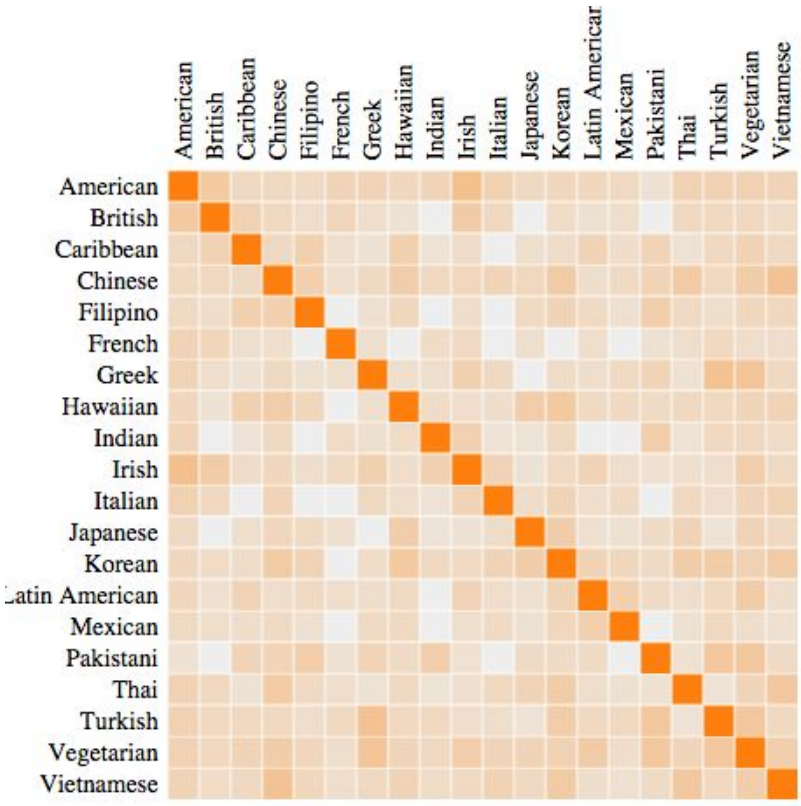
GETTING DATA

First Yelp dataset was loaded into mongodb. Using node.js script I generated datasets for 20 different cuisines. After I applied LDI model, on each of them to generate topic files.

All calculations were done with use of multiple node.js packages, and scripts i wrote. Visualisations were done with d3.js libraries. All source code is available on github.com, licensed under MIT, and you are free to use it to replicate the results of this report.

TASK 2.1

First I wrote script to get numeric representation of each topic, by simply giving each topic unique numerical value. This allowed me to use *cosine similarity* algorithm to calculate similarity between cuisines.

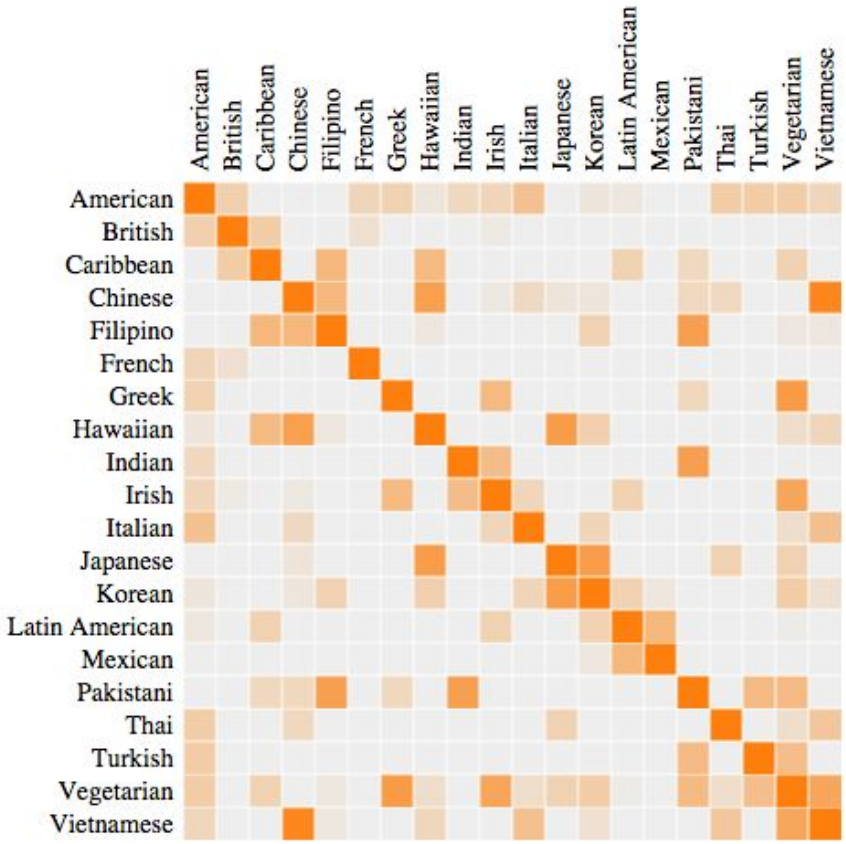


Above is a similarity matrix, with each cell representing similarity between two

cuisines. The opacity of each cell is the similarity - with a higher opacity for a higher similarity.

TASK 2.2

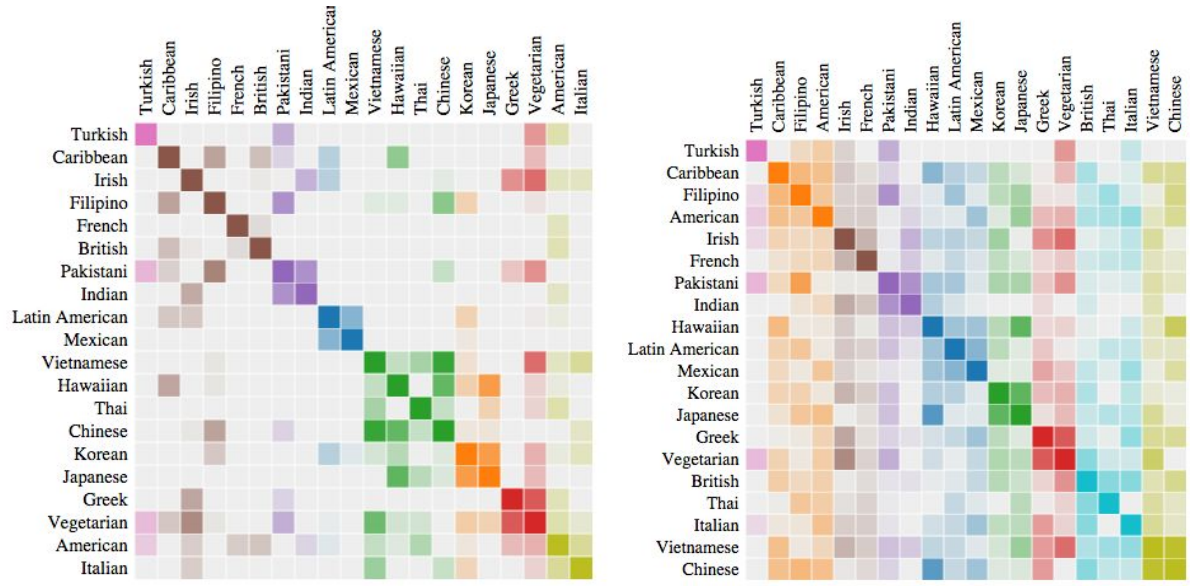
After analyzing topics for each of the cuisine, i made a list of topics that were present in majority of the categories, but were not significant for analyze, as they were not representing a cuisine, but a representation of a restaurant as an establishment (eg. table, order, menu). After removing those words, topic lists were left only with topics that were describing meal, or ingredients that are used in cuisine.



Above visualisation shows more prominent similarities between cuisines, and has less noise.

TASK 2.3

I wrote my own clustering algorithms. It's somehow based on k-mean clustering but instead of using point coordinates it uses cuisine similarity values. Below you can see results of the visualisation. Different clusters represented with different colours.



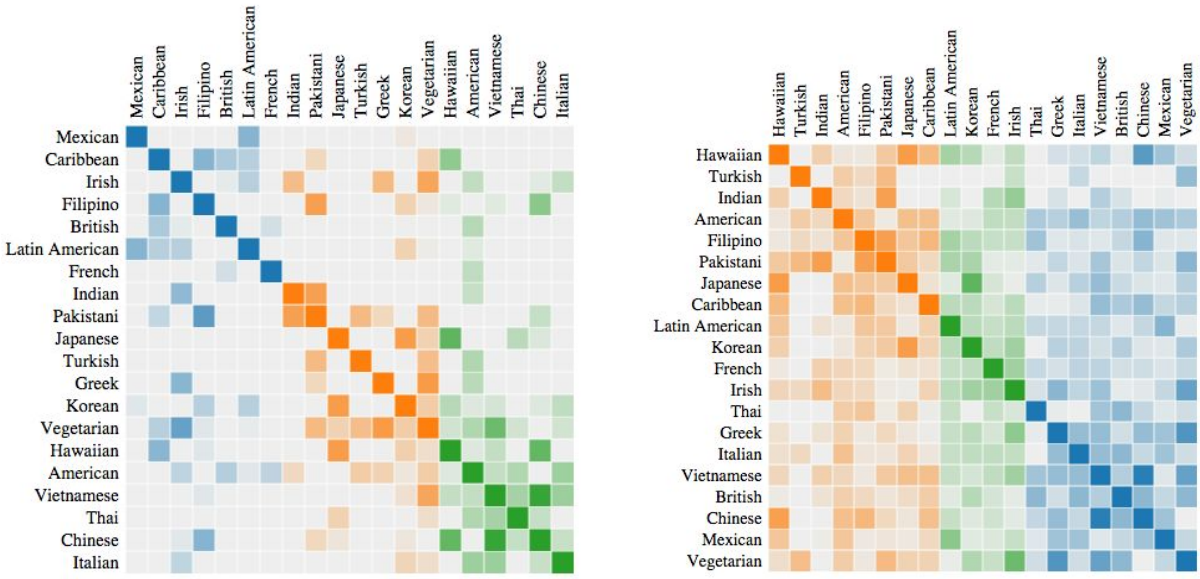
Left visualisation is using improved topic modeling weights from the task 2.2, and right visualisation is using default weights from task 2.1.

As you can see algorithm was able to distinct multiple clusters with most prominent asian food clusters:

- Indian, Pakistani
- Japanese, Korean
- Vietnamese, Thai, Chinese, Hawaiian(not really asian, and still a surprise for me)

In this visualisation clustering were set to limit number of generated clusters to 3. You still can see similarities in the cuisines that were chosen in each of the cluster, but it's obvious that limiting this algorithm to such low amount of clusters giving results that

could seems not logical to human judgement.



Left visualisation is using improved topic modeling weights from the task 2.2, and right visualisation is using default weights from task 2.1.