JIGSAW ACADEMY
THE ONLINE SCHOOL OF ANALYTICS

Case Study: Finding Similar Products

Business Case: A retail company has been struggling in the recent past, trying to optimise product placement in one of its largest stores in the city. On a close inspection, the management found out that, part of the problem is the fact, that similar kind of products are placed in shelves far apart from each other. Manually figuring out which products are similar, will be a very time-consuming task. The store manager wants a solution in which manual intervention is minimal. The store manager checked with his IT team and the IT team could extract data on the current product placement in the store. His IT team can fetch data on:

(i) A text description of each product
(ii) Current Aisle Id of each product
(iii) Current Department Id of each product

The IT team could extract a sample of data. This sample has been made available in the file named **"prods.csv".** Can you come up with a solution that is data driven? An ideal solution will help in finding out for a given product what are top 5 most similar ones and which Aisle and Department they belong to.

Hints: You can try to find out cosine similarity of each product pair. To compute cosine similarity, you will need to represent text in a vector form. Use tfidf, to represent description of products in the vector form. Once you have vector representation of each product, you will be able to compute cosine similarity.

Before you create tfidf representation of product descriptions, you will need to clean these descriptions by getting rid of stop words, unnecessary special characters.

One way to structure your final output is the following:

product_id	product_nam	aisle_id	department_	suggested_products	suggested_pro_ids	suggested_dept_ids
1	Chocolate Sa	61	19	Oreo Cookies and Cream Chocolate Frozen Dairy Dessert,Vanilla Sugar C	5,91,57,61,72,559	1,19,19,19
2	All-Seasons S	104	13	Thin Stackers Brown Rice Salt Free,Sardines in Water Salt Added,Salt Fre	2,73,42,54,63,240	19,6,13,13
3	Robust Golde	94	7	Almond Breeze Unsweetened Almond Coconut Milk Blend,Unsweetened C	7,38,86,25,60,569	16,16,7,19
4	Smart Ones (	38	1	Classic coke,Ice Cream, Cookies & Cream,Mini Double Chocolate Ice Crea	6,49,43,19,97,774	7,1,1,9
5	Green Chile /	5	13	Petite Green Peas,Caramel Sauce,Apple Green Cups,Chile Con Queso Pota	9,69,27,58,46,851	1,19,17,19
6	Dry Nose Oil	11	11	Dry Pasta Lasagne,Dry Roasted Pistachios,Super Dry Beer,Argan Oil of Mo	9,13,30,43,79,214	9,19,5,11
7	Pure Coconu	98	7	Pompelmo Water,Coconut Drink, Vanilla,Organic Apple Spinach Kale Coc	4,91,67,02,99,920	21,16,7,16
8	Cut Russet Pc	116	1	Instant Mashed Potatoes,Gold Potatoes,Zita Cut, No. 118,Fresh Cut Specia	2,78,30,61,39,799	9,4,9,15
9	Light Strawb	120	16	Creamline Yogurt Wild Blueberry,Organic Blueberry Pomegranate Whole	3,39,58,44,19,132	16,16,16,16
10	Sparkling Ora	115	7	Sparkling Blush Grape Juice,Organic Orange Turmeric Juice,Sparkling Ras	1,67,48,83,34,618	7,7,7,7
11	Peach Mango	31	7	Soft Eating Mango Flavor Liquorice,All Natural 100% Apple Juice,Peach Al	70,98,26,77,552	19,7,7,1
12	Chocolate Fu	119	1	No Sugar Added The Original Fudge Pops,6\" Organic Carrot Cake,Flourle	7,58,21,88,24,362	1,3,3,13
13	Saline Nasal	11	11	Cheese Creations Four Cheese Cheese Sauce,Complete Health Deboned C	34,13,42,32,91,000	9,8,16,10
14	Fresh Scent D	74	17	Automatic Toilet Bowl Cleaner Bleach & Blue Rain Clean Scent Pack,with	8,65,54,17,32,186	17,17,17,17
15	Overnight Di	56	18	Snack Size Candy,Swaddlers Diapers Jumbo Pack Size Newborn,Cruisers D	6,30,76,56,82,879	19,18,18,18
16	Mint Chocola	103	19	Mint Dental Floss,Orange Flavored Sports Drink,Slow Churned Mint Choco	2,91,81,54,80,205	11,7,1,19
17	Rendered Du	35	12	Traditional No Fat Refried Beans,85% Lean, 15% Fat Ground Turkey (0130	7,03,78,26,21,953	15,12,15,16
18	Pizza for One	79	1	Brooklyn Pizza Dough,Pepperoni Party Pizza,Ristorante Pizza Spinaci Thin	3,22,74,33,56,735	1,1,1,1

You can use this article as a reference point to approach this problem:
https://towardsdatascience.com/how-i-used-text-mining-to-decide-which-ted-talk-to-watch-dfe32e82bffd

Deliverables:

1. The python code that you used to create the analysis (Make sure, you comment the code well)
2. The final data file with additional columns on: Recommended Products, Aisle Id and Department Id of respective products recommended.