Summary of Tasks

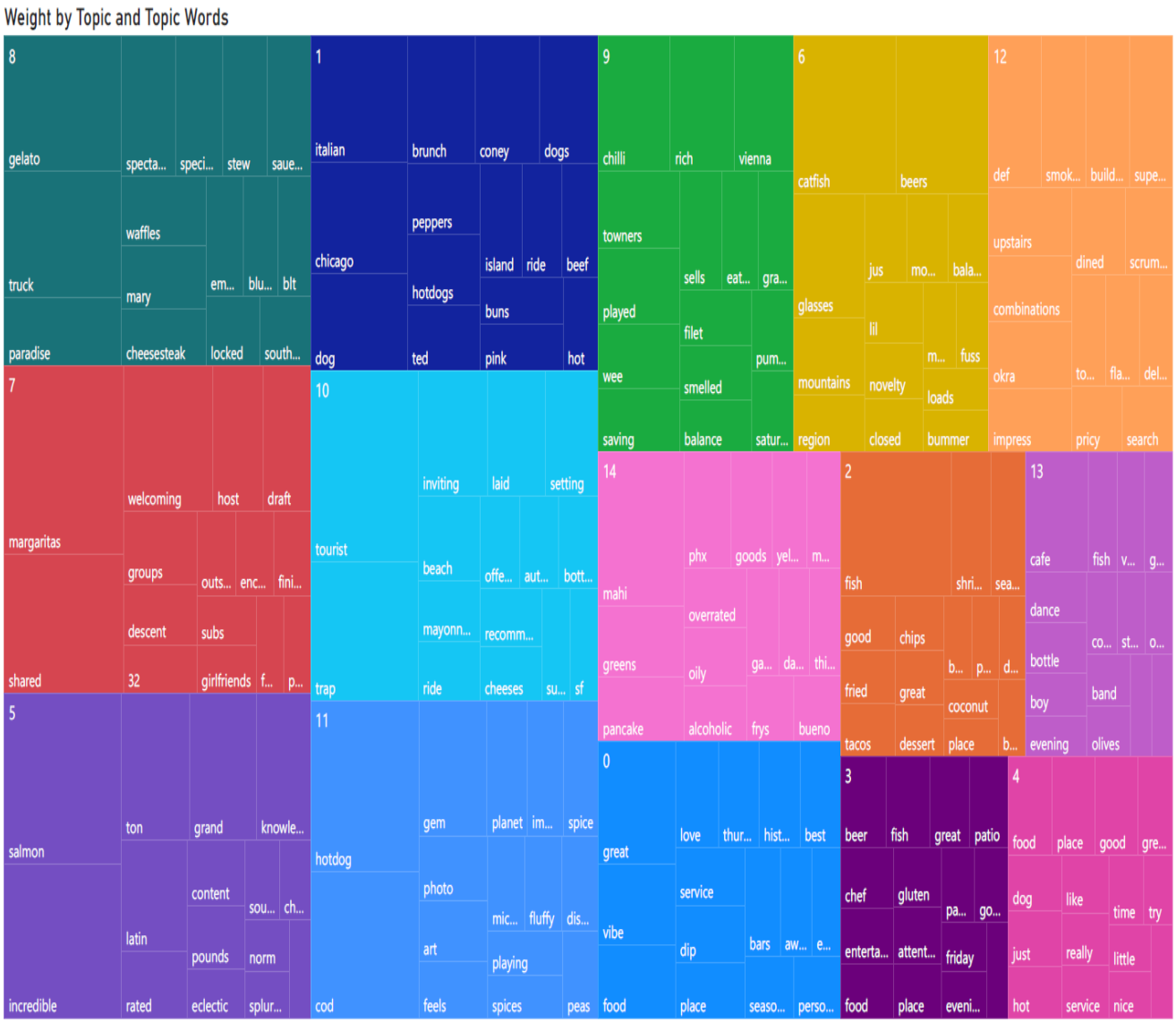
Task 1: Topic Modelling from Reviews

Description of Topic Modelling Process

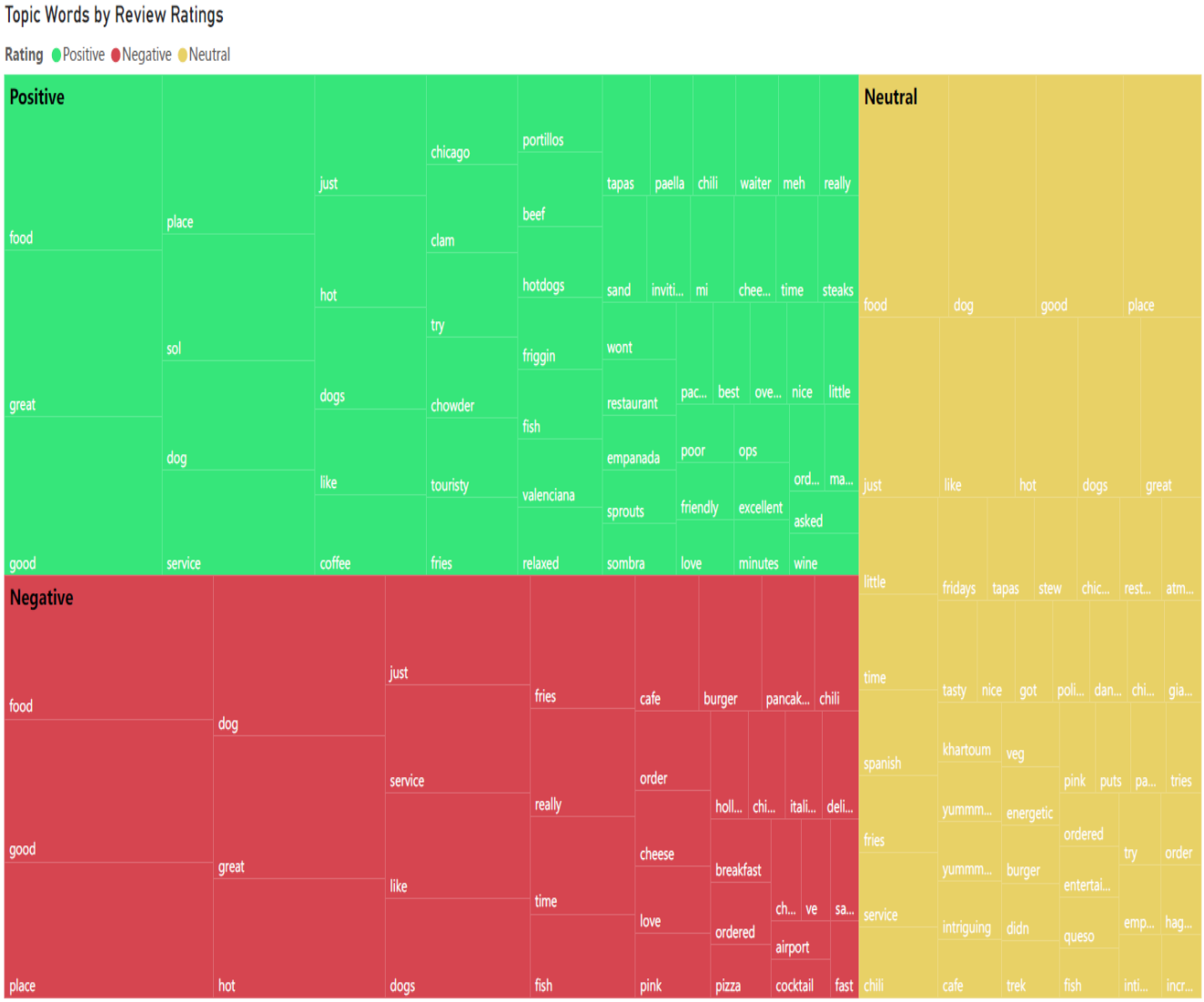
Initially, the Yelp dataset is loaded and pre-processed in Python by using Windows adapted version of py27_processYelpRestaurants.py. It loads the yelp review data and generates a random sample of 100,000 reviews from it, so topic modelling is quicker on a smaller dataset. After pre-processing, p27_ldaTopicModelling.py is used to apply LDA algorithm on the sample file. The parameters are set so 15 Topics and their weighted vocabulary is generated. The data produced from LDA is cleaned up in Excel to get a table containing topic numbers, words and their weights. Excel workbook is then fed to PowerBI to visualise the data.

For the second task, the sample of reviews are merged with their ratings in Excel. Then, the reviews are sorted into three groups : rating 1 and 2, rating 3, rating 4 and 5 and are named negative , neutral and positive. These groups are then fed to LDA algorithm to find topics and then visualized in Power BI just like in task 1.

Visualisations



The topics and their words for 100,000 reviews randomly selected from Yelp restaurant reviews dataset are displayed as a treemap. The bigger the weight of the term, the more area it occupies. Each color represents the 15 topics generated through LDA topic modelling.



In this visualisation, the reviews are categorised into positive, negative and neutral shown by green, pink and yellow colours. The topic words used in the three groups are shown by different sizes. The bigger the size of the rectangle, the bigger is the weight.

Findings

Visualisations show an interesting pattern in the reviews. There are common themes and topics that can be seen in all three groups of reviews for example, food, dog, place etc. and some words that are specific to one category of reviews only like time, great and ok. This is expected as the domain words that are specific to dining and restaurants appear frequently in all types of reviews and do not indicate the intention of the user. It is interesting to note that words like just, order really appear in negative reviews and friendly, love etc. appear as a part of positive reviews topic modelling. Similarly, word “ok” stands out in neutral reviews. This gives an indication of how humans use language vocabulary to express negative and positive feedback.

Task 2: Visualization of the Cuisine Map Description

In order to create a visualisation for a cuisine map, categories of cuisines are generated by running Python program p27-processYelpRestaurants.py with parameter “–cuisine”. The

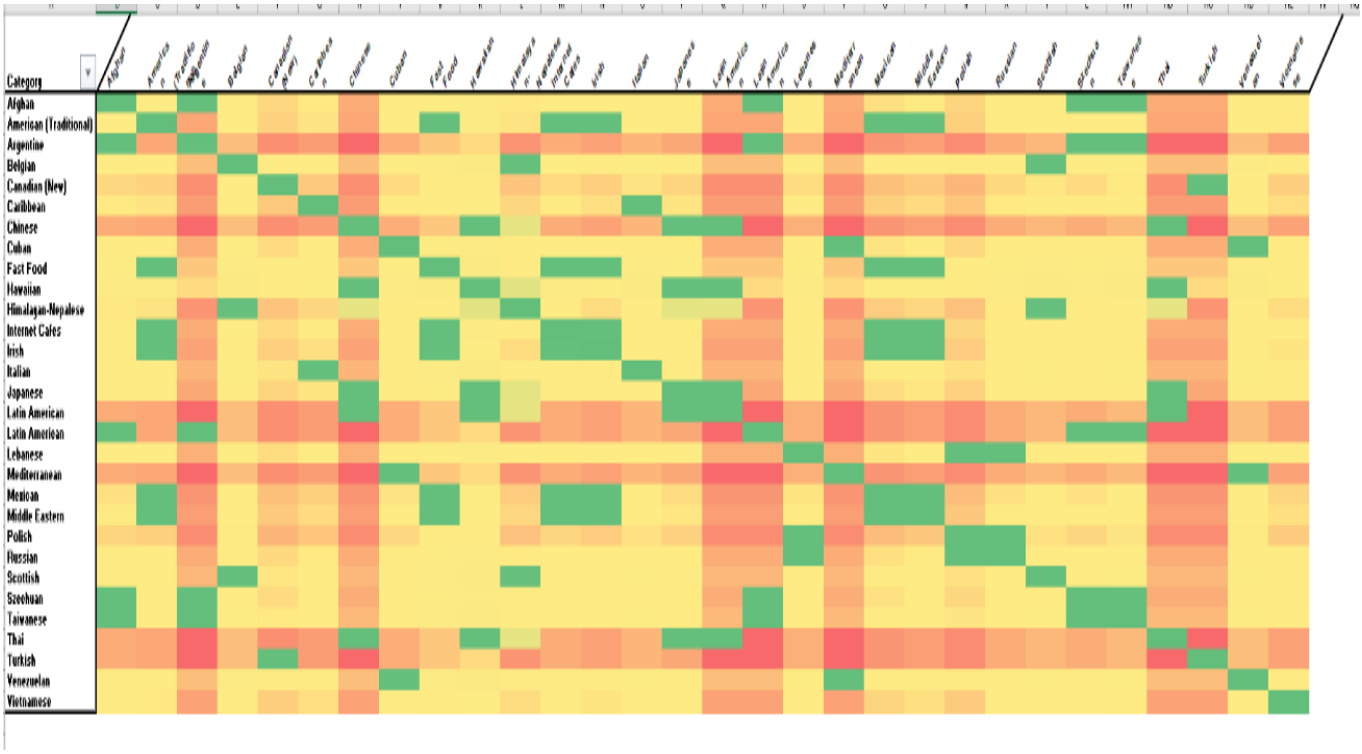
program extracts reviews from the restaurants that are tagged with a particular category and creates a collection of the reviews of such restaurants in a file named the same as the cuisine category in the category folder. These categories are then narrowed down to cuisines that belong to different countries/cultures. Then, Python program p27-processYelpRestaurants.py with parameter value "-matrix" is used to generate a random sample of 30 cuisines. The reviews for these cuisines are extracted as documents and converted to a corpus using a vectorizer. LDA model is then used to generate topic distribution for these cuisine reviews. 10 topics are assessed for each category. Then, cosine similarity measure is used to measure the similarity of the topic distribution in cuisines. The similarity matrix is used to generate visualisation of cuisine map.

Visualization of the Cuisine Map and Improvement

The visualisation is based on a colour gradient that goes from red (0 or close values) to yellow (mid-range values) to green (1 or close values). It is heatmap representation of similarity matrix for 30 cuisines generated in Excel.

In the first visualisation a count vectorizer is used to generate corpus for LDA. Most of the values are 0 or close as indicated by pink color.

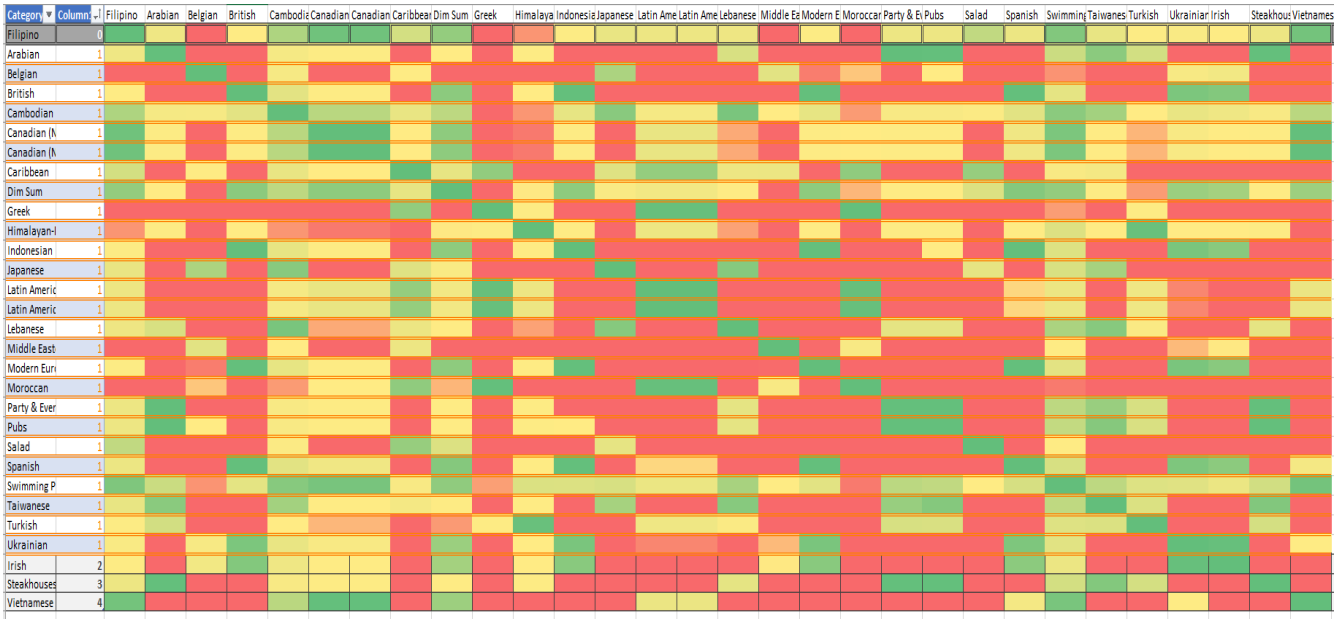
In the second visualisation TDFIDF vectorizer is used to generate the bag of words for LDA. The similarity values seem to be varied in this visualisation and hence give a better understanding of similarity between cuisines.



Task 2.3: Incorporating Clustering in Cuisine Map

In order to perform clustering, kmeans and agglomerative clustering algorithms are used by importing the sklearn python library. The first visualisation shows the results of kmean making 5 clusters without IDF vectorization. The results are very poor with most of the cuisines clumping up in one cluster only. This is because similarity matrix had mostly similar values.

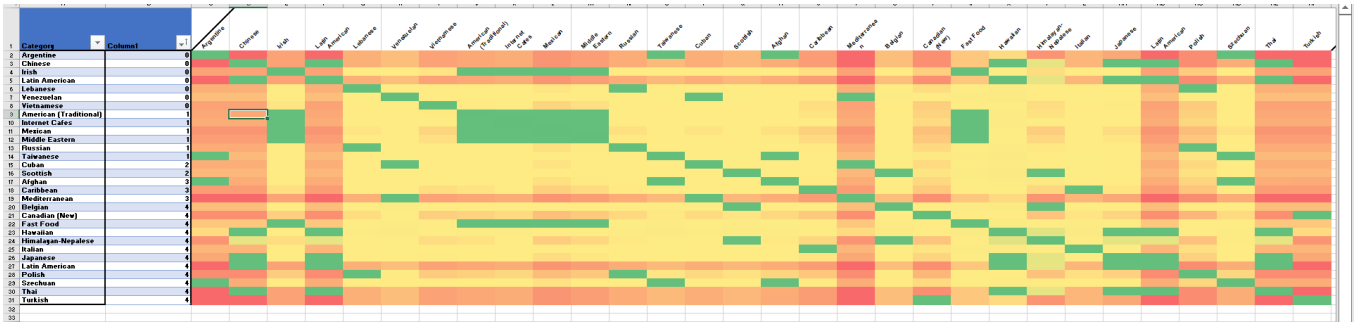
KMean clustering on Similarity Matrix without IDF



Resulting clusters:

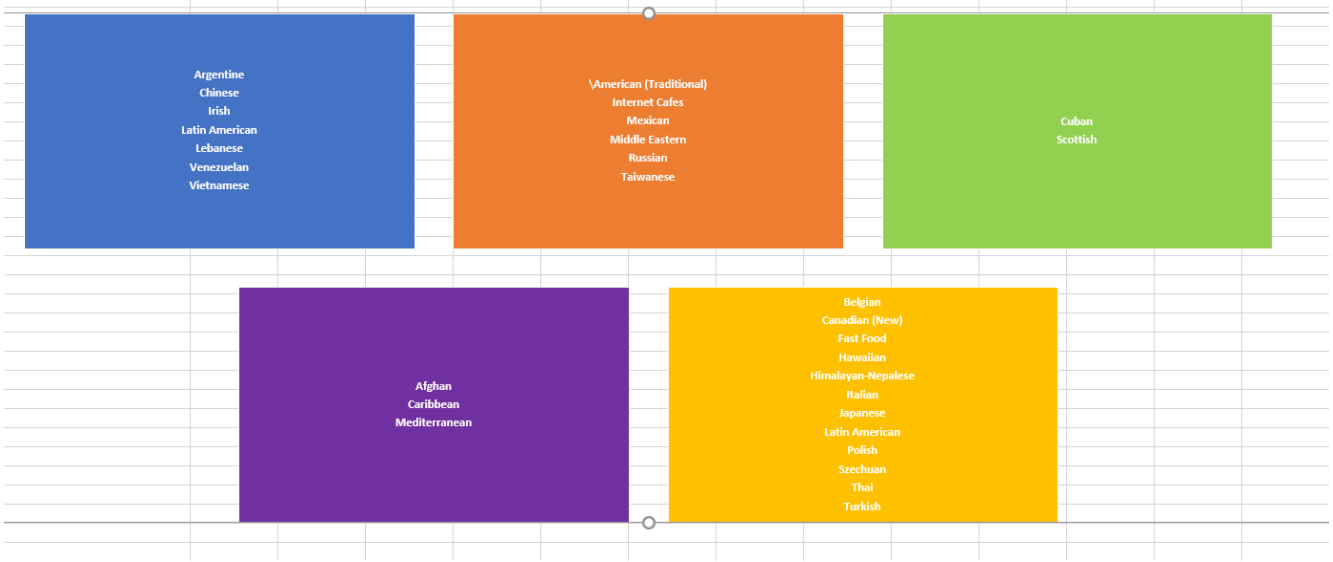
To improve the results TFIDF is incorporated.

KMean clustering on Similarity Matrix with IDF



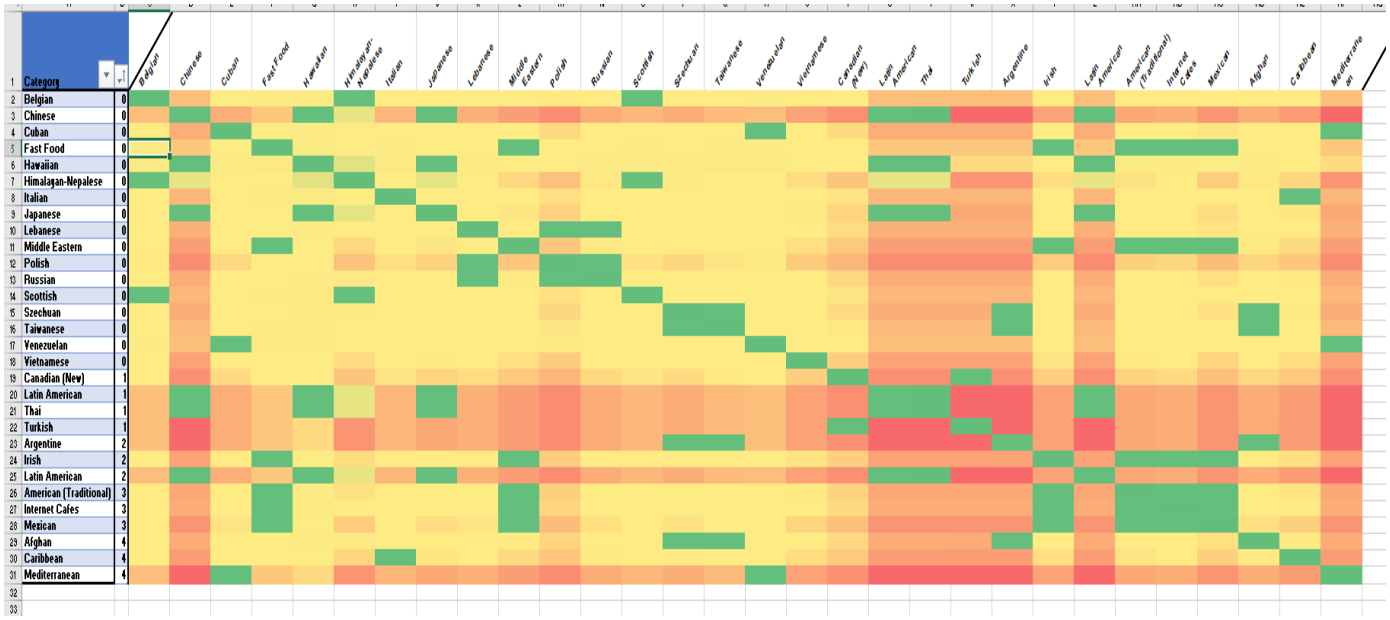
Category	Column1
Argentine	0
Chinese	0
Irish	0
Latin American	0
Lebanese	0
Venezuelan	0
Vietnamese	0
American (Traditional)	1
Internet Cafes	1
Mexican	1
Middle Eastern	1
Russian	1
Taiwanese	1
Cuban	2
Scottish	2
Afghan	3
Caribbean	3
Mediterranean	3
Belgian	4
Canadian (New)	4
Fast Food	4
Hawaiian	4
Himalayan-Nepalese	4
Italian	4
Japanese	4
Latin American	4
Polish	4
Szechuan	4
Thai	4
Turkish	4

Results improved and the following clusters are made.



In order to improve results further another algorithm agglomerative clustering is used. This significantly improved the quality of clusters as shown below.

Agglomerative clustering on Similarity Matrix with IDF

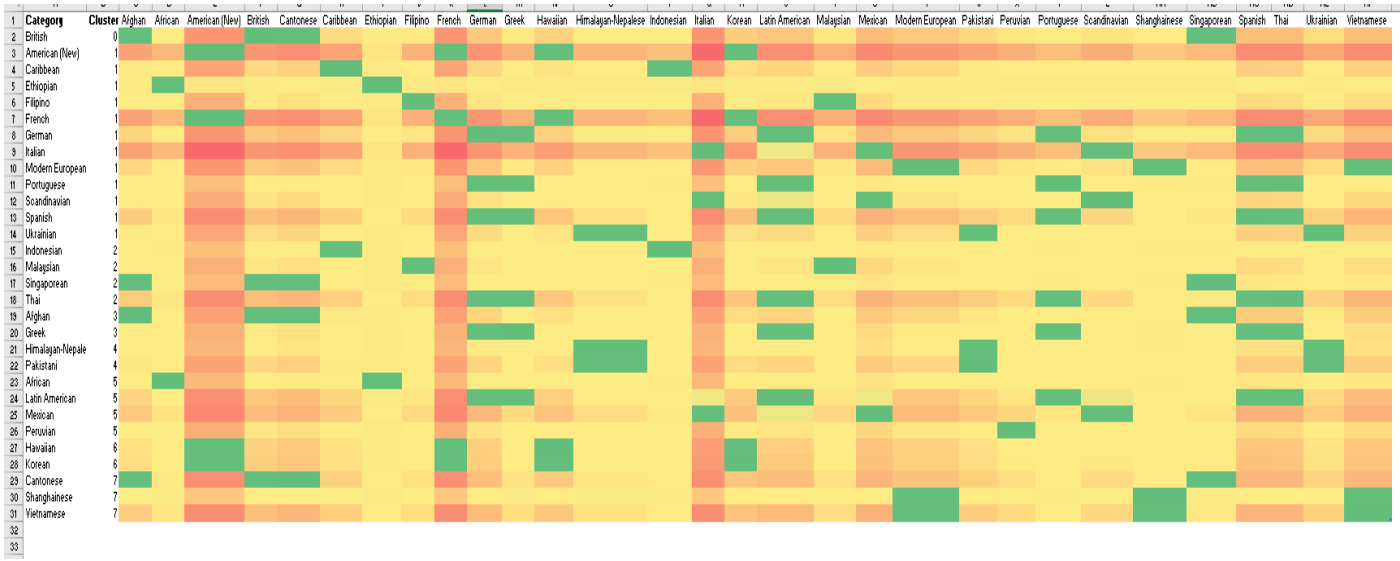


Category	
Belgian	0
Chinese	0
Cuban	0
Fast Food	0
Hawaiian	0
Himalayan-Nepalese	0
Italian	0
Japanese	0
Lebanese	0
Middle Eastern	0
Polish	0
Russian	0
Scottish	0
Szechuan	0
Taiwanese	0
Venezuelan	0
Vietnamese	0
Canadian (New)	1
Latin American	1
Thai	1
Turkish	1
Argentine	2
Irish	2
Latin American	2
American (Traditional)	3
Internet Cafes	3
Mexican	3
Afghan	4
Caribbean	4
Mediterranean	4

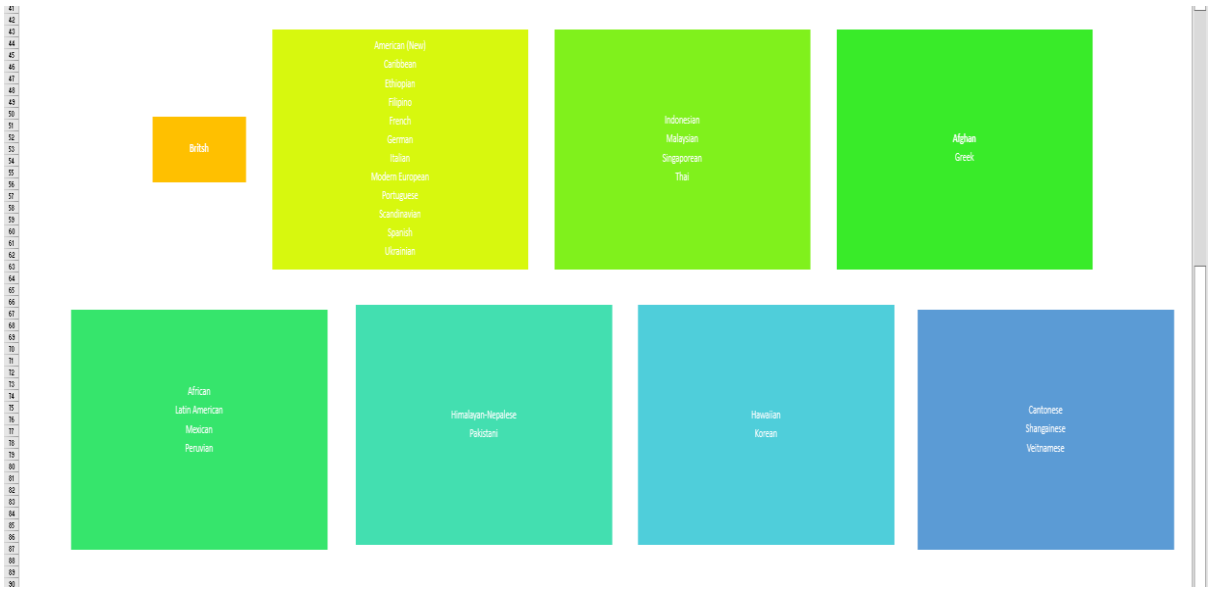
The clustering results:

Another approach is used and number of clusters is changed from 5 to 8. It definitely improves Kmean clustering results.

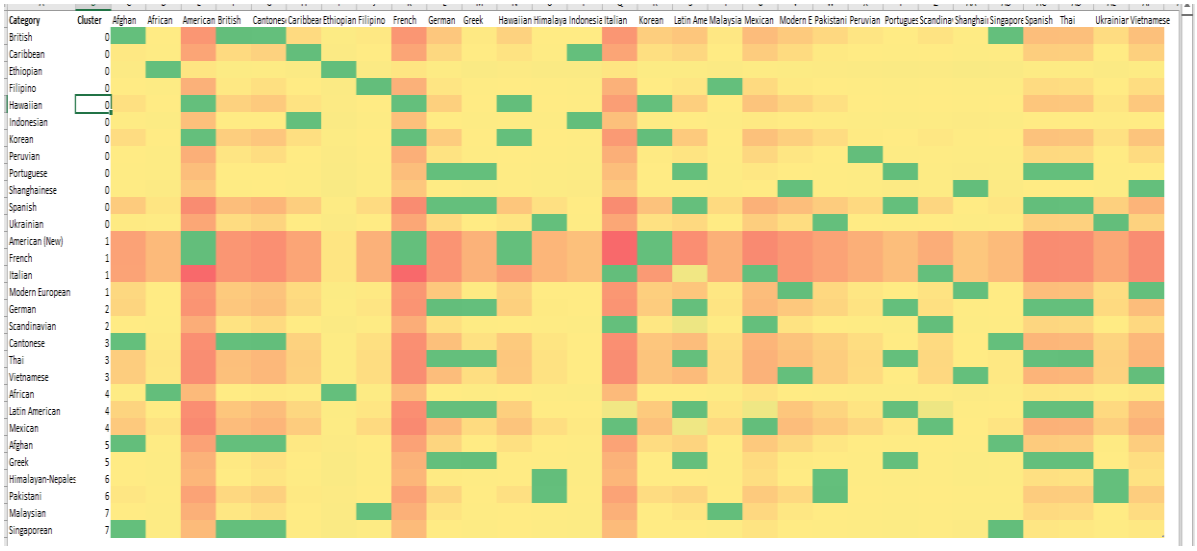
KMEANS clustering with 8 clusters



Category	Cluster	Afghan	African	American (New)	British	Cantonese	Caribbean	Ethiopian	Filipino	French	German	Greek	Hawaiian	Himalayan-Nepalese	Indonesian	Italian	Korean	Latin American	Malaysian	Mexican	Modern European	Pakistani	Peruvian	Portuguese	Scandinavian	Shanghainese	Singaporean	Spanish	Thai	Ukrainian	Vietnamese
British	0																														
American (New)	1																														
Caribbean	1																														
Ethiopian	1																														
Filipino	1																														
French	1																														
German	1																														
Italian	1																														
Modern European	1																														
Portuguese	1																														
Scandinavian	1																														
Spanish	1																														
Ukrainian	1																														
Indonesian	2																														
Malaysian	2																														
Singaporean	2																														
Thai	2																														
Afghan	3																														
Greek	3																														
Himalayan-Nepale	4																														
Pakistani	4																														
African	5																														
Latin American	5																														
Mexican	5																														
Peruvian	5																														
Hawaiian	6																														
Korean	6																														
Cantonese	7																														
Shanghainese	7																														
Vietnamese	7																														



Agglomerative clustering results with 8 clusters also worked well too. The following are the results:



The K-mean clustering with 8 clusters seem to be more accurate. Cuisines from geographically close areas have ended up in same clusters for example the cluster with Pakistani and Himalayan cuisines and cluster with Malaysian , Singaporean, Thai cuisines. This shows that reviews from tagged restaurants have been a useful representation for developing a cuisine map.

Task 3: Mining Dish Names

Manual Tagging

The following list is the result of manual tagging of “Indian” cuisines. False positives are removed from the list and the tag of false negatives are changed from 1 to 0.

chick peas	1	the lunch	0	not the best i've had	0
chicken tikka	1	all excellent	0	have never	0
flat bread	1	over priced	0	with lots of	0
tandoori chicken	1	i really like this place	0	the selection of	0

rogan josh 1	i'm in 0	well presented 0
gulab jamun 1	don't remember the 0	with our meal 0
basmati rice 1	thus far 0	the naan was 0
rice pudding 1	food is one 0	i will be returning 0
hot sauce 1	writing this 0	i've had 0
iced tea 1	last time i 0	would be a 0
fried rice 1	and for 0	no idea 0
ice cream 1	to cook 0	friendly staff and 0
tomato soup 1	was very tasty and 0	half an hour 0
chicken tikka masala 1	in my opinion 0	tikka masala 1
chinese food 1	decided to check 0	place stars 0
white rice 1	was great 0	a bad experience 0
tomato sauce 1	indian restaurant and 0	we had dinner 0
brown rice 1	there was a 0	food i've ever had 0
chicken wings 1	of the best indian buffets 0	coconut chicken 1
that's not 0	to arrive and 0	was out of 0
if there was 0	are way 0	at the end of 0
a traditional 0	too small 0	recommend this place to anyone 0
hands down the best 0	cuisine of india 0	water glasses 0
the reviews on yelp 0	is bland 0	ala carte 0
extra spicy 0	the indian sampler 0	sit down 0
happy and 0	not so good 0	the set 0
first time at 0	just moved 0	i suppose 0
amazing service 0	any case 0	did not disappoint 0
so after 0	to my 0	for dinner and 0
entr e 0	of white 0	the many 0
to pay the 0	the man 0	was flavorful and 0
pleased with 0	restaurant for the 0	appetizer we 0
we come 0	dipping sauce 0	
could have been better 0	on the menu 0	
guru palace 0	visit this place 0	
gets the 0	great service 0	
the non veg 0	it like 0	
have eaten 0	in las 0	
biz photos 0	in for the lunch 0	
a plus 0	great with 0	
a table 0		

Mining Additional Dish Names

In order to mine missing dish names, the reviews from yelp dataset were filtered to separate out the reviews for all the restaurants that are tagged “Indian”. The reviews are used to build a word2vec model in python. It is a procedure that learns the word association from the text and is used to guess similarity between words based on their semantic and pragmatic use in sentences. It is noticed that most of the dishes in the form of two-word phrases. The text of reviews is pre-processed into words. After this “Phrases” model from genism is used to generate all phrases from the text so that vocabulary for the model includes compound words and phrases of longer length. These phrases are used to train the word2vec model. The trained model is then programmed to generate all similar phrases to the manually tagged cuisine list. All these words/phrases are then saved in a file. The following is the result generated by trained model with default parameters.

chick peas	dal makhni	channa masala	followed by	saag aloo	green peppers
mixed vegetables	green peas	chana masala	mutter paneer	mango ice	chicken tikka
chicken tiki	chicken tika	paneer tikka	chicken tikki	dosi	chana
bhindi	channa	paprika	chala	flat bread	fresh baked
well seasoned	freshly made	mango chutney	garlic nan	garnished with	mango ice
side salad	plain naan	mango pudding	tandoori chicken	tikka masala	saag paneer
butter chicken	goat curry	palak paneer	lamb curry	lamb vindaloo	malai kofta
chicken korma	aloo gobi	rogan josh	aloo gobi	chana masala	fish curry
chicken biryani	bhindi masala	veggie korma	chicken makhani	shrimp vindaloo	vegetable biryani
chili chicken	gulab jamun	rice pudding	mint chutney	perfectly spiced	garlic nan
plain naan	lentil soup	carrot halwa	aloo gobi	tomato soup	chana masala
basmati rice	vegetable korma	saag paneer	plain naan	aloo gobi	chicken makhani
chana masala	along with	lamb curry	garlic nan	gulab jamun	rice pudding
gulab jamun	mint chutney	carrot halwa	plain naan	aloo gobi	saag paneer
chicken biryani	lamb curry	garlic nan	lentil soup	hot sauce	naan bread
spicy	sour	crispy	savory	right amount	spicies
heat	mild	creamy	iced tea	samosa chat	free mango
chickpea ceviche	total bill	my companion	shortly after	my son	father day
my girlfriend	my daughter	fried rice	mango chutney	tomato based	chili chicken
mango ice	mutter paneer	chana masala	dal makhani	fish curry	roasted
vegetable biryani	ice cream	raita yogurt	garbanzo beans	puffy bread	homemade cheese
mint chutney	passion fruit	minced lamb	tamarind sauce	chili sauce	paratha bread
tomato soup	lentil soup	chili chicken	dal makhani	mango ice	gulab jamun

veggie korma	tamarind chutney	mint chutney	creamy spinach	tikki masala	chicken tikka masala
lamb	butter chicken	wings	tikka masala	korma	saag
tandoori chicken	lamb vindaloo	paneer	chicken vindaloo	chinese food	cuisine
pakistani	indian cuisine	cooking	authentic	south	fare
northern	lunch buffets	other places	white rice	chili sauce	deep fried
garbanzo beans	homemade cheese	fried dough	ice cream	black lentils	creamy sauce
covered in	minced lamb	tomato sauce	creamy sauce	your mouth	garbanzo beans
creamy tomato	boneless chicken	chili sauce	cheese cubes	ground beef	white rice
hint of	brown rice	jasmine rice	each dish	lentil soup	ice cream
mint chutney	aloo tikki	mango ice	white rice	tandori chicken	homemade cheese
chicken wings	fried cheese	galub jamun	lentil soup	aloo tikki	goat karahi
goat biryani	vegetable samosa	palek paneer	onion bhaji	baingan bharta	tikka masala
tandoori chicken	butter chicken	saag paneer	goat curry	lamb curry	lamb vindaloo
chicken korma	aloo gobi	palak paneer	chicken vindaloo	coconut chicken	shrimp
okra	butter chicken	fish	tikka masala	tandoori chicken	saag paneer
dal	tofu	beef			

Opinion

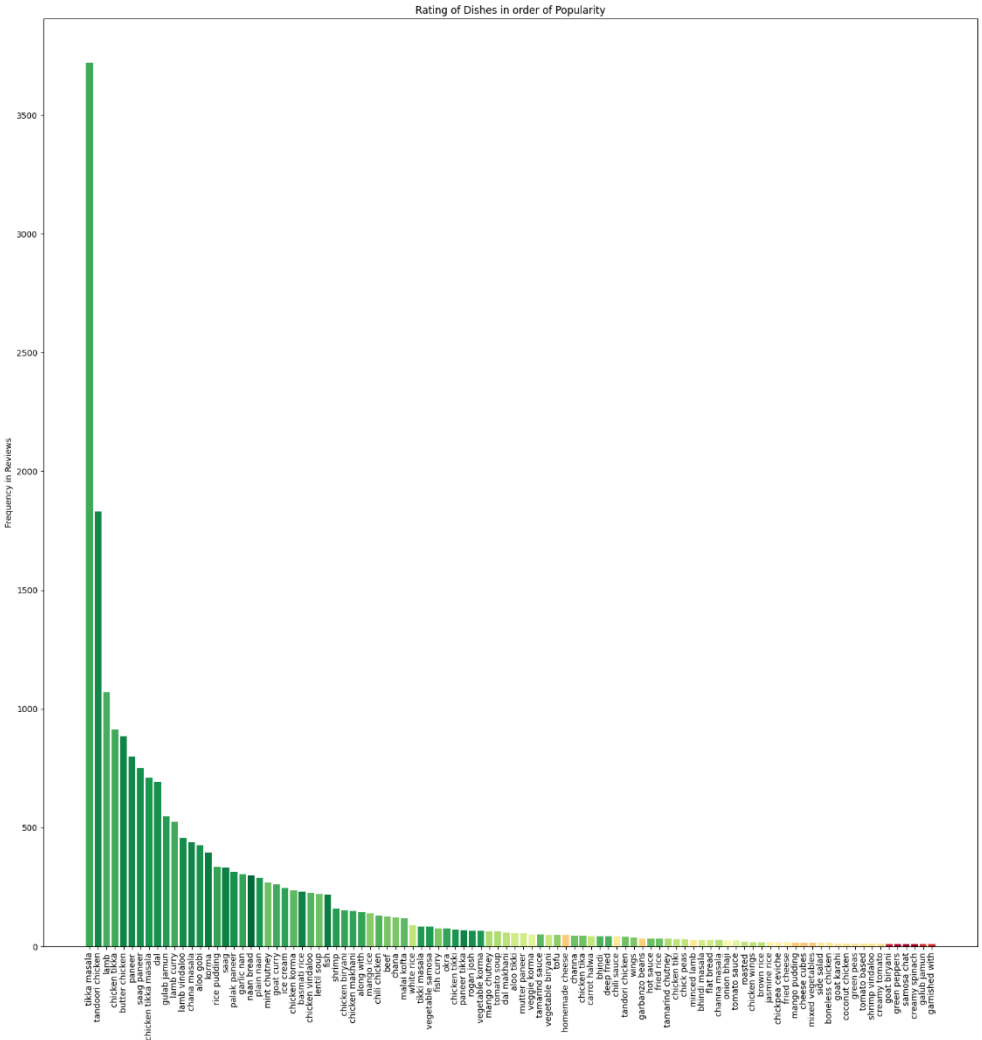
The results achieved from using word association based word2vec model are high quality. The model has been able to generate a long and mostly accurate list of cuisines. Since the model is trained on Indian Restaurant reviews only the list is seeming to be a good reflection of an Indian cuisine menu. The results also include different spelling variations of the same cuisine. Such a model can be useful in generating a cuisine list and can be used a basis for a dish recommendation system.

Task 4: Mining Popular Dishes

The task is based on mining a list of popular dishes in “Indian” cuisine. The list of Indian cuisine from task 3 and reviews from Yelp datasets are combined. This is done by extracting reviews and their ratings specific for the restaurants that are tagged in the “Indian” category. For each cuisine, the number of times it is mentioned in reviews and the sentiment of reviews is recorded. Dishes are then ranked based on the frequency of their appearance in the reviews. An average rating of reviews that mention the dish name is also recorded to reflect the overall sentiment behind these reviews. All the data pre-processing and visualization is done in Python.

The following visualisation shows the popularity/frequency of the dishes based on the reviews of Indian restaurants. The colour of the bars that goes from green-yellow-red reflects the overall sentiment of the reviews that include the dish name. For example, naan bread being dark green and having a bar height of around 350 means that “naan bread” was mentioned in 350 reviews and all

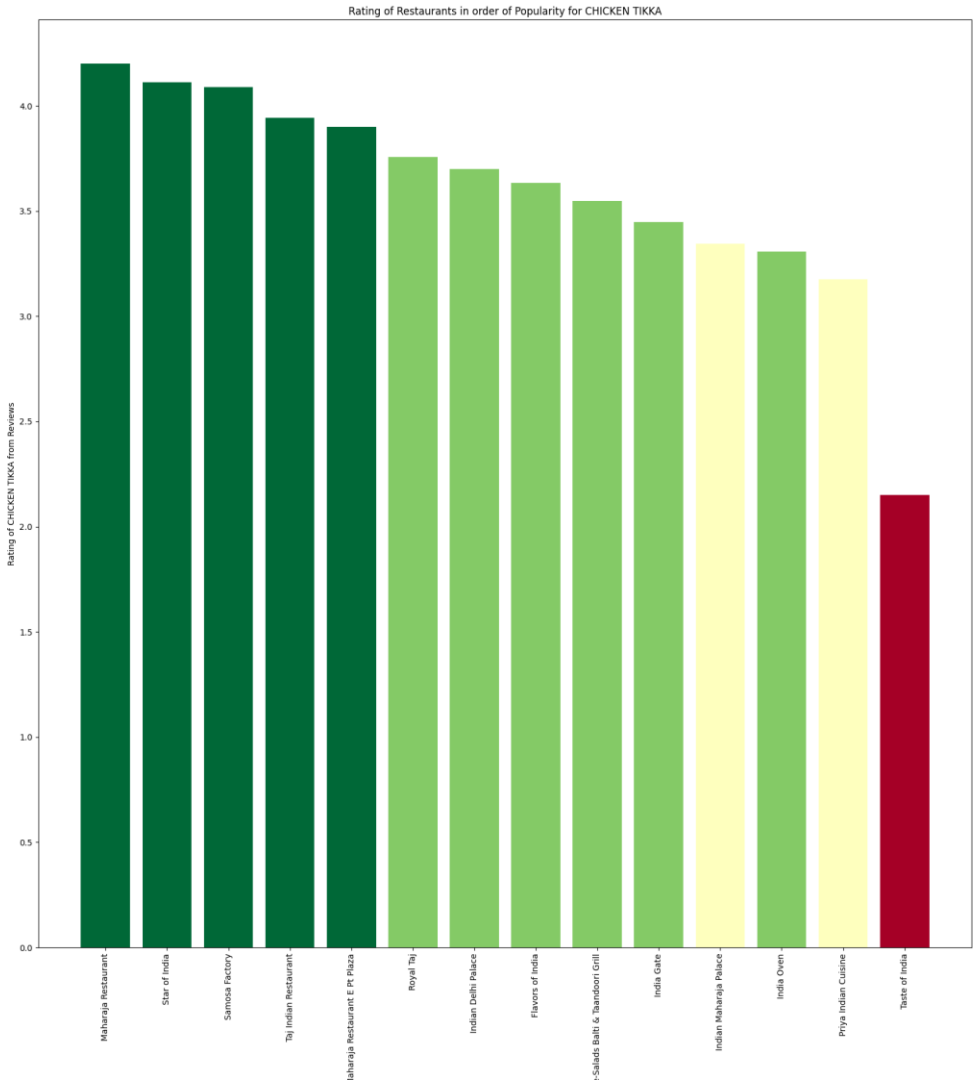
these reviews on average were positive. This system can easily be upgraded as a recommender system for any type of cuisine by replacing the dish list to another category.



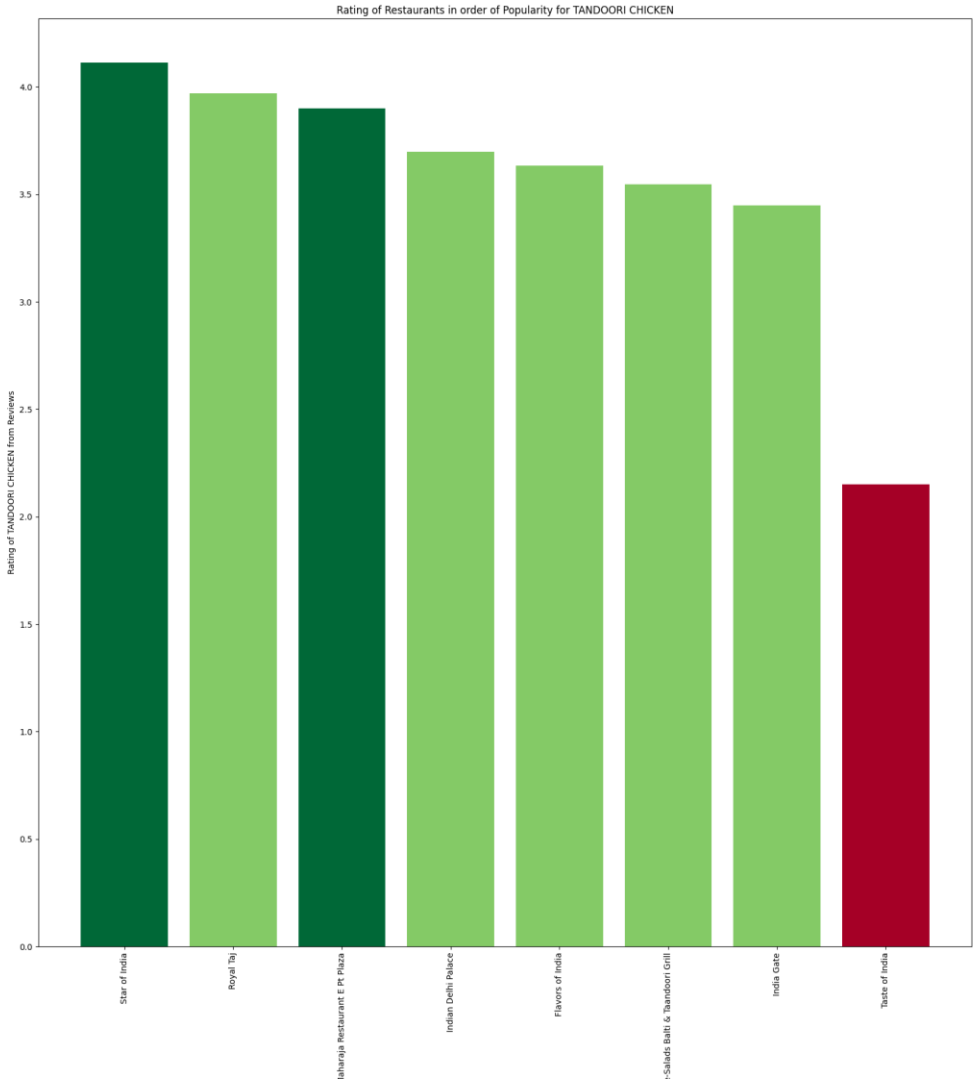
Task 5: Restaurant Recommendation

Task 5 involves developing a rating system for restaurants for a particular dish in a category. The dataset used is based on the “Indian” cuisine list from task 3 and yelp dataset. For each cuisine the restaurants and their ratings are recorded. All reviews are filtered for each cuisine and restaurant and average rating of these reviews is calculated. The ranking is done based on the average review rating of a cuisine for all the retaurants separately.

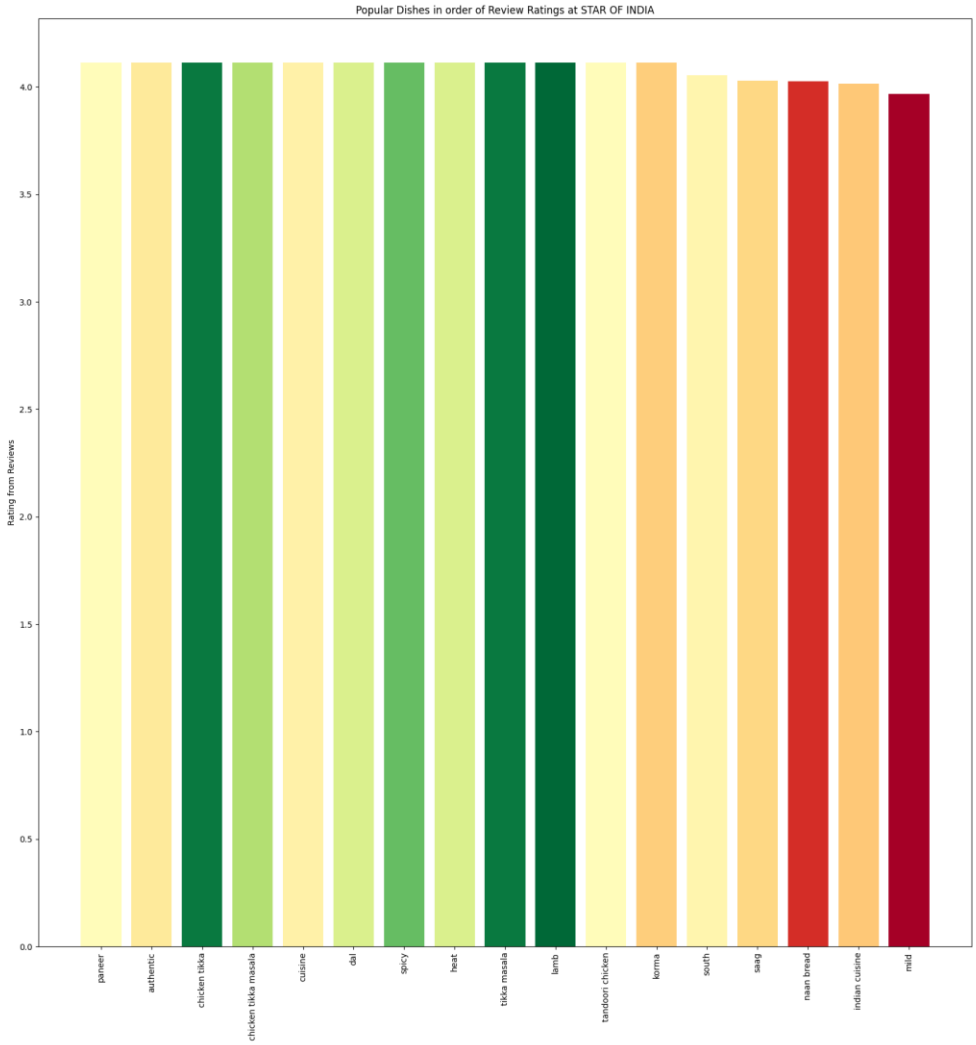
The following visualization shows the average review rating of restaurants for “Chicken tikka” . The color of the bars go from green to yellow to red and indicate the overall rating of the restaurant. For example, based on the reviews rating “Maharaja Restaurant” seems to be the best choice for “Chicken tikka” and also has a good reputation itself.



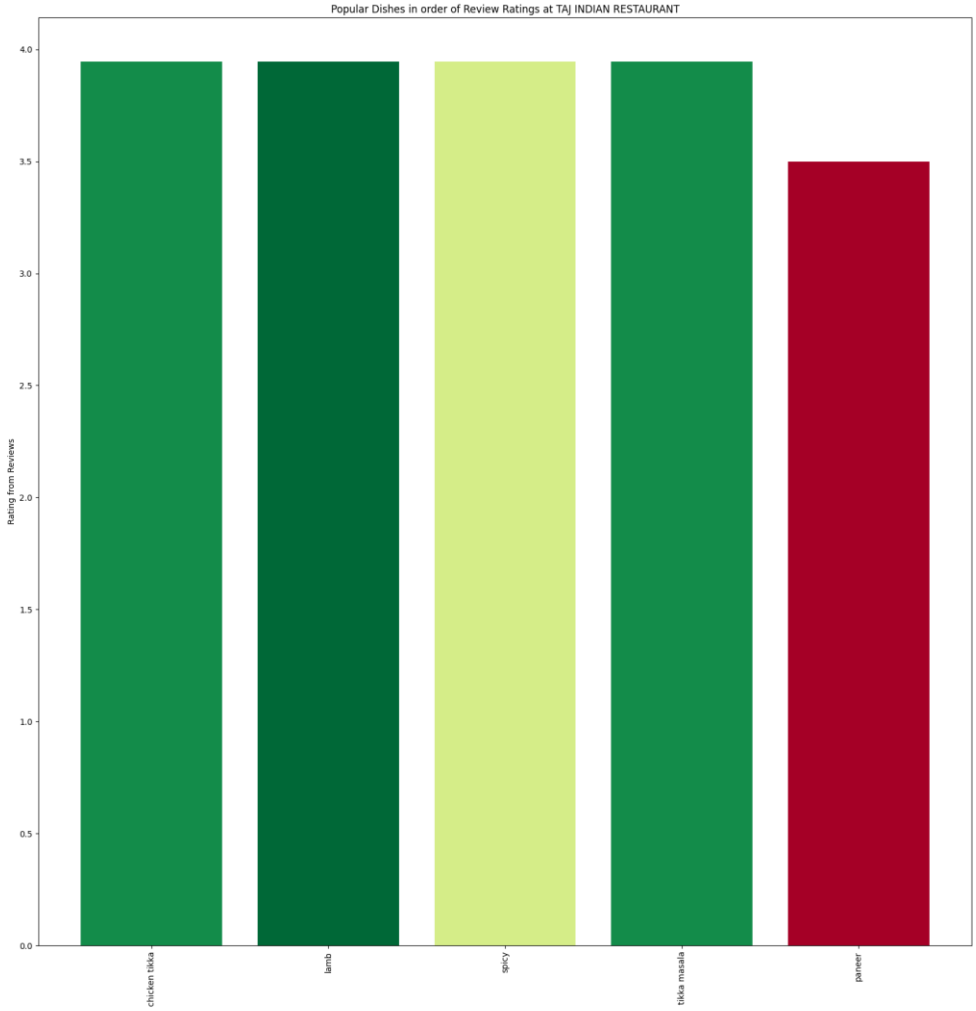
“Taste of India” has “Tandoori Chicken” with a low review rating and not many people recommend the restaurant either.



This can for the basis of a recommender system for a dish. This system can be extended to recommend a dish from a certain restaurant as shown in the visualisation below. The height of the bars shows the average review rating for different dishes and the colour of the bar is based on the popularity of dish in the “Indian” category that is calculated based on the number of times the dish is mentioned in reviews of all the restaurants. For example, paneer is getting good reviews at “Star of India” but it is not popularly reviewed.


Popular Dishes in order of Review Ratings at STAR OF INDIA
Rating from Reviews
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
paneer
authentic
chicken tikka
chicken tikka masala
cuisine
dal
spicy
heat
tikka masala
lamb
tandoori chicken
korma
south
saag
naan bread
indian cuisine
mild

Lamb seems to be a very popular dish in “Indian Taj Restaurant” with good reviews.



Task 6: Predict Hygiene Inspection Results

Text Representation

The text of reviews is represented in two ways. The first one is the unigram model and the second is bigram model. TDFIDF vectorizer Is used to convert reviews into a vector. All reviews in the file are first cleaned up by removing English stop-words and punctuation. Then they are converted into a cleaned list of stem words. These words are then converted into TDFIDF vector by using preprocessing module of Scikit-Learn library from Python. Different number of features were selected and tested out.

Features Selection

Python dataframe is used to store the features of the reviews and the target value that represents whether the restaurant has passed hygiene test or not. The features of data X include the values zip code, average rating, number of reviews, cuisines and text of reviews. Each cuisine listed in the data

file is converted into a separate feature and is given 1 or 0 value for each restaurant with 1 representing that cuisine being offered by the restaurant and 0 denoting not offered in the menu. These values are all concatenated with the reviews vector representation as described earlier in a datframe with each row corresponding to one restaurant.

Models:

The prepared dataframe and target values are then used to train machine learning models from SciKit Learn library of Python. The learning models and parameters used are:

Logistic Regression: regularization parameter C = 1

Gradient Boosting Regression: learning_rate = 0.1, n_estimators = 100

Gaussian Naïve Bayes

Support Vector Machine (SVM): C=1, kernel= ”rbf”

Decision Tree

Results

The data was split into training and test set to gauge performance.

The following were the results of F1 score based on macro averaging for different models:

Text Representation	Number of features	Logistic Regression	Gradient Boosting Regression	Gaussian Naïve Bayes	Support Vector Machine (radial)	Decision Tree
Unigram	50	0.612	0.365	0.368	0.423	0.576
Unigram	100	0.663	0.380	0.392	0.364	0.540
Bigram	20	0.620	0.394	0.458	0.401	0.562
Bigram	50	0.599	0.347	0.385	0.414	0.577

The results show that in general Logistic Regression performs the best on test data. Decision Tree algorithm also showed promising results. The data also shows that involving more features in developing the model does not always improve the F1 score. The failed cases might have been reduced by using a bigger dataset to improve learning and accuracy.

Usefulness of Results

- Task 1 Visualisations results show an interesting pattern in the reviews. There are common themes and topics that can be seen in all three groups of reviews for example, food, dog, place etc. and some words that are specific to one category of reviews only like time, great and ok. This is expected as the domain words that are specific to dining and restaurants appear frequently in all types of reviews and do not indicate the intention of the user. It is interesting to note that words like just, order really appear in negative reviews and friendly, love etc. appear as a part of positive reviews topic modelling. Similarly, word “ok” stands out in neutral reviews. This gives an indication of how humans use language vocabulary to express negative and positive feedback and is useful in predicting ratings for a restaurant based on reviews.
- In Task 2 cuisines from geographically close areas have ended up in same clusters for example the cluster with Pakistani and Himalayan cuisines and cluster with Malaysian,

Singaporean, Thai cuisines. This shows that reviews from tagged restaurants have been a useful representation for developing a cuisine map and can be useful basis for a recommender system that suggests a new cuisine to a diner based on their current preferences.

- Task 3 results achieved from using word association based word2vec model are high quality. The model has been able to generate a long and mostly accurate list of cuisines. Since the model is trained on Indian Restaurant reviews only the list is seeming to be a good reflection of an Indian cuisine menu. The results also include different spelling variations of the same cuisine. Such a model can be useful in generating a cuisine list and can be used a basis for a dish recommendation system.
- Task 4 results show the popularity/frequency of the dishes based on the reviews of Indian restaurants. This ranking system can easily be upgraded as a recommender system for any type of cuisine by replacing the dish list to another category of cuisine. The ranking system can also be useful is case of a search query to rank the dishes in order.
- Task 5 results can be used by a recommender system for a dish to try from a particular cuisine. This system can be extended to recommend a dish from a certain restaurant as well.
- Task 6 provides a supervised machine learning model to make predictions about restaurants. This model can be useful in recommending restaurants based on hygiene and can be a part of a bigger prediction system that evaluates if a user will like to dine in a restaurant or not.

Novelty of Exploration

To generate the visualisations and clean dataset, different tools are used in coherence e.g., excel, PowerBI, Python libraries. It was an interesting experience to see how different tools can be synched to get appropriate data mining results. Task 5 was extended to generate a list of popular dishes of a cuisine based on their popularity with sentiment analysis included. Restaurant recommendation included rating indication for the restaurant as well as the sentiment of the reviews it received for a particular dish. A system was also developed to recommend dishes in order of their popularity given a particular restaurant. In task 4 word2vec model with bigrams was used to generate dish names with multiple words.

Contribution of New Knowledge

I think this project has provided a basis for a recommender system for diners. They can search up similar cuisines as well as dishes. They can get restaurant recommendations for dish and a dish recommendation for a restaurant. I think these tasks have all the building blocks for building a search engine and recommendation system for restaurants.