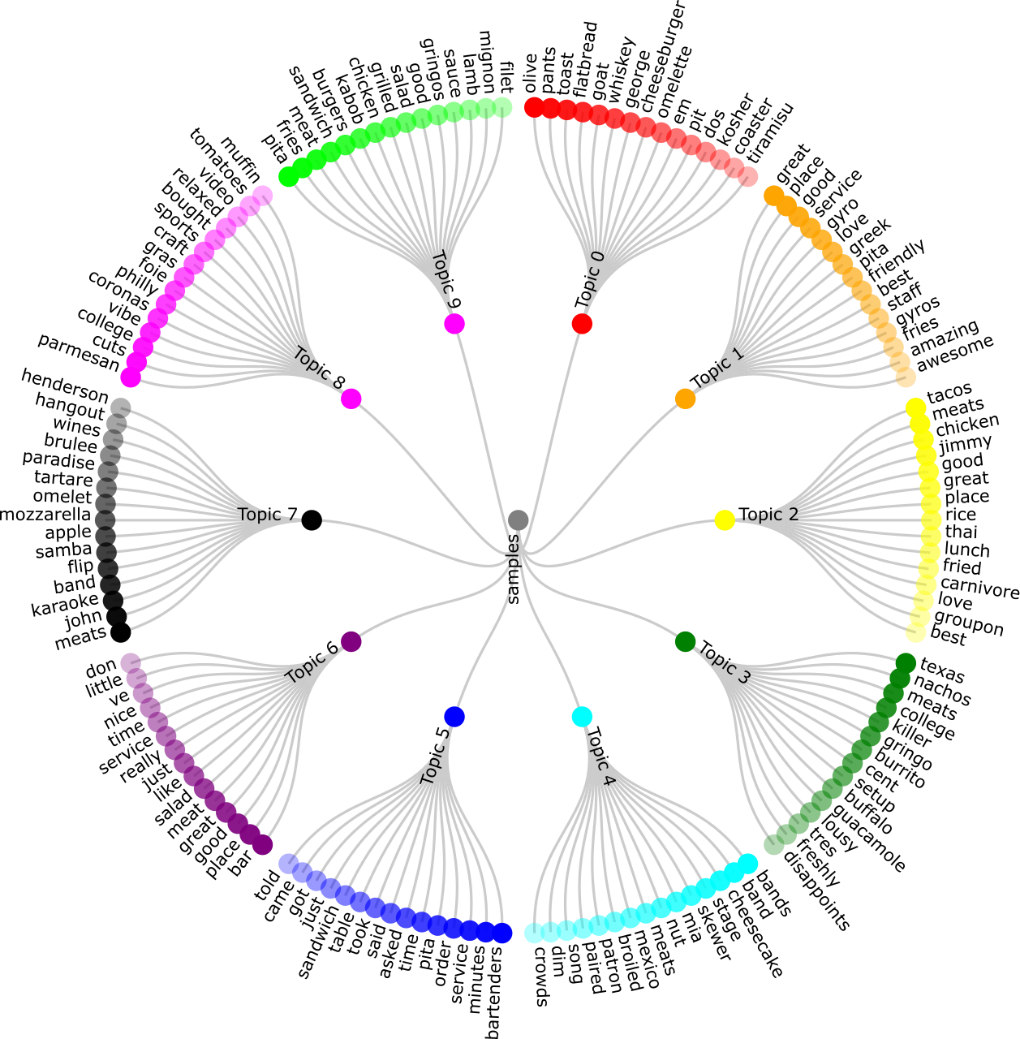
Final Report

# Activities Summary

## Task 1: Exploration of Data Set

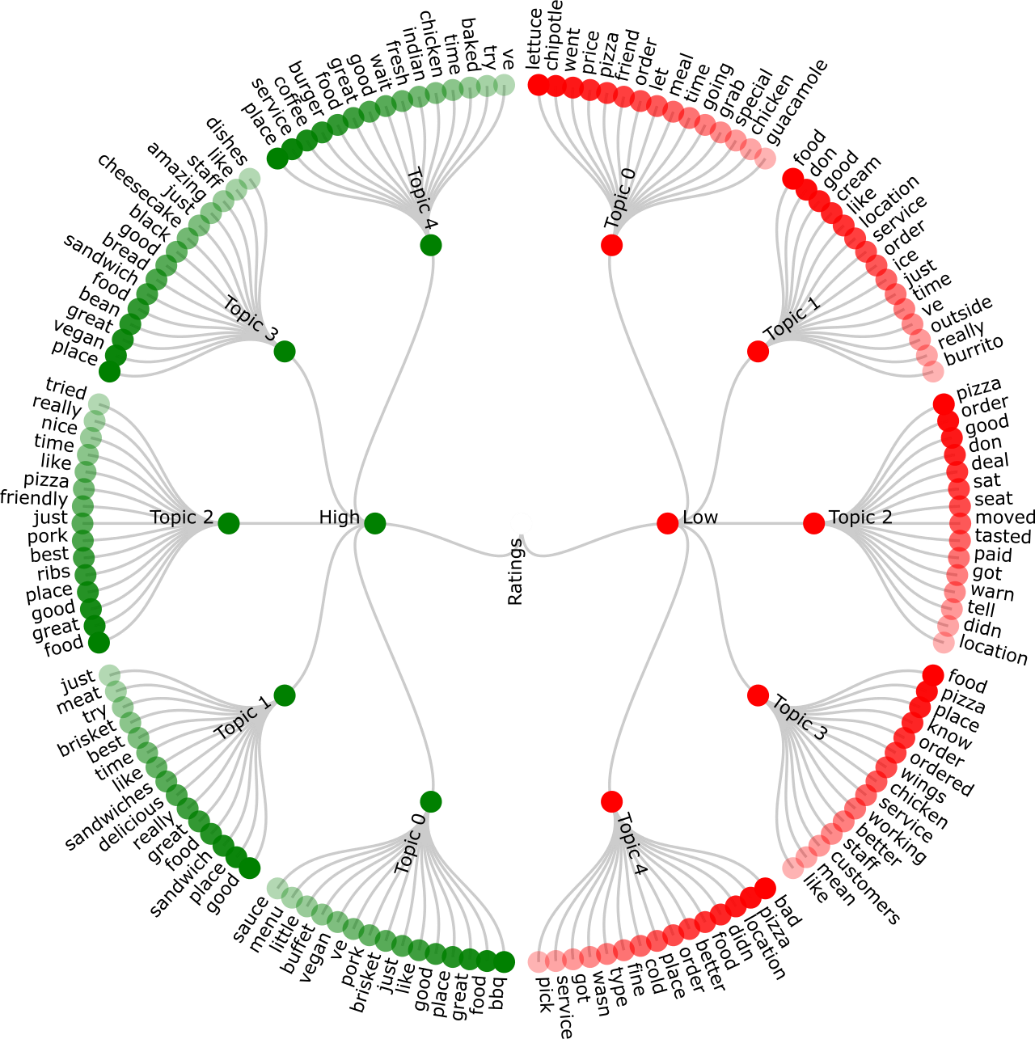
Full Task 1 Report: <https://github.com/pauldeng/MOOC/blob/master/Data%20Mining%20Capstone/Task%201/Task%201.docx>

### Task 1.1



|  |  |
| --- | --- |
| Data | 100000 samples reviews from the Yelp review |
| Algorithms Applied | LDA topic model |
| Results | As above |

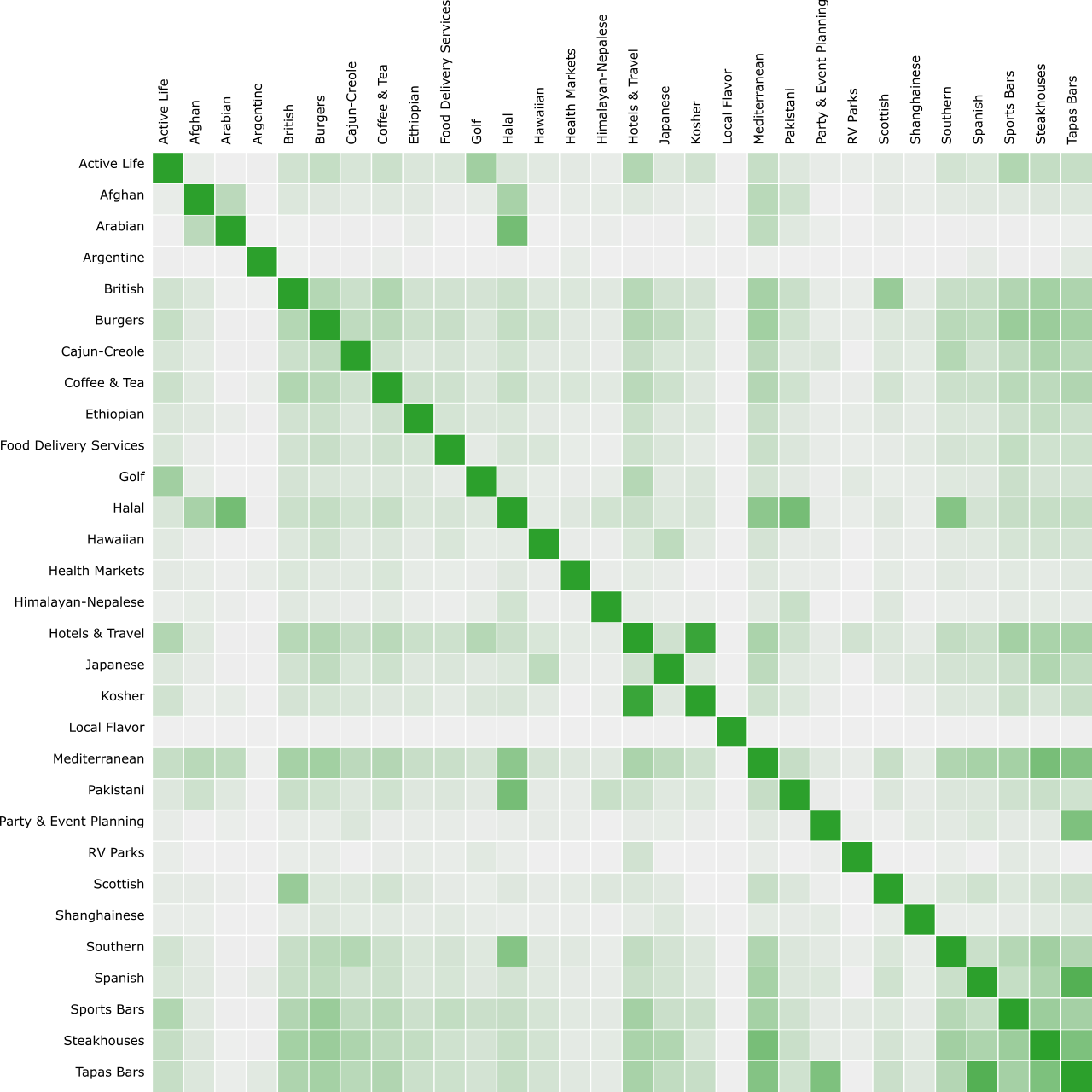
### Taks 1.2



|  |  |
| --- | --- |
| Data | Maximum 10000 samples review per each rating group from the Yelp review   * High Ratings == 5 * Low Ratings == 1 |
| Algorithms Applied | LDA topic model |
| Results | As above |

## Task 2: Cuisine Clustering and Map Construction

Full Task 2 Report: <https://github.com/pauldeng/MOOC/blob/master/Data%20Mining%20Capstone/Task%202/Task%202.docx>



|  |  |
| --- | --- |
| Data | 30 cuisine reviews randomly selected from Task 2 Cuisines Sample Dataset provided. |
| Algorithms Applied | LDA, IDF, K-Means |
| Results | As above |

## Task 3: Dish Recognition

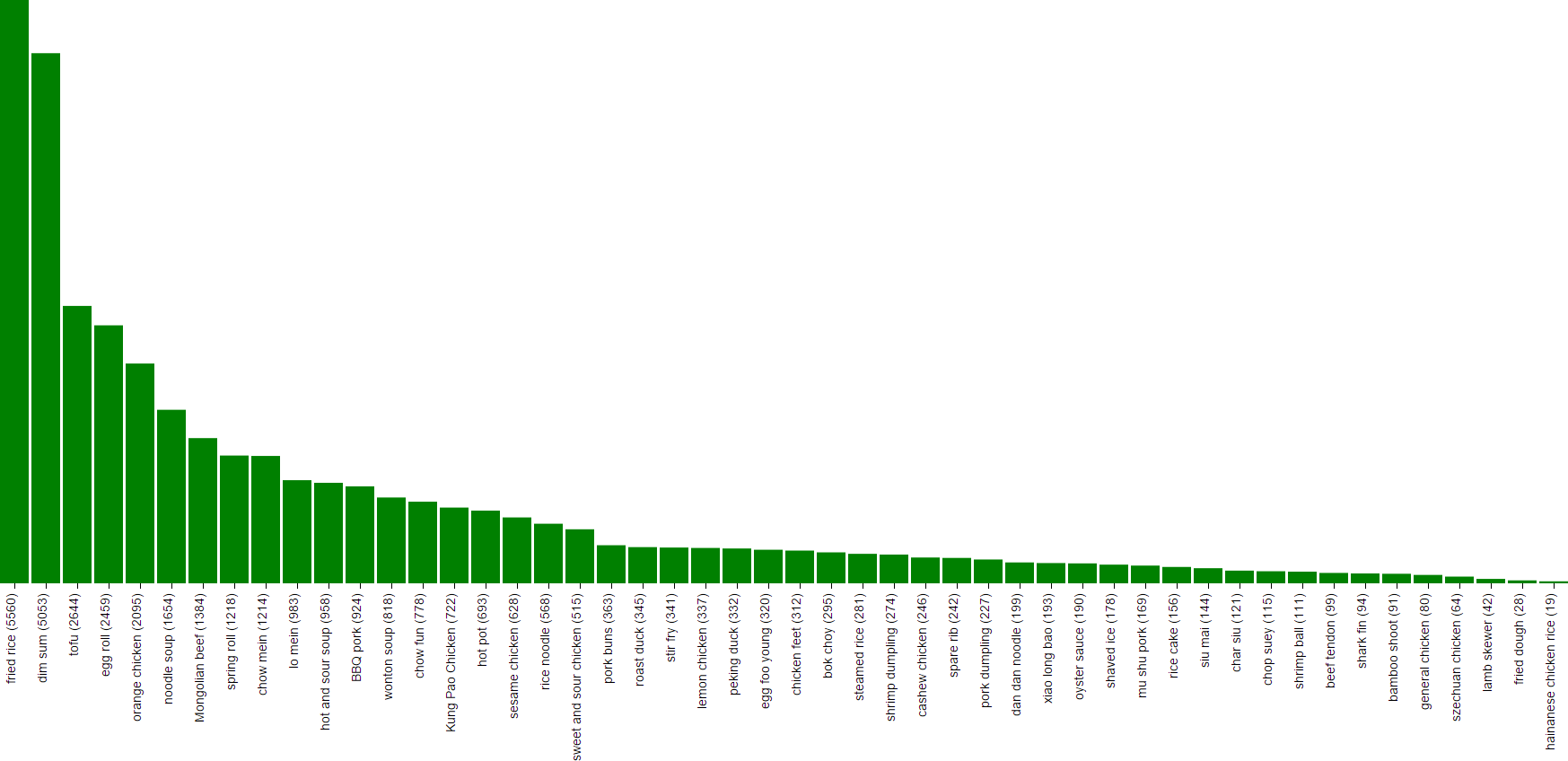
Full Task 3 Report: <https://github.com/pauldeng/MOOC/blob/master/Data%20Mining%20Capstone/Task%203/Task%203.docx>

|  |  |
| --- | --- |
| Data | Chinese cuisine |
| Algorithms Applied | Manually tagged, Mutual information, TOPMINE, SegPhrase, Combined |
| Results | Auto grader gave me 10/10 on the Combined method |

## Task 4 and Task 5: Popular Dishes and Restaurant Recommendation

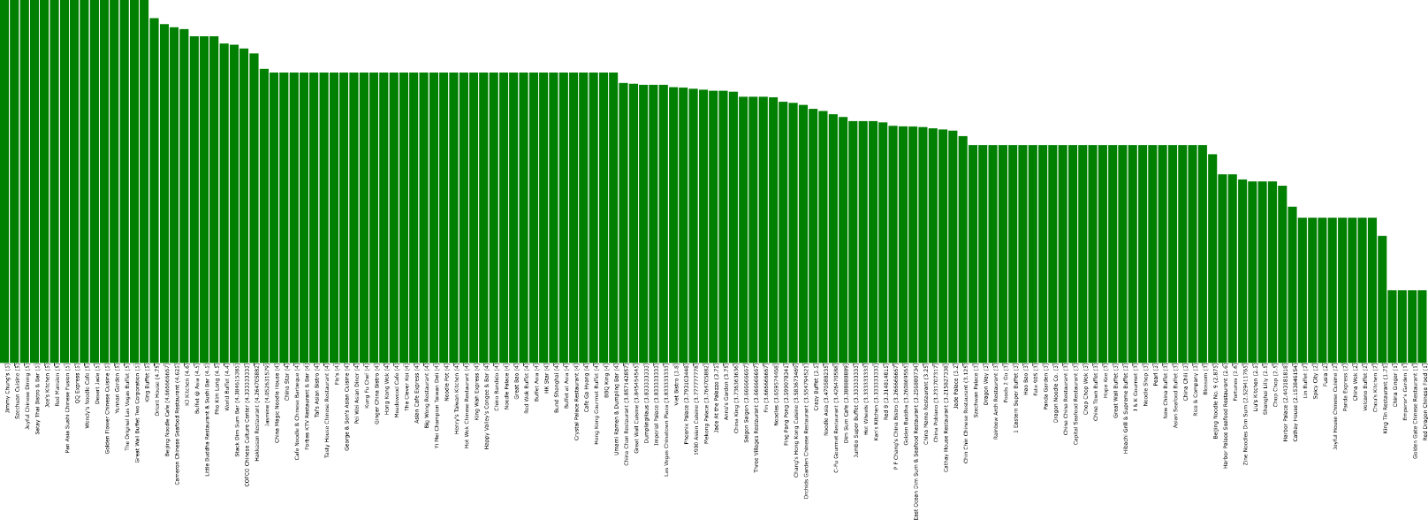
Full Task 4 and Task 5 Report: <https://github.com/pauldeng/MOOC/blob/master/Data%20Mining%20Capstone/Task%204%20%26%20Task%205/Task%204%20%26%20Task%205.docx>

### Task 4



|  |  |
| --- | --- |
| Data | Chinese cuisine |
| Algorithms Applied | Occurrence based, Count Occurrence + Star Rating |
| Results | As above |

### Task 5



|  |  |
| --- | --- |
| Data | Chinese cuisine |
| Algorithms Applied | Occurrence based, Count Occurrence + Star Rating |
| Results | As above |

## Task 6: Hygiene Prediction

Full Task 6 Report: <https://github.com/pauldeng/MOOC/blob/master/Data%20Mining%20Capstone/Task%206/Task%206.docx>

|  |  |
| --- | --- |
| Data | Hygiene data |
| Algorithms Applied | SVM, NaiveBayes, NaiveBayesMultinominal, RandomForest, J48, OneR |
| Results | Auto grader gave me 15/15 on NaiveBayesMultinominal algorithm which produced F1 = 0.55 |

# Project Highlights

## Usefulness of Results

* Task 1 results:
  + extract topics from Yelp data set to get a sense about what the data look like and their characteristics. It helps customer to what people talk about in general
  + extracts topics from good and bad reviews to get a general ideal about what makes good restraint good and why people loving it and vice versa. It will help customer where to go and for restraint to spot things that they can improve.
* Taks 2 results:
  + construct a cuisine map to visually understand the landscape of different types of cuisines and their similarities. The cuisine map can help users understand what cuisines are available and their relations, which allows for the discovery of new cuisines, thus facilitating exploration of unfamiliar cuisines
* Task 3 results:
  + discover the common/popular dishes of a particular cuisine. It helps customer try new cuisine.
* Task 4 results:
  + leverage recognized dish names to further help people making dining decisions. It mines popular dishes in a cuisine that are liked by people; this can be very useful for people who would be interested in trying a cuisine that they might not be familiar with.
* Task 5 results:
  + leverage recognized dish names to further help people making dining decisions.  It recommends restaurants to people who would like to have a particular dish or a certain type of dishes. This is directly useful to help people choose where to dine.
* Task 6 results:
  + Predict whether a set of restaurants will pass the public health inspection tests given the corresponding Yelp text reviews along with some additional information such as the locations and cuisines offered in these restaurants.

## Novelty of Exploration

### Tools used

* Weka 3.7
* D3.js
* MeTa
* InkScape
* Notepad++
* Python
* Java

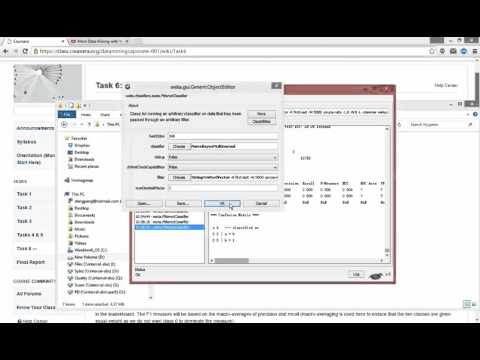
I used different algorithms and parameters to experiment for different tasks, e.g. Task 2, Taks 3, Task 4, Task 5 and Task 6.

## Contribution of New Knowledge

I wrote 2 blog posts regarding the new knowledge could possible useful to others:

### How to use Weka 3.7 Video Tutorial

Blog address: <http://sensorapp.net/text-classification-using-weka-3-7/>

[](https://www.youtube.com/watch?v=wg5KOhvTqmE)

### How to Embed D3 into WordPress Tutorial

Blog address: <http://sensorapp.net/tutorial-how-to-add-d3-to-wordpress/>