represented as 16-vector of probabilities on topics
- Calculating similarity between items by a distance metric on vectors
    - Euclid_distance: temporary use this metric. In fact, the selection of distance metric must be considered with the distribution of products on topics. Advantages of euclid distance is not considering order of topics, well reflecting the distance between items in the same topics.
- Recommending items with lowest distance

Recommending products with topic=16 can be wrong, because all probabilities on topics are 0.

| | start_code | rec_code | score | start_des | rec_des |
|---|---|---|---|---|---|
| 0 | 23344 | 23344 | 0.000000 | JUMBO BAG 50'S CHRISTMAS | JUMBO BAG 50'S CHRISTMAS |
| 1 | 23344 | 23343 | 0.000019 | JUMBO BAG 50'S CHRISTMAS | JUMBO BAG VINTAGE CHRISTMAS |
| 2 | 23344 | 22910 | 0.000026 | JUMBO BAG 50'S CHRISTMAS | PAPER CHAIN KIT VINTAGE CHRISTMAS |
| 3 | 23344 | 23313 | 0.000045 | JUMBO BAG 50'S CHRISTMAS | VINTAGE CHRISTMAS BUNTING |
| 4 | 23344 | 23202 | 0.000047 | JUMBO BAG 50'S CHRISTMAS | JUMBO BAG VINTAGE LEAF |
| 5 | 23344 | 22086 | 0.000050 | JUMBO BAG 50'S CHRISTMAS | PAPER CHAIN KIT 50'S CHRISTMAS |

Recommend for StockCode=23344, description='JUMBO bAG 50'S CHRISTMAS'. Distance to itself=0 is in the first line.

| | start_code | rec_code | score | start_des | rec_des |
|---|---|---|---|---|---|
| 0 | 10002 | 21769 | 0.000000e+00 | INFLATABLE POLITICAL GLOBE | VINTAGE POST OFFICE CABINET |
| 1 | 10002 | 90196B | 0.000000e+00 | INFLATABLE POLITICAL GLOBE | BLACK GEMSTONE NECKLACE 45CM |
| 2 | 10002 | 85146 | 0.000000e+00 | INFLATABLE POLITICAL GLOBE | JARDIN ETCHED GLASS SMALL BELL JAR |
| 3 | 10002 | 21278 | 0.000000e+00 | INFLATABLE POLITICAL GLOBE | VINTAGE KITCHEN PRINT PUDDINGS |
| 4 | 10002 | 16169P | 0.000000e+00 | INFLATABLE POLITICAL GLOBE | WRAP GREEN RUSSIAN FOLKART |
| 5 | 10002 | 90077 | 0.000000e+00 | INFLATABLE POLITICAL GLOBE | BLACK DIAMOND CLUSTER EARRINGS |

Recommend for a product in topic 16 with StockCode=10002. Beside of distance to itself, other items have 0-distance.

# Graph Memory-based Collaborative Filtering

Mechanism: Extract information “People buy x also buy y” - implicit hidden relationship.

Step-by-step:

- Constructing graph
- Calculating the score reflecting relationship between users by similarity function
- Calculating the score reflecting relationship between items and users selected them -> user-item recommending
- Calculate the score reflecting relationship between items -> item-item recommending

I re-write these step by matrix computing on the right.

For more information about the method, check here.

Let $k_\alpha$ is purchased times of items $\alpha$.

Let call $A$ is history as matrix. if user $j$ selected item $\alpha$:

$$a_{j\alpha} = 1$$

Calculating $T$ is weights of products $\alpha$ in history of each user $j$:

$$t_{j\alpha} = \frac{a_{j\alpha} k_\alpha^\gamma}{\sum_\beta a_{j\beta} k_\beta^\gamma}$$

Calculating $H$ includes the weight when recommeding $\alpha$ for $\beta$:

$$h_{\alpha\beta} = \frac{1}{k_\alpha^{1-\lambda} k_\beta^\lambda}$$

Calculating $W$ is item2item matrix when recommeding $\alpha$ for $\beta$:

$$w_{\alpha\beta} = h_{\alpha\beta} \sum_j a_{j\alpha} a_{j\beta} t_{j\beta}$$

means:

$$W = H \circ (A^T \cdot (A \circ T))$$

Making Recommendation by $W$, larger score is better.

To recommend for all users:

$$R = AW^T$$

While Content-based uses distance to compute score, so that the lower is the better. In contrast, CF computes similarity score directly, thus the higher is the better.

It doesn't meet the challenge of zero-probability vector as Content-based.

| | start_code | rec_code | score | start_des | rec_des |
|---|---|---|---|---|---|
| 0 | 23344 | 23344 | 0.018848 | JUMBO BAG 50'S CHRISTMAS | JUMBO BAG 50'S CHRISTMAS |
| 1 | 23344 | 23343 | 0.008916 | JUMBO BAG 50'S CHRISTMAS | JUMBO BAG VINTAGE CHRISTMAS |
| 2 | 23344 | 45013 | 0.007101 | JUMBO BAG 50'S CHRISTMAS | FOLDING SHOE TIDY |
| 3 | 23344 | 23582 | 0.005697 | JUMBO BAG 50'S CHRISTMAS | VINTAGE DOILY JUMBO BAG RED |
| 4 | 23344 | 23532 | 0.005500 | JUMBO BAG 50'S CHRISTMAS | WALL ART WORK REST AND PLAY |
| 5 | 23344 | 23201 | 0.005070 | JUMBO BAG 50'S CHRISTMAS | JUMBO BAG ALPHABET |

Recommend for StockCode=23344.
Similarity score to itself is the highest, which is in the first line

| | start_code | rec_code | score | start_des | rec_des |
|---|---|---|---|---|---|
| 0 | 10002 | 10002 | 0.005555 | INFLATABLE POLITICAL GLOBE | INFLATABLE POLITICAL GLOBE |
| 1 | 10002 | 84881 | 0.002099 | INFLATABLE POLITICAL GLOBE | BLUE WIRE SPIRAL CANDLE HOLDER |
| 2 | 10002 | 21826 | 0.001689 | INFLATABLE POLITICAL GLOBE | EIGHT PIECE DINOSAUR SET |
| 3 | 10002 | 10123C | 0.001558 | INFLATABLE POLITICAL GLOBE | HEARTS WRAPPING TAPE |
| 4 | 10002 | 84745A | 0.001276 | INFLATABLE POLITICAL GLOBE | PINK HANGING GINGHAM EASTER HEN |
| 5 | 10002 | 84745B | 0.001276 | INFLATABLE POLITICAL GLOBE | BLUE HANGING GINGHAM EASTER HEN |

Recommend for a product in topic 16 with StockCode=10002.
Score to itself is still the highest, thus it still has recommending to itself first.

# Recommending with user history

Recommending with user history is helpful to improve personalization ability of system. Instead of recommending from current items only, items which are more related to history of users could have more priority.

The graph memory-based method can compute score of items from history directly. With content-based, this mechanism can be applied either:

- Compute recommending score for all items from each item on history of the user.
- Summarize (getting mean or sum) score to have the final recommending score for the user.

Note that we need to inverse score of content-based to obtain the same logic “higher is better”, then scale values to fixed range.

# The hybrid

- Combining any kind of score, regardless of score from models.
- Manually: After scaling all score to the same range, multiply score with weights and sum all of them.

Recommending $\alpha$ given $\beta$:

$$r_\alpha(\beta) = \sum_i w_i * s_i(\alpha, \beta)$$

where $s_i$ is score of one of previous models.

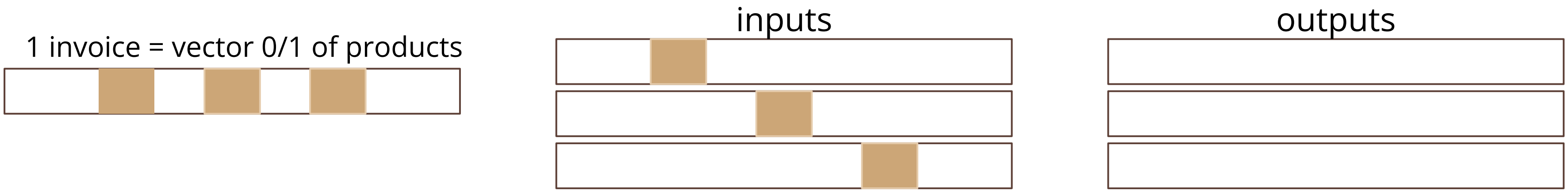
| | start_code | rec_code | score_bytopic | score_bygraph_item | score_bygraph_user | score | start_des | rec_des |
|---|---|---|---|---|---|---|---|---|
| 0 | 23344 | 23344 | 1.000000 | 1.000000 | 0.000000 | 0.900000 | JUMBO BAG 50'S CHRISTMAS | JUMBO BAG 50'S CHRISTMAS |
| 1 | 23344 | 23343 | 0.821380 | 0.401838 | 0.000000 | 0.529471 | JUMBO BAG 50'S CHRISTMAS | JUMBO BAG VINTAGE CHRISTMAS |
| 2 | 23344 | 22910 | 0.754555 | 0.004967 | 0.048553 | 0.309161 | JUMBO BAG 50'S CHRISTMAS | PAPER CHAIN KIT VINTAGE CHRISTMAS |
| 3 | 23344 | 22086 | 0.520251 | 0.088940 | 0.406638 | 0.293234 | JUMBO BAG 50'S CHRISTMAS | PAPER CHAIN KIT 50'S CHRISTMAS |
| 4 | 23344 | 23202 | 0.549560 | 0.109421 | 0.000000 | 0.274534 | JUMBO BAG 50'S CHRISTMAS | JUMBO BAG VINTAGE LEAF |
| 5 | 23344 | 23313 | 0.569009 | 0.017765 | 0.000000 | 0.236486 | JUMBO BAG 50'S CHRISTMAS | VINTAGE CHRISTMAS BUNTING |

The results of the hybrid considers more than 1 score which comes from different strategies. The example uses 0.4 * Content-based score + 0.5 * Graph CF item-item score + 0.1 * Graph CF user-item score

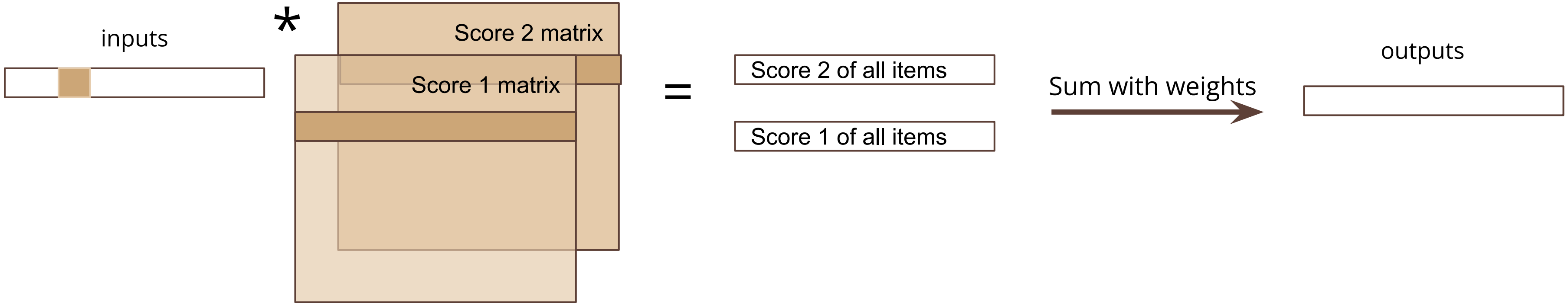
# The hybrid

- Training weights: instead of manually assigning values for weights, they can be obtained by learning from data. Input and output of model are generated by purchases (as the figure).
- Neural network: multiclass classification
  - V0 (Simple version): each type of score has a weight, similar to manually hybrid
  - V1: Each pair of score and product has a weights
  - Multi layer model with advanced techniques: BatchNorm, Dropout,...
- Loss function: CrossEntropyLoss, BCEWithLogitLoss are popular for multiclass classification. Mean Square Error (MSE) or Mean Absolute Error (MAE) can be applied also, then the problem becomes linear regression. This makes the model more flexible, since multiclass classification with imbalanced classes (because of >90% invoice length <=50) is challenged.

*I don’t have enough time, so that I’ve not try many types of loss function and optimized the model yet.*




Parsing purchases to data for training models




Explain computational steps in the simple model

# Simple version

The simple model is trained to learn weights of 4 types of score: item-item graph (CF), user-item graph (CF), content-based (CB) and user-item of CB.

CF scores have negative impact on the final score.

| | start_code | rec_code | item_graph_score | item_topic_score | user_graph_score | user_topic_score | score | start_des | rec_des |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 22139 | 23843 | 0.0 | 0.885991 | 0.0 | 0.990790 | 0.875275 | RETROSPOT TEA SET CERAMIC 11 PC | PAPER CRAFT , LITTLE BIRDIE |
| 1 | 22139 | 90133 | 0.0 | 0.887748 | 0.0 | 0.990476 | 0.875231 | RETROSPOT TEA SET CERAMIC 11 PC | TEAL/FUSCHIA COL BEAD NECKLACE |
| 2 | 22139 | 21736 | 0.0 | 0.889435 | 0.0 | 0.990855 | 0.875211 | RETROSPOT TEA SET CERAMIC 11 PC | GOLD SCROLL GLASS T-LIGHT HOLDER |
| 3 | 22139 | 84856S | 0.0 | 0.889321 | 0.0 | 0.990839 | 0.875202 | RETROSPOT TEA SET CERAMIC 11 PC | SMALL TAHITI BEACH BAG |
| 4 | 22139 | 90214Z | 0.0 | 0.887197 | 0.0 | 0.990477 | 0.875141 | RETROSPOT TEA SET CERAMIC 11 PC | LETTER "Z" BLING KEY RING |

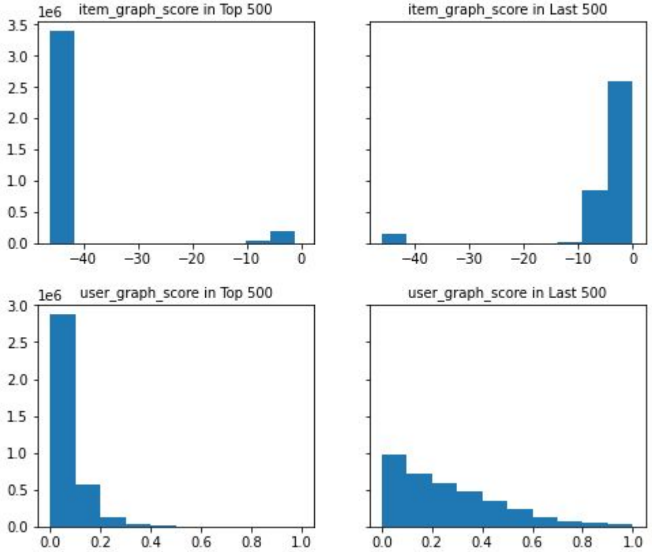
Top score of the model recommending for stock 22139

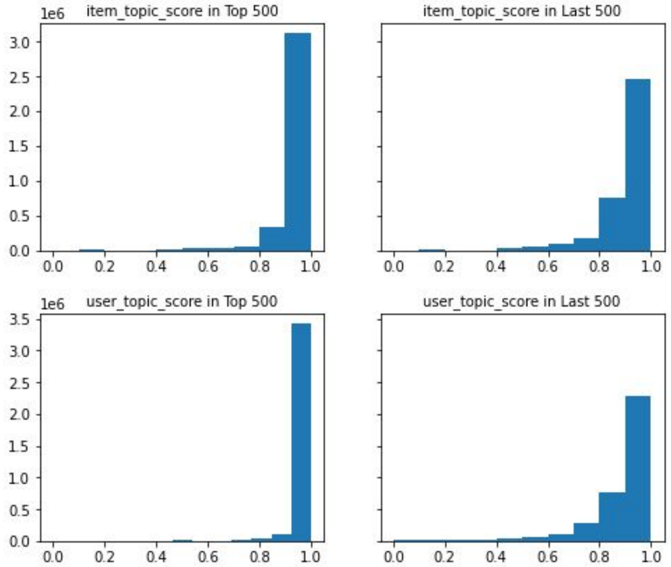
| | start_code | rec_code | item_graph_score | item_topic_score | user_graph_score | user_topic_score | score | start_des | rec_des |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 22139 | 85099B | 0.062929 | 0.530401 | 0.365359 | 0.157267 | 0.361837 | RETROSPOT TEA SET CERAMIC 11 PC | JUMBO BAG RED RETROSPOT |
| 1 | 22139 | POST | 0.057050 | 0.441938 | 0.377080 | 0.000000 | 0.405936 | RETROSPOT TEA SET CERAMIC 11 PC | POSTAGE |
| 2 | 22139 | 22720 | 0.070957 | 0.445397 | 0.420483 | 0.063872 | 0.414901 | RETROSPOT TEA SET CERAMIC 11 PC | SET OF 3 CAKE TINS PANTRY DESIGN |
| 3 | 22139 | 22960 | 0.075113 | 0.466284 | 0.397774 | 0.098664 | 0.428121 | RETROSPOT TEA SET CERAMIC 11 PC | JAM MAKING SET WITH JARS |
| 4 | 22139 | 22139 | 1.000000 | 1.000000 | 0.968411 | 0.847262 | 0.429055 | RETROSPOT TEA SET CERAMIC 11 PC | RETROSPOT TEA SET CERAMIC 11 PC |

Last score of the model recommending for stock 22139

Visualizing histogram of all score in top 500 recommending and all score in last 500 recommending of all items:

- High CB score (topic_score) place on top 500, High CF score place on last 500.



# Check the example case StockCode=23344

Recommend for StockCode=23344

| | start_code | rec_code | score | start_des | rec_des |
|---|---|---|---|---|---|
| 0 | 23344 | 22220 | 0.562569 | JUMBO BAG 50'S CHRISTMAS | CAKE STAND LOVEBIRD 2 TIER WHITE |
| 1 | 23344 | 21555 | 0.556116 | JUMBO BAG 50'S CHRISTMAS | CERAMIC STRAWBERRY TRINKET TRAY |
| 2 | 23344 | 22042 | 0.555378 | JUMBO BAG 50'S CHRISTMAS | CHRISTMAS CARD SINGING ANGEL |
| 3 | 23344 | 21189 | 0.555149 | JUMBO BAG 50'S CHRISTMAS | WHITE HONEYCOMB PAPER GARLAND |
| 4 | 23344 | 23070 | 0.554874 | JUMBO BAG 50'S CHRISTMAS | EDWARDIAN HEART PHOTO FRAME |
| 5 | 23344 | 46776D | 0.554805 | JUMBO BAG 50'S CHRISTMAS | WOVEN SUNSET CUSHION COVER |

Recommend for StockCode=23344 with CustomerID= 14911

| | start_code | rec_code | score | start_des | rec_des |
|---|---|---|---|---|---|
| 0 | 23344 | 90199B | 1.169672 | JUMBO BAG 50'S CHRISTMAS | 5 STRAND GLASS NECKLACE AMETHYST |
| 1 | 23344 | 21655 | 1.169275 | JUMBO BAG 50'S CHRISTMAS | HANGING RIDGE GLASS T-LIGHT HOLDER |
| 2 | 23344 | 90027D | 1.168822 | JUMBO BAG 50'S CHRISTMAS | GLASS BEAD HOOP EARRINGS AMETHYST |
| 3 | 23344 | 90134 | 1.168603 | JUMBO BAG 50'S CHRISTMAS | OLD ROSE COMBO BEAD NECKLACE |
| 4 | 23344 | 90013B | 1.168575 | JUMBO BAG 50'S CHRISTMAS | BLACK VINTAGE EARRINGS |
| 5 | 23344 | 23603 | 1.168435 | JUMBO BAG 50'S CHRISTMAS | SET 10 CARD KRAFT REINDEER 17084 |

# Version 1

The simple model is trained to learn weights of 4 types of score: item-item graph (CF), user-item graph (CF), content-based (CB) and user-item of CB.

CF scores have high positive impact on the final score.

| | start_code | rec_code | item_graph_score | item_topic_score | user_graph_score | user_topic_score | score | start_des | rec_des |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 22139 | 22633 | 0.105268 | 0.932371 | 0.283248 | 0.670111 | 1.0 | RETROSPOT TEA SET CERAMIC 11 PC | HAND WARMER UNION JACK |
| 1 | 22139 | 22563 | 0.162723 | 0.893098 | 0.240241 | 0.975546 | 1.0 | RETROSPOT TEA SET CERAMIC 11 PC | HAPPY STENCIL CRAFT |
| 2 | 22139 | 20754 | 0.165542 | 0.890693 | 0.254511 | 0.991168 | 1.0 | RETROSPOT TEA SET CERAMIC 11 PC | RETROSPOT RED WASHING UP GLOVES |
| 3 | 22139 | 22494 | 0.108484 | 0.885568 | 0.221975 | 0.979264 | 1.0 | RETROSPOT TEA SET CERAMIC 11 PC | EMERGENCY FIRST AID TIN |
| 4 | 22139 | 82600 | 0.114497 | 0.890245 | 0.320297 | 0.964981 | 1.0 | RETROSPOT TEA SET CERAMIC 11 PC | NO SINGING METAL SIGN |

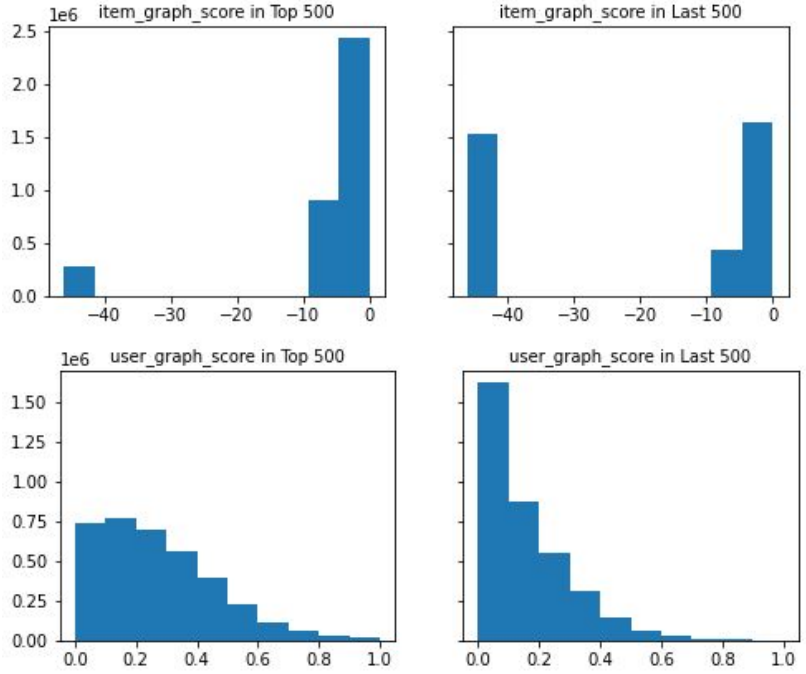
Top score of the model recommending for stock 22139

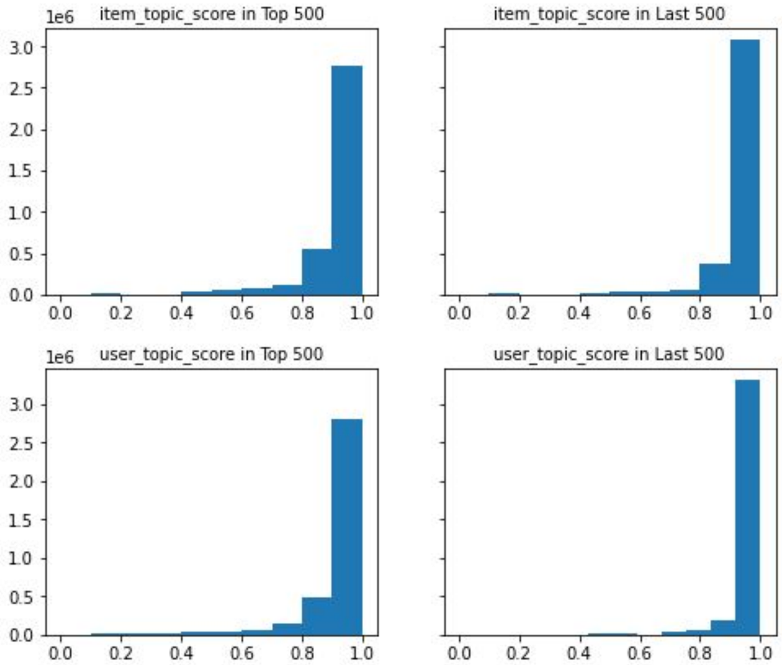
| | start_code | rec_code | item_graph_score | item_topic_score | user_graph_score | user_topic_score | score | start_des | rec_des |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 22139 | 84527 | 0.000000 | 0.886669 | 0.052293 | 0.990425 | 5.612948e-20 | RETROSPOT TEA SET CERAMIC 11 PC | FLAMES SUNGLASSES PINK LENSES |
| 1 | 22139 | 37489A | 0.012163 | 0.886086 | 0.064951 | 0.991523 | 5.840373e-20 | RETROSPOT TEA SET CERAMIC 11 PC | YELLOW/PINK FLOWER DESIGN BIG MUG |
| 2 | 22139 | 23311 | 0.148495 | 0.887868 | 0.388875 | 0.935223 | 6.012289e-20 | RETROSPOT TEA SET CERAMIC 11 PC | VINTAGE CHRISTMAS STOCKING |
| 3 | 22139 | 23414 | 0.078011 | 0.896558 | 0.244140 | 0.996642 | 6.045452e-20 | RETROSPOT TEA SET CERAMIC 11 PC | ZINC BOX SIGN HOME |
| 4 | 22139 | 23436 | 0.033786 | 0.901022 | 0.293640 | 0.986571 | 6.229080e-20 | RETROSPOT TEA SET CERAMIC 11 PC | GIFT BAG LARGE VINTAGE CHRISTMAS |

Last score of the model recommending for stock 22139

Visualizing histogram of all score in top 500 recommending and all score in last 500 recommending of all items:

- High CF score (topic_score) place on top 500, High CB score place on last 500.



# Check the example case StockCode=23344

Recommend for StockCode=23344

| | start_code | rec_code | score | start_des | rec_des |
|---|---|---|---|---|---|
| 0 | 23344 | 84824 | 0.022645 | JUMBO BAG 50'S CHRISTMAS | DANISH ROSE UMBRELLA STAND |
| 1 | 23344 | 78033 | 0.018501 | JUMBO BAG 50'S CHRISTMAS | FLAG OF ST GEORGE CHAIR |
| 2 | 23344 | 21412 | 0.015623 | JUMBO BAG 50'S CHRISTMAS | VINTAGE GOLD TINSEL REEL |
| 3 | 23344 | 23522 | 0.015461 | JUMBO BAG 50'S CHRISTMAS | WALL ART DOG AND BALL |
| 4 | 23344 | 90177E | 0.015369 | JUMBO BAG 50'S CHRISTMAS | DROP DIAMANTE EARRINGS GREEN |
| 5 | 23344 | 22910 | 0.015251 | JUMBO BAG 50'S CHRISTMAS | PAPER CHAIN KIT VINTAGE CHRISTMAS |

Recommend for StockCode=23344 with CustomerID= 14911

| | start_code | rec_code | score | start_des | rec_des |
|---|---|---|---|---|---|
| 0 | 23344 | 22372 | 0.032912 | JUMBO BAG 50'S CHRISTMAS | AIRLINE BAG VINTAGE WORLD CHAMPION |
| 1 | 23344 | 85049H | 0.031808 | JUMBO BAG 50'S CHRISTMAS | URBAN BLACK RIBBONS |
| 2 | 23344 | 22749 | 0.030397 | JUMBO BAG 50'S CHRISTMAS | FELTCRAFT PRINCESS CHARLOTTE DOLL |
| 3 | 23344 | 84992 | 0.030142 | JUMBO BAG 50'S CHRISTMAS | 72 SWEETHEART FAIRY CAKE CASES |
| 4 | 23344 | 21379 | 0.030087 | JUMBO BAG 50'S CHRISTMAS | CAMPHOR WOOD PORTOBELLO MUSHROOM |
| 5 | 23344 | 22791 | 0.030016 | JUMBO BAG 50'S CHRISTMAS | T-LIGHT GLASS FLUTED ANTIQUE |

# Conclusion

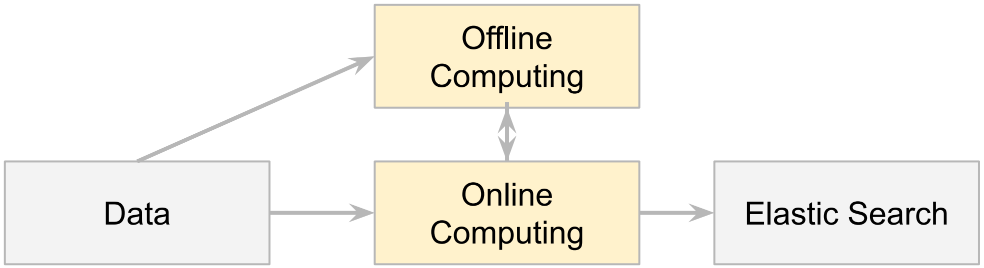
- Good clusters from topics lead to good results on content-based.
- With graph memory-based Collaborative Filtering, item-item recommending brings good results as well.
- The hybrid has many advantages such as utilizing as much as models or strategies, combining analysis results like prioritizing low product_cnt items
- The hybrid also has many problems and results seem to be difficult to explain by common sense. It need to be carefully designed and tuned.

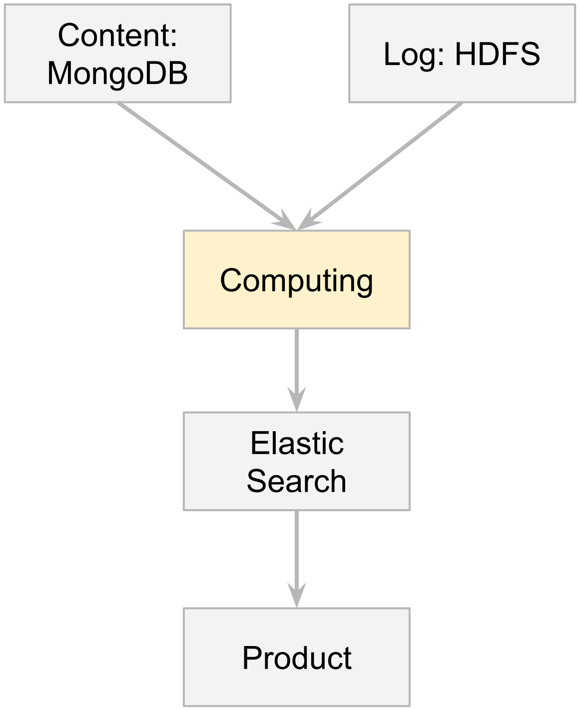
=> The hybrid also unfolds many further works.

# 6. System design

- Content: stable.
- Log: large size, fast update
- Output of computing: basic schema, need fast query

- Offline computing: computing daily, weekly, monthly, includes feature engineering, training, evaluating…
- Online computing: fast computing, even realtime or semi-realtime, catch update of log and content.



# Summary

- Given the task and target, this presentation shows an overview of data, how to exploit it and techniques and models to process the data. The presentation also briefly overview of recommender system methods in research.
- The beginning assumptions are simple, so these works the necessary starting step to further development.
- Limitation: fully evaluation is missing and system design is still basic.