DS 5230 - Unsupervised Machine Learning

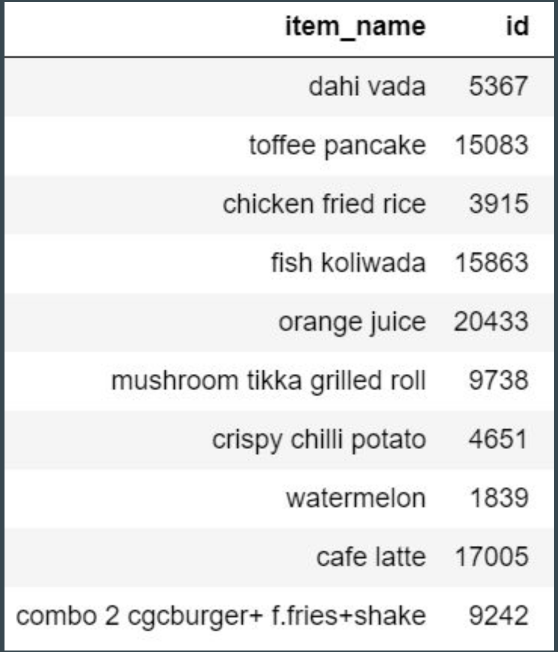
Classifying Food Items into Appropriate Categories

•••

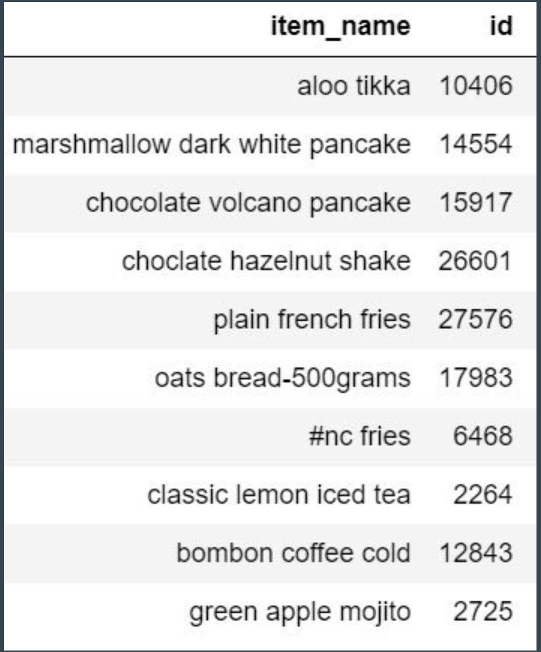
Pranshu Diwan (001347331) and Shagun Jindal (001596493)

About the dataset

The dataset contains more than 13,000 rows of food items from Indian restaurants.



item_name	id
dahi vada	5367
toffee pancake	15083
chicken fried rice	3915
fish koliwada	15863
orange juice	20433
mushroom tikka grilled roll	9738
crispy chilli potato	4651
watermelon	1839
cafe latte	17005
combo 2 cgcburger+ f.fries+shake	9242



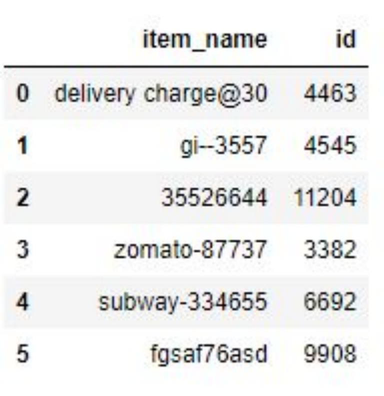
item_name	id
aloo tikka	10406
marshmallow dark white pancake	14554
chocolate volcano pancake	15917
choclate hazelnut shake	26601
plain french fries	27576
oats bread-500grams	17983
#nc fries	6468
classic lemon iced tea	2264
bombon coffee cold	12843
green apple mojito	2725

Data Cleaning

The dataset contained several garbage values. We performed some basic text processing like:

- Unwanted Punctuations
- Special Symbols
- Converted to lowercase
- Food names starting with certain patterns

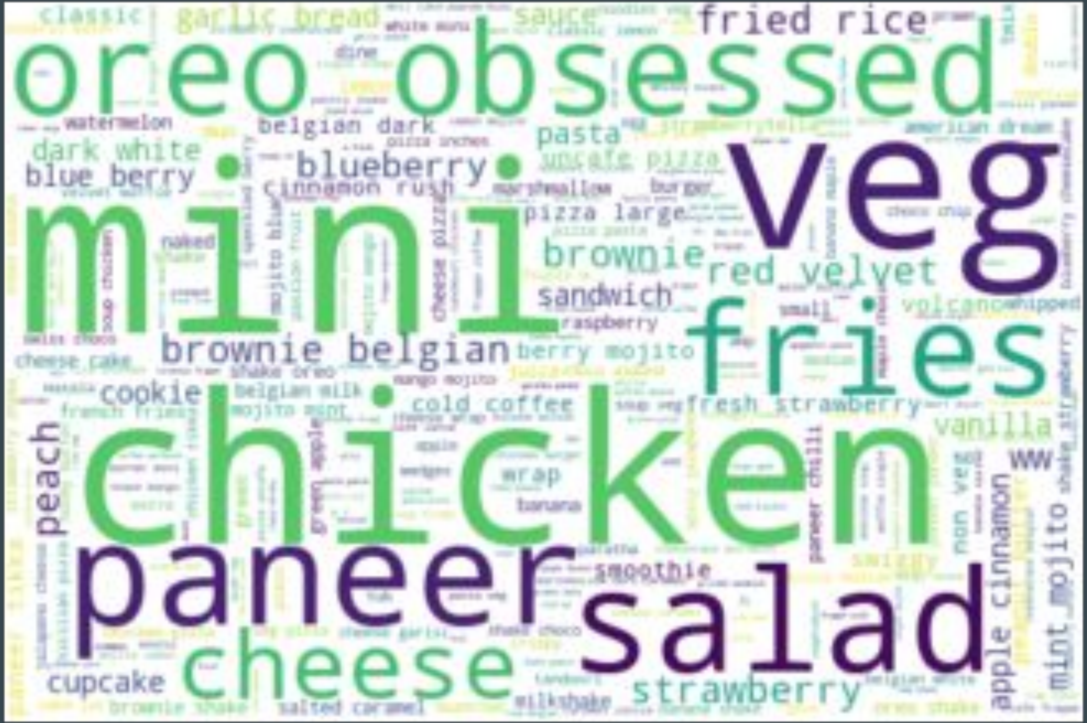
Some restaurants did not provide with the food names so garbage values were filled in for every food id



	item_name	id
0	delivery charge@30	4463
1	gi--3557	4545
2	35526644	11204
3	zomato-87737	3382
4	subway-334655	6692
5	fgsaf76asd	9908

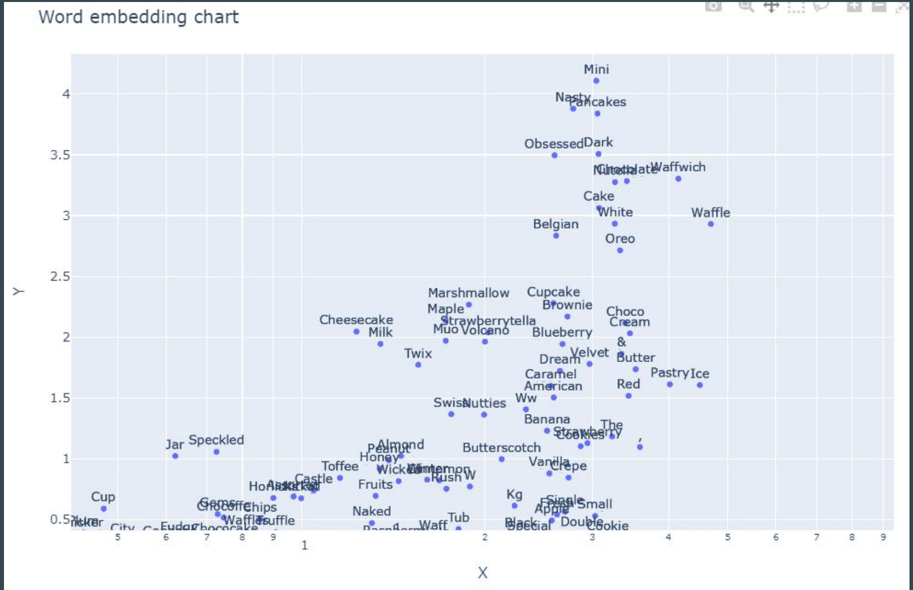
Exploratory Data Analysis - I

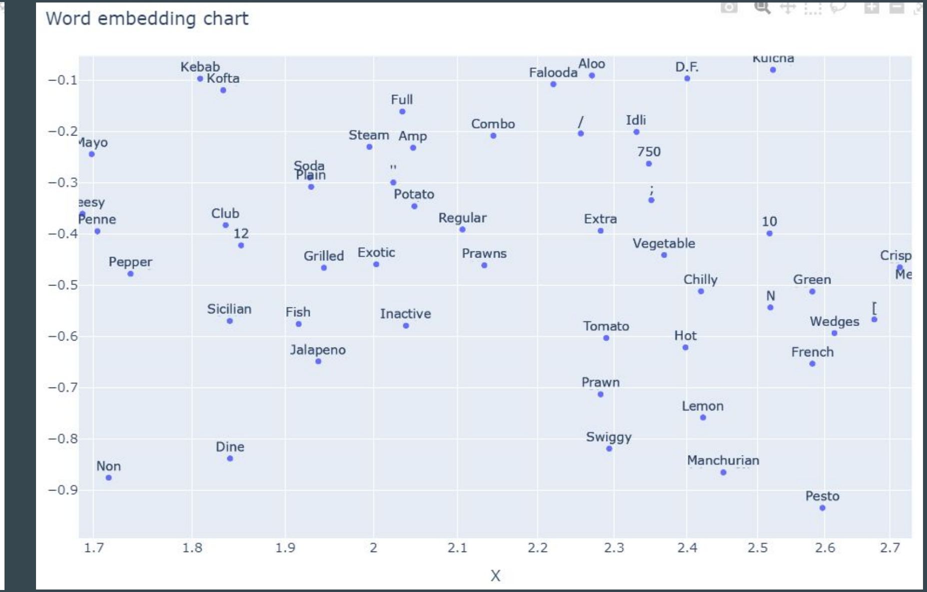
Visualizing the most prevalent food dishes through word clouds



Exploratory Data Analysis - II

Visualizing the word vectors in a 2-D space

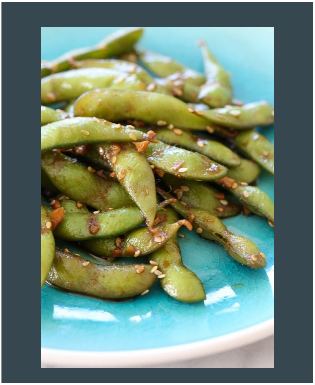




Formulating the problem statement using K-Means

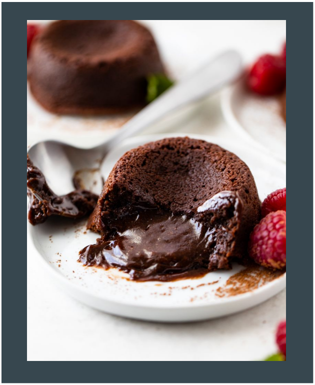
Trying to cluster the word vectors in 5 categories using K-Means algorithms

Vegetarian



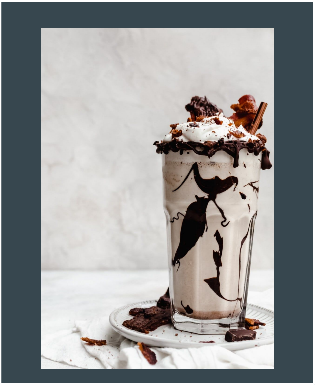
Non-Vegetarian

Desserts



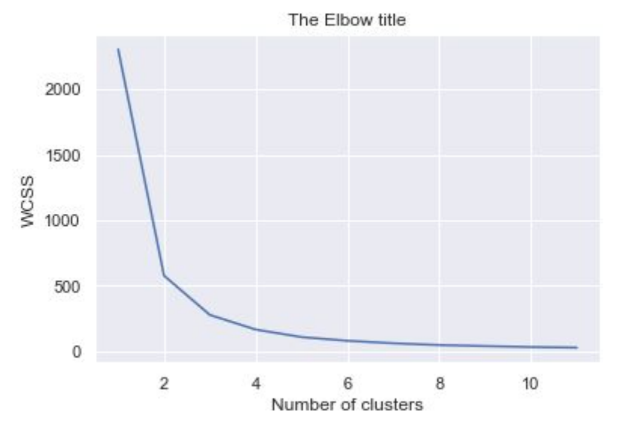
Alcoholic Beverages

Non-Alcoholic Beverages



K-Means Clustering

Optimizing the number of clusters using the elbow method

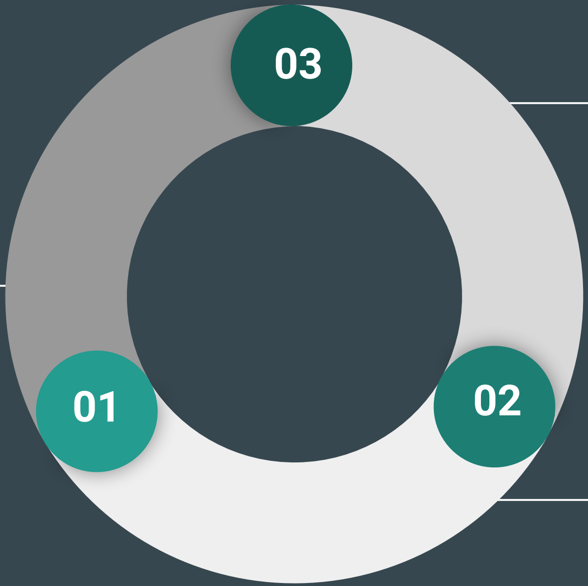


Elbow method suggests number of clusters = 3

Evaluating accuracy

- We manually created a test dataset of 100 food items with their correct categories.
- We then evaluated the K-Means algorithm on these 100 items.
- Accuracy: 56/100

Using Cosine Similarity



Deciding the top keywords for each category

We define our categories and the top keywords associated for that category. For example,

Topics = {‘veg’, ‘non-veg’, ‘desserts’}

Keywords = {‘paneer tofu’, ‘chicken mutton fish’, ‘choco milkshake’ }

Generating word Vectors using Word2Vec

We then generate a word vector for every category and a vector for every word in that category.

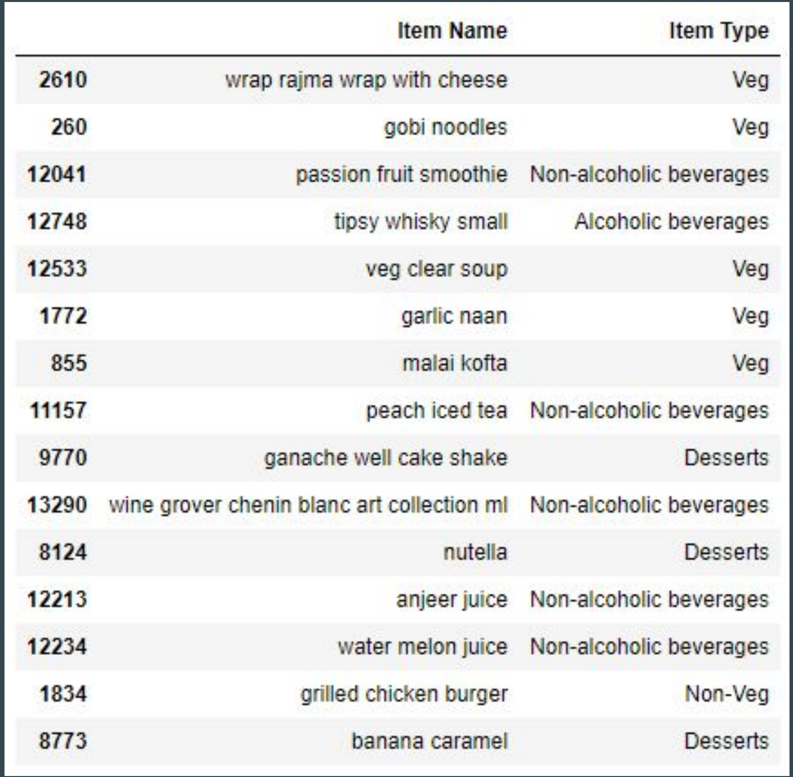
Calculating cosine similarity for each dish

We then calculate the cosine similarity for each item in the dataset with the keywords and pick the one with the highest cosine similarity

Results - Using Cosine Similarity

Evaluating accuracy

- We manually created a test dataset of 100 food items with their correct categories.
- We then evaluated the Cosine Similarity method on these 100 items.
- Accuracy: 84/100



	Item Name	Item Type
2610	wrap rajma wrap with cheese	Veg
260	gobi noodles	Veg
12041	passion fruit smoothie	Non-alcoholic beverages
12748	tipsy whisky small	Alcoholic beverages
12533	veg clear soup	Veg
1772	garlic naan	Veg
855	malai kofta	Veg
11157	peach iced tea	Non-alcoholic beverages
9770	ganache well cake shake	Desserts
13290	wine grover chenin blanc art collection ml	Non-alcoholic beverages
8124	nutella	Desserts
12213	anjeer juice	Non-alcoholic beverages
12234	water melon juice	Non-alcoholic beverages
1834	grilled chicken burger	Non-Veg
8773	banana caramel	Desserts

Conclusions and next steps

Use advanced clustering algorithms

Use different clustering algorithms like K-medoids, hierarchical clustering, etc.

+

Convert into a semi-supervised problem

Using the results obtained from Clustering developed a dataset for supervised ML classification

+

Perform Hyperparameter Tuning

Optimize hyperparameters leading to better clustering results

Conclusions

- Simple K-Means clustering alone does not yield good results.
- Changing the number of categories (clusters) will vary the results.
- Generating more complex word vectors might yield better results.