**Capstone Project 1**

**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:

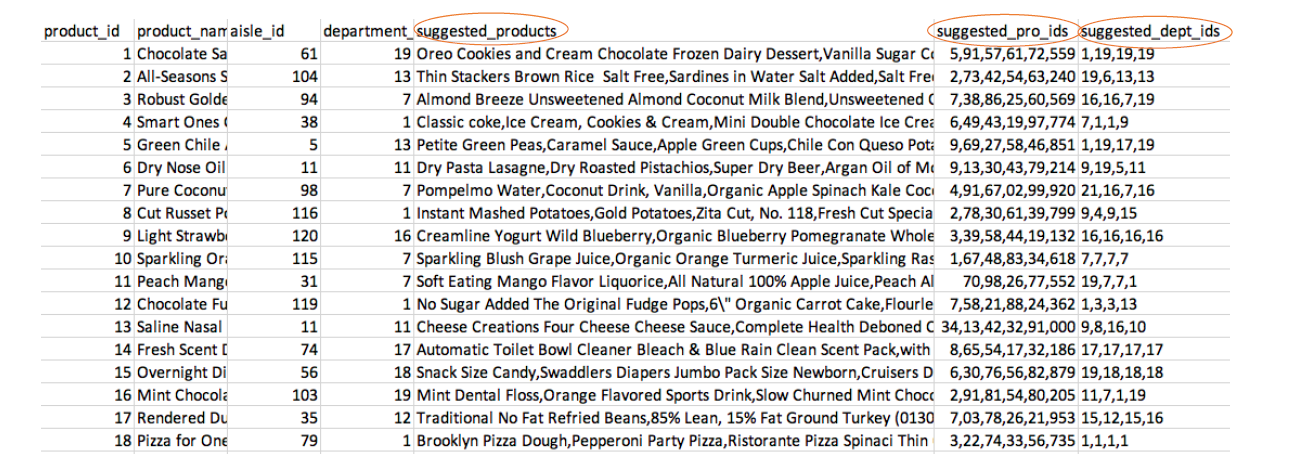
1. (i) A text description of each product
2. (ii) Current Aisle Id of each product
3. (iii) Current Department Id of each product

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:



**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.