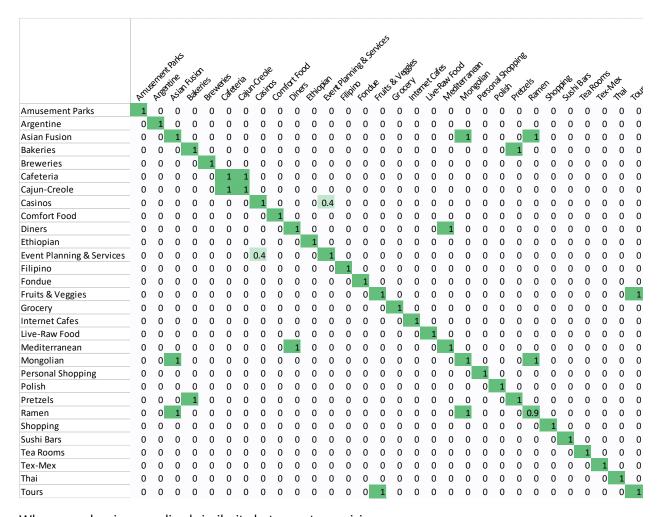
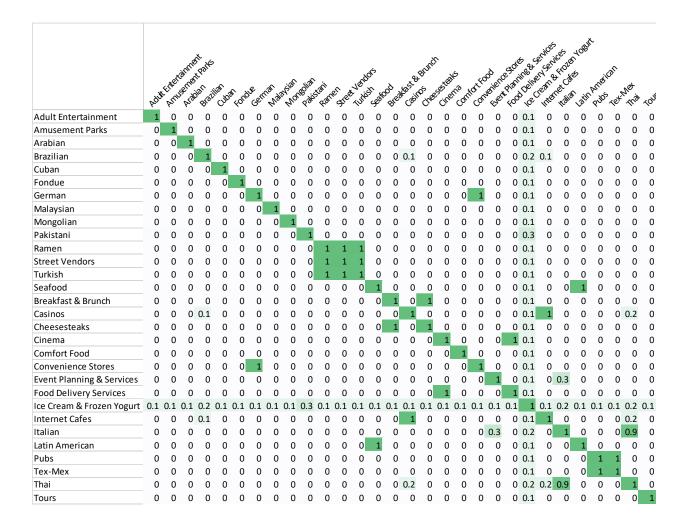
$\label{eq:task-2.1} \text{This is a similarity matrix between different cuisines from IDF}$



Where number is normalized similarity between two cuisines.

Here it indicates that There are barely any similarities between topics. This is because IDF is too sparse and it is difficult to identify similarity in such a sparse vector.

I also used LDA model to generate topics model and compare similarity between different cuisines based on topics and its corresponding weights, which is already provided in sample code.

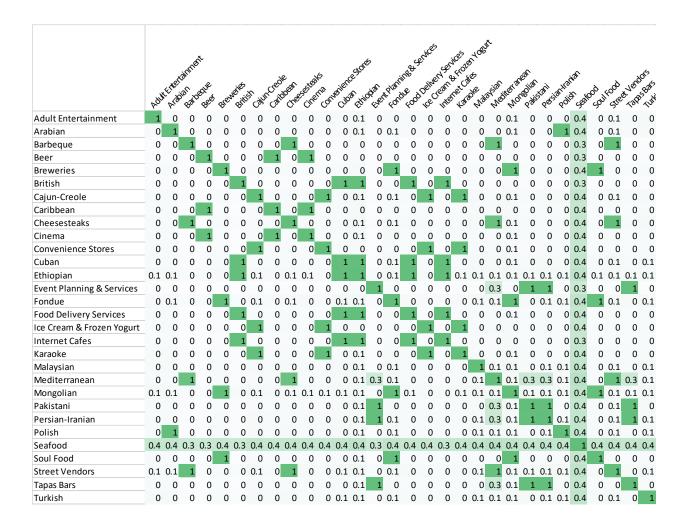


Here it indicates that more similarity is discovered using LDA model. This is mainly because topic models are more condense vectors, and more significant topics will provide higher weight when comparing similarities.

But the result by no mine is meaningful, because there are many similarities doesn't make sense. For example, "Ice Cream and Frozen Yogurt" cuisine has almost identical similarity to all other cuisines which is not intuitive. "Turkish" and "Ramen" cuisines are also similar to each other which is not a meaningful result.

Task 2.2

To Improve similarity discovery, I decreased the topics generated from LDA model, and new similarity matrix is as following:



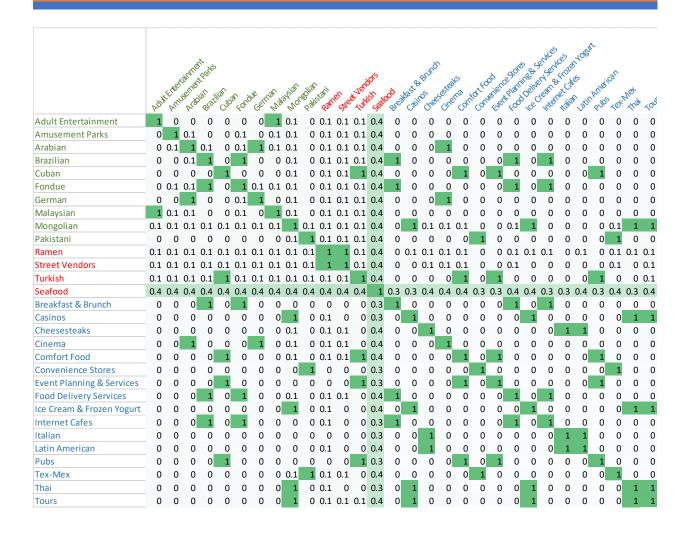
As seen, there are more similarities discovered in this topic model setting.

But similar topics are not grouped together in this confusion matrix. This is improved in Task 2.3

Task 2.3

I used K-means to cluster different cuisines. Feature of teach cuisine is the normalized sum of weights in different topics.

The first trial is to cluster all cuisines to 3 different groups. And this is the similarity matrix



The second trial is to cluster all cuisines to 4 different groups. And this is the similarity matrix



Both clustering results improved the quality of this confusion matrix as similar cuisines are grouped together.