Hepsiburada Recommendation Team Data Scientist Assignment

Design Document

By

Asem Okby

Table of contents

1. Understanding The Problem

The purpose/goal of this project is recommending products related to the products in the customers' carts. To solve this problem, many methods may be used, and the problem is better approached by breaking it into smaller problems.

The similarity between two products can be inferred from the features of the two products. In the data given, those features are text. Therefore, a subproblem to be solved is representing the text in a numerical way to be able to perform calculations on them. There are many methods for converting text into vector representations, therefore, some of them should be tried out and compared. Once the text features are represented by vectors, another problem to tackle is calculating the similarity between the vectors of each product.

Finally, another major subproblem is ranking the results of a recommendation system. Although ranking could be simply based on the similarity measure, there are other ways that could yield better results. For example, this could be done based on a feature such as price or any other feature. In the following sections, those subproblems are discussed in details.

2. Data Discovery

1. Initial thoughts and assumptions

- The features of the products are text, hence, a NLP technique may be of use in this problem.
- There seems to be a **category** which is written in English, Pet Shop.
- The data might need to merged at some point on the **productid** attribute if needed.
- The **brand** name seems to be stated again in the **name** column.

2. Data Types

All the columns are of type **object**. The column **price** may be converted to float. The column **price** may be used in many ways. For instance, since customers

like whole numbers, the prices of the items that are recommended to the user may have a great factor in attracting the user if the they complement the price in the cart making it a whole number.

3. Missing Values

There are missing values to be handled in some way. The values may be dropped or imputed. A decision should be made here and what may help in making such a decision is things suchs as the number of the null values and the importance of the feature.

4. Value Counts

By looking at the value counts of some columns, the following is noted:

- Some users put similar products to their carts, hence, a user-based recommendation technique may be considered.
- The categories and subcategories sets' sizes is small, hence, a content-based recommendation technique may be effective here.

3. Data Processing

1. Handling Missing Values

Missing values could be handled in many ways such as deleting the rows with the missing values or imputing techniques. Since the number of rows with missing values in the **events** dataset is small, deleting those rows would be an okay way to handle the missing values. For **meta** dataset, the column **brand** that has the most missing values, won't be used here, since the **brand** is already stated again in the **name** column. Therefore, no need to handle those missing values. However, there is a single row in the **meta** dataset where all the values are missing. This row is simply removed.

2. Processing Text Features

As mentioned earlier, the features of the products are text. Before carrying on and using a NLP technique the text is processed.

In the columns **category** and **subcategory** there seem to be a pattern in the text, either a **comma**, a **ve** or an **empty space** separate any two words. Therefore, to separate the words of each product those separators are taken into consideration.

For column **name**, some further things should be taken care of:

- There are **punctuation** marks other than **commas**, so they all are removed.
- There are **stopwords** other than **ve**. All are removed using a Turkish stop-words list.
- There are **numbers** in the text that are removed, although there could be used in some way.
- There are **units** in the text. A list of units could be used to remove them all, but for this project an assumption made is that units found in the dataset are 2 chars or less. Therefore, any word with length which is less than or equal to 2 is removed.

3. Converting Data Types

Here, only the type of the column **price** is converted from **object** to **float**.

4. Approach 1: Content-Based Filtering

In this section, the methods will be content-based methods, hence, the focus in this section will be on the **meta** dataset.

1. Method 1: CountVectorizer

1. Description

As discussed earlier, a problem that should be solved is representing the text features as vectors. In this method, **CountVectorize** is used to achieve this. Before this is applied, first, all the text of all text features (category, subcategory, and name) is combined and put under one single feature. Then, each set of words is converted to a vector representation on the basis of the frequency of each word in the entire text. The vectors are used then to calculate the similarities with the **cosine similarity** technique and a similarity matrix is created. Finally, using the **similarity matrix** the top n products similar to a products may inferred.

2. Example

Input product

	productid	brand	category	subcategory	name
10225	ZYBICN9286868	Lipton	İçecekler	Gazsız İçecekler	LİPTON İİCE TEA ŞEFTALI AROMALI TNK 500 ML

Recommended products

	productid	brand	category	subcategory	name	similarity
0	ZYHPPEPSIGZS019	LIPTON ICE TEA	İçecekler	Gazsız İçecekler	Lipton Ice Tea Şeftali 1,5 L	0.889499
1	ZYHPCOCACGZS010	Fuse Tea	İçecekler	Gazsız İçecekler	Fuse Tea Şeftali Pet 1 Lt	0.859338
2	ZYBICN9287068	Lipton	İçecekler	Gazsız İçecekler	LİPTON İCE TEA DOUBLE ŞEFTALİ & KAYISI AROMALI...	0.787726
3	HBV00000NFHOB	Lipton	İçecekler	Gazsız İçecekler	Lipton Ice Tea Şeftali 4 x 250 ml	0.739600
4	ZYBICN9286873	Lipton	İçecekler	Gazsız İçecekler	Lipton İce Tea Şeftali-Kayısı 1.5 Lt	0.701646
5	HBV00000PQKHY	Lipton	İçecekler	Gazsız İçecekler	Lipton İce Tea Şeftali Aromalı 6*330 Ml	0.701646
6	ZYBICN9286869	Lipton	İçecekler	Gazsız İçecekler	LİPTON İCE TEA LİMON AROMALI TNK 500 ML	0.647150
7	ZYBICN9286870	Lipton	İçecekler	Gazsız İçecekler	LİPTON İCE TEA DOUBLE ÇİLEK & KAVUN AROMALI TN...	0.647150
8	ZYBICN9310832	Lipton	İçecekler	Gazsız İçecekler	Lipton İce Tea Şeftali 2 Lt	0.647150
9	HBV00000PQKHW	Lipton	İçecekler	Gazsız İçecekler	Lipton İce Tea Şeftali Aromalı 500 Ml	0.647150

3. Discussing The Results

It is fascinating that such a simple method can produce such results. The top 10 products all have the same value for the columns **category** and **subcategory** and the values of the column **name** are also very close to the value of the input product. Despite that, It cannot be said that the method works perfectly. This is merely a single example. Indeed, an interesting example is provided in the notebook showing that this method may not be perfect, yet it works just fine.

4. Pros

- Easy to implement and understand.
- Provides meaningful results.

5. Cons

- CounVectorizer is biased in favor of the most frequent words, hence, ignores the rare words.
- CounVectorizer cannot capture semantic and analogy relations between the words like methods such as Word2Vec.

2. Method 2: Doc2Vec

1. Description

Another method that could be used in converting text into vectors is **Doc2Vec**. Doc2Vec captures semantic relations between documents. To find the similarity between the combined sentences (category, subcategory and name), Doc2Vec is first used to assign vectors to documents where each product's features combined represent a document. After assigning each

product a vector, **cosine similarity** is used to find the similarity between documents/products.

2. Example

1. Input product

	productid	brand	category	subcategory	name
2454	PTINQUIK-078	Quik	[pet, shop]	[kuş]	[quik, kumlu, tünek]

- **Recommended products**

	productid	brand	category	subcategory	name	similarity
0	HBV00000ILFYH	Fit Fly	Pet Shop	Kuş	Fit Fly Kapalı Lüx Tül Small	0.936506
1	PTLIENJ13	Enjoy	Pet Shop	Kedi	Enjoy Cat Food Yetişkin Kedi Maması 1 Kg	0.928902
2	HBV00000NG8TV	Pet Craft	Pet Shop	Köpek	Pet Craft Reflektörlü Göğüs Tasma XS	0.928716
3	HBV00000NVZ6O	Quik	Pet Shop	Kuş	Quik Kuş Tüneği 4'lü	0.925733
4	HBV0000056GDH	Pro Line	Pet Shop	Kedi	Proline Marsilya Sabunlu Topaklaşan Kedi Kumu ...	0.922589
5	PTSRENJ-25	Enjoy	Pet Shop	Kuş	Enjoy Parpağan Yemi 700 Gr	0.921027
6	HBV000009GWZK	Felix	Pet Shop	Kedi	Felix Yetişkin Kuzu Etli Kedi Pouch 100 Gr	0.918381
7	PTINJNG-005	Jungle	Pet Shop	Kuş	Jungle Kanarya Yemi 400 Gr	0.916194
8	HBV00000QX1W5	Felix	Pet Shop	Kedi	Felix Balıklı 4 x 100 gr	0.915048
9	HBV00000PQSBR	Jungle	Pet Shop	Kuş	Jungle Paraket Yemi 500 g	0.914106

2. Input product

	productid	brand	category	subcategory	name
752	HBV00000NG8KC	Universal	[spor, outdoor, oto]	[spor, topları]	[universal, voleybol]

- **Recommended products**

	productid	brand	category	subcategory	name	similarity
0	SPORUS22342	Voit	Spor, Outdoor ve Oto	Kondisyon Aletleri	Voit Yoga Block	0.949698
1	HBV00000PGMYX	Aytaç	Et, Balık, Şarküteri	Şarküteri	Aytaç Dana Sosis 5'li 190 g	0.947816
2	HBV00000US4BG	Wells	Spor, Outdoor ve Oto	Kamp Malzemeleri	Wells Katlanır Kamp Sandalyesi	0.935428
3	HBV00000QU3WQ	Damla	Su	Su	Damla Su 1.5 lt	0.929969
4	HBV000000LRY8	Molfix	Bebek	Bebek Bezi	Molfix Bebek Bezi 3 Beden Midi Süper Fırsat Pa...	0.927490
5	HBV00000PVR4Q	Voit	Spor, Outdoor ve Oto	Kondisyon Aletleri	Voit DB107 Dipping Dumbell 5 kg Gri	0.926815
6	HBV00000PQIOP	Baby Turco	Bebek	Bebek Bezi	Baby Turco Bebek Bezi Yenidoğan 60 Adet	0.922928
7	HBV00000QU3YT	Erikli	Su	Su	Erikli Su 1 lt	0.922686
8	HBV00000PQIXT	Bingo	Ev Bakım ve Temizlik	Çamaşır Yıkama	Bingo Yumuşatıcı Konsantre Lovely 1400 Ml	0.922556
9	HBV00000QU3ZX	HAYAT SU	Su	Su	Hayat Su 0,5 Lt	0.921938

3. Discussing The Results

In the first example, the recommended products are well related to the input product. However, the second example is very interesting. The second recommended product, for example, does not relate with the input product directly at all. Further analysis may be conducted here to understand why such results are obtained. Also, it is good to keep in mind that Doc2Vec does not work very well on small datasets. The published paper itself uses tens-of-thouthands to millions of text. Therefore, certainly the results would become more meaningful as the dataset gets bigger. Finally, the hyperparameters of the model may be tuned and the results are compared to get more reliable results.

4. Pros

- Doc2Vec captures semantic relations between document's features.
- Doc2Vec produces reliable results when given enough data.

5. Cons

- Doc2Vec requires huge amount of text to yield good results.
- Doc2Vec is more complex than CountVectorizer.

5. Future Work

Due to the time constraint many methods and ideas were not considered or implemented here. For example the following could have been tested for converting the text to vectors and finding similar products:

- TF-IDF
- Word2Vec
- Word Mover's Distance

Also, many user-based methods could have been tested using the events dataset. Some methods that could have been tried out are:

- Clustering
- Matrix Factorization
- Deep Learning

Besides, some time could have been dedicated to experimenting many ways of ranking the results and even combining more than one method together to get a better recommendation system.