Task_2_Report

For this task, I pick 50 categories from the given sample set of reviews randomly. The picked data is used for all this task.

Data collection and preprocessing

In this task, nltk package is used for data collection and preprocessing. String is used to recognize and remove punctuations. The steps are as follows:

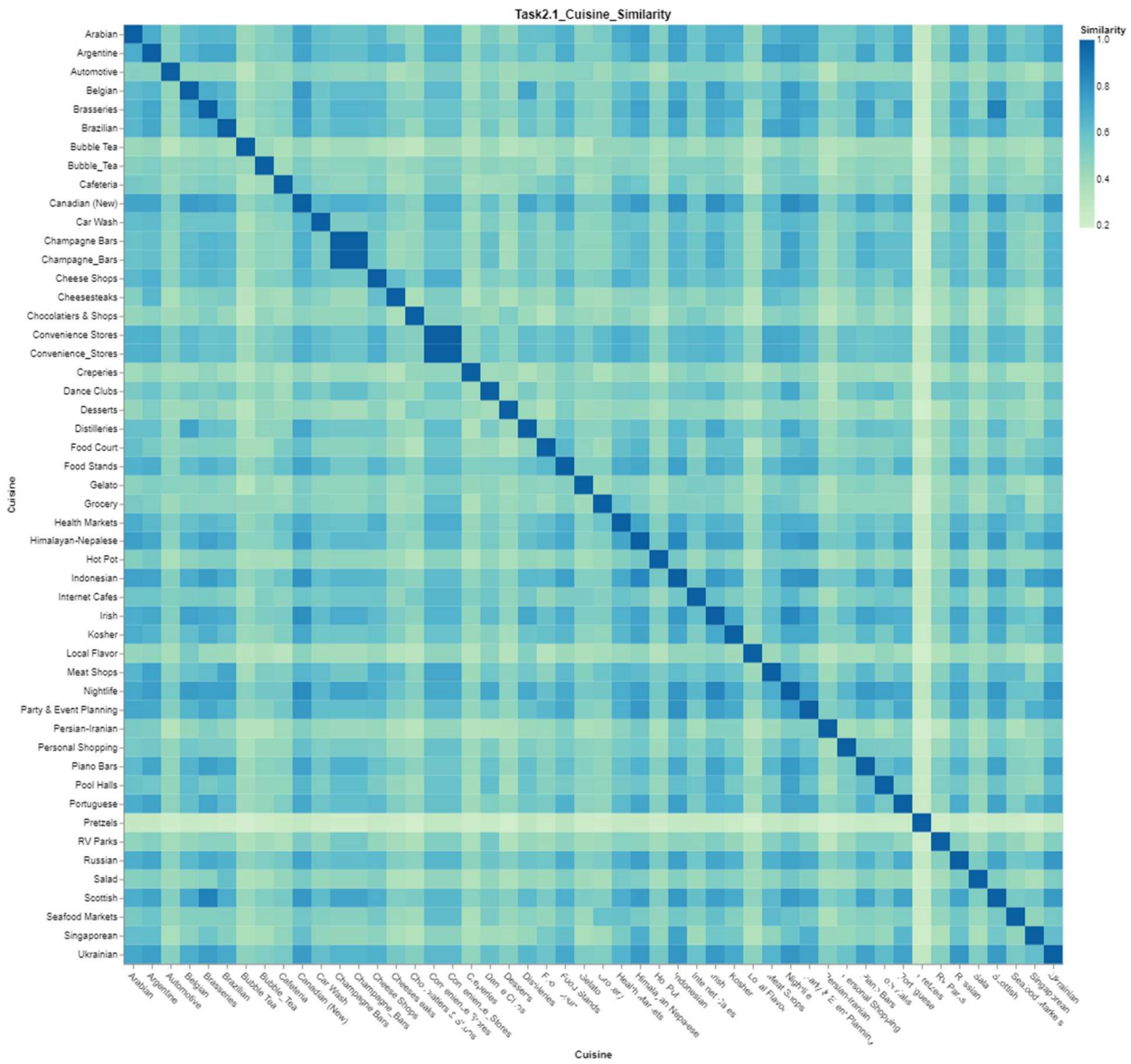
1. nltk.tokenize is used for tokenizing.
2. The stopwords in nltk.corpus is used for removing stopwords from collected data.
3. String.punctuation is used for recognizing and removing punctuations from data.
4. For stemming step, I use porterstemmer from nltk.stem to uniform text and removing affixes.

Data Visualization tools

In this task, I use Altar, a python visualization tool, to generate heat map to visualize the difference of similarity between different categories.

Task2.1: Visualization of the Cuisine Map

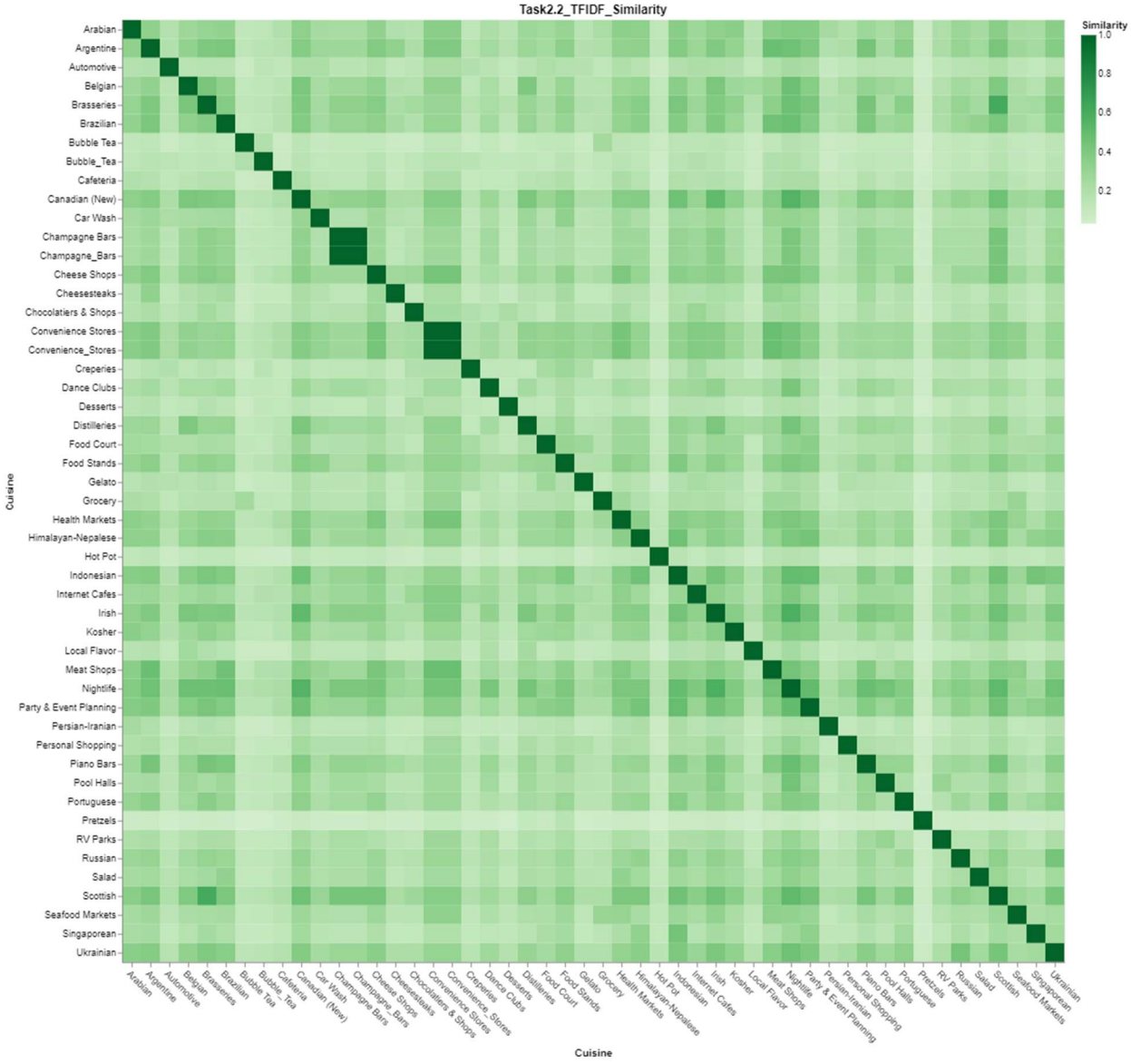
After cleaning the data, I use CountVectorizer to calculate the cuisine similarity of all selected reviews. In this section, I do not use tfidf. Cosine distance is used to calculate similarity in this section.



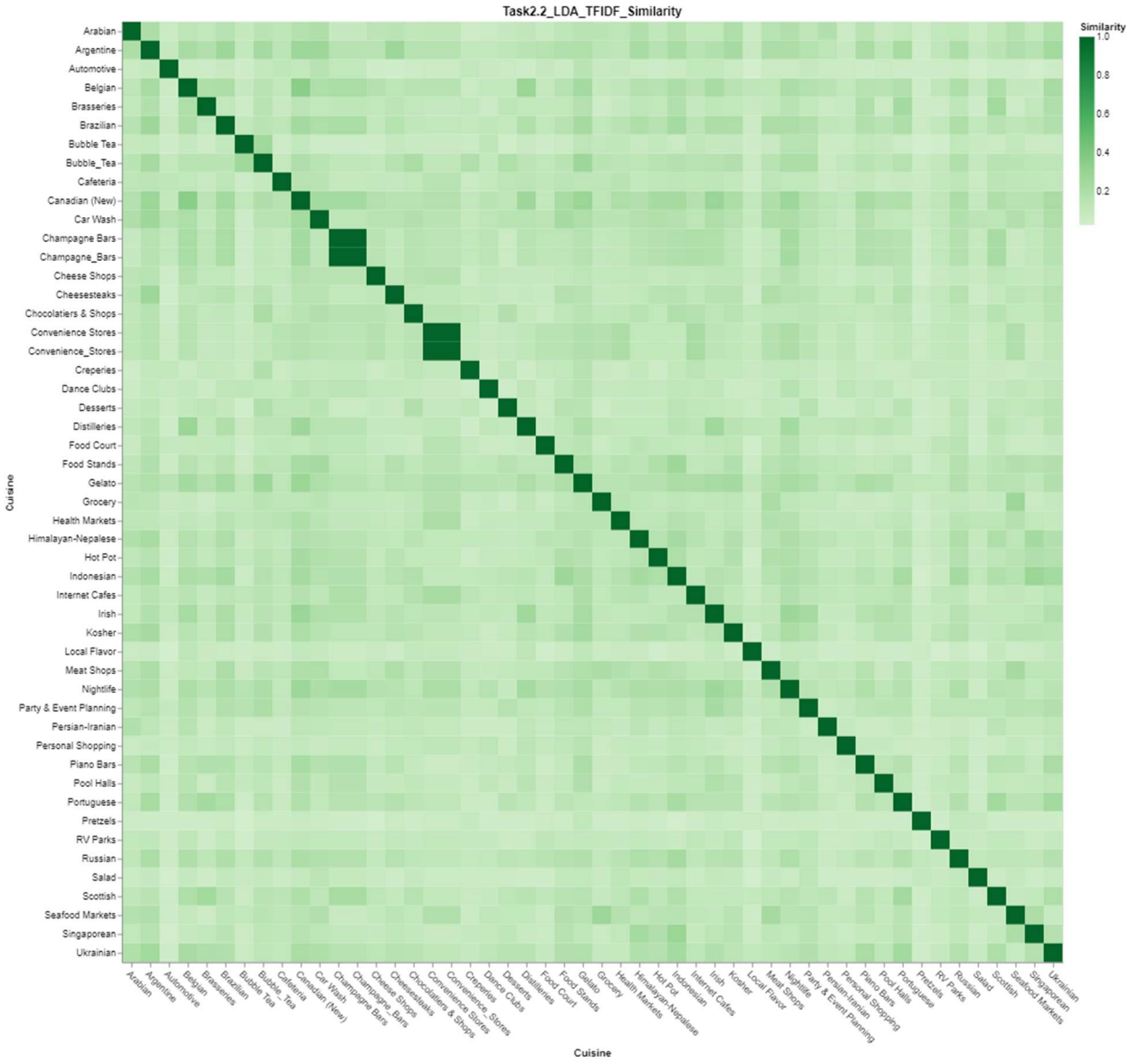
Task2.2: Improving Cuisine Map

In this section, TF-IDF was used to process selected categories. This section has two parts. In the first part, I USE TF-IDF instead of cosine distance to calculate similarity. In the section part, I first use LDA, a topic extraction method, to aggregate 10 topics of each category. Then, TF-IDF was used to calculate the similarity of all selected categories.

Similarity based on TF-IDF



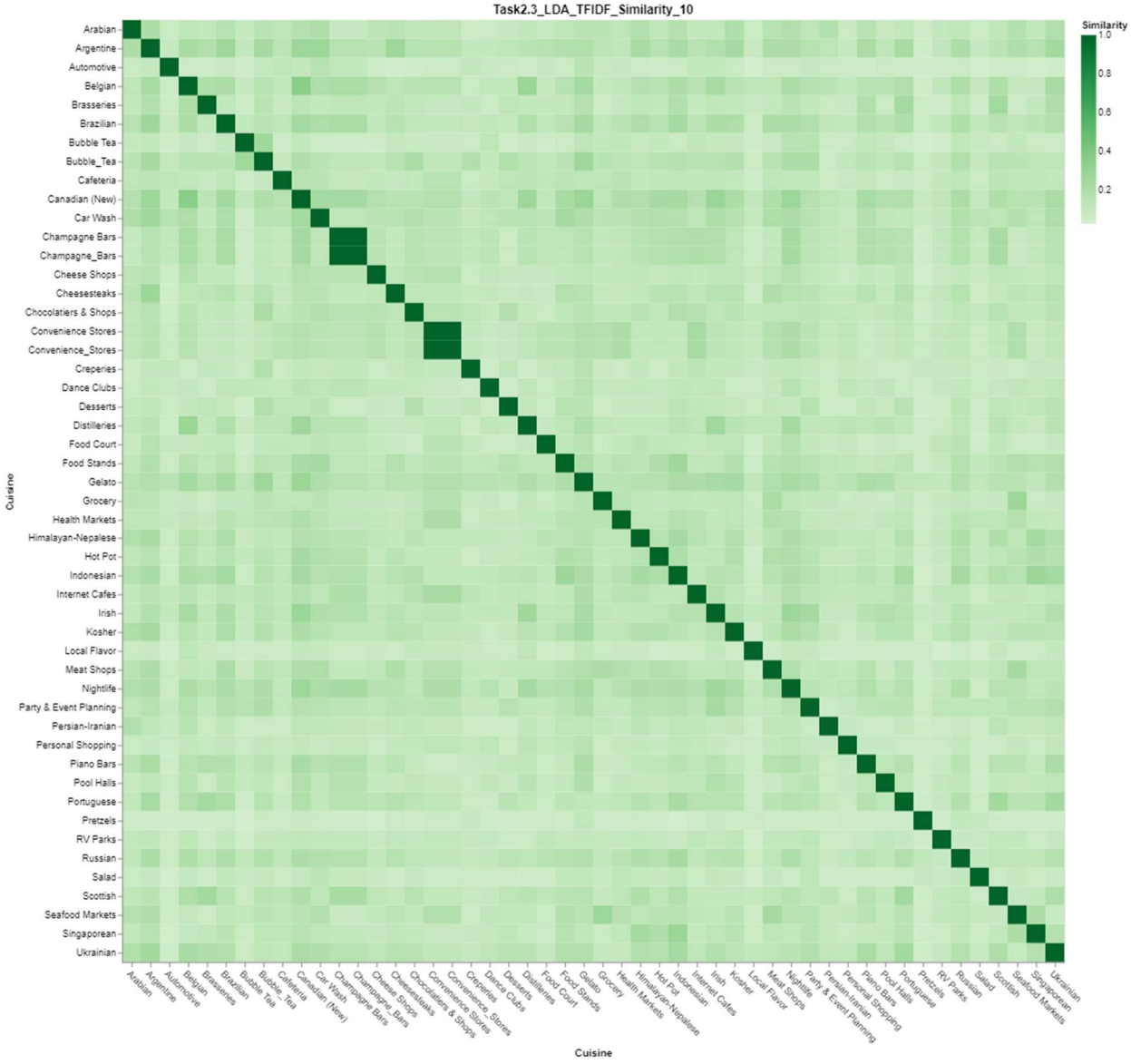
Topic similarity based on LDA and TF-IDF



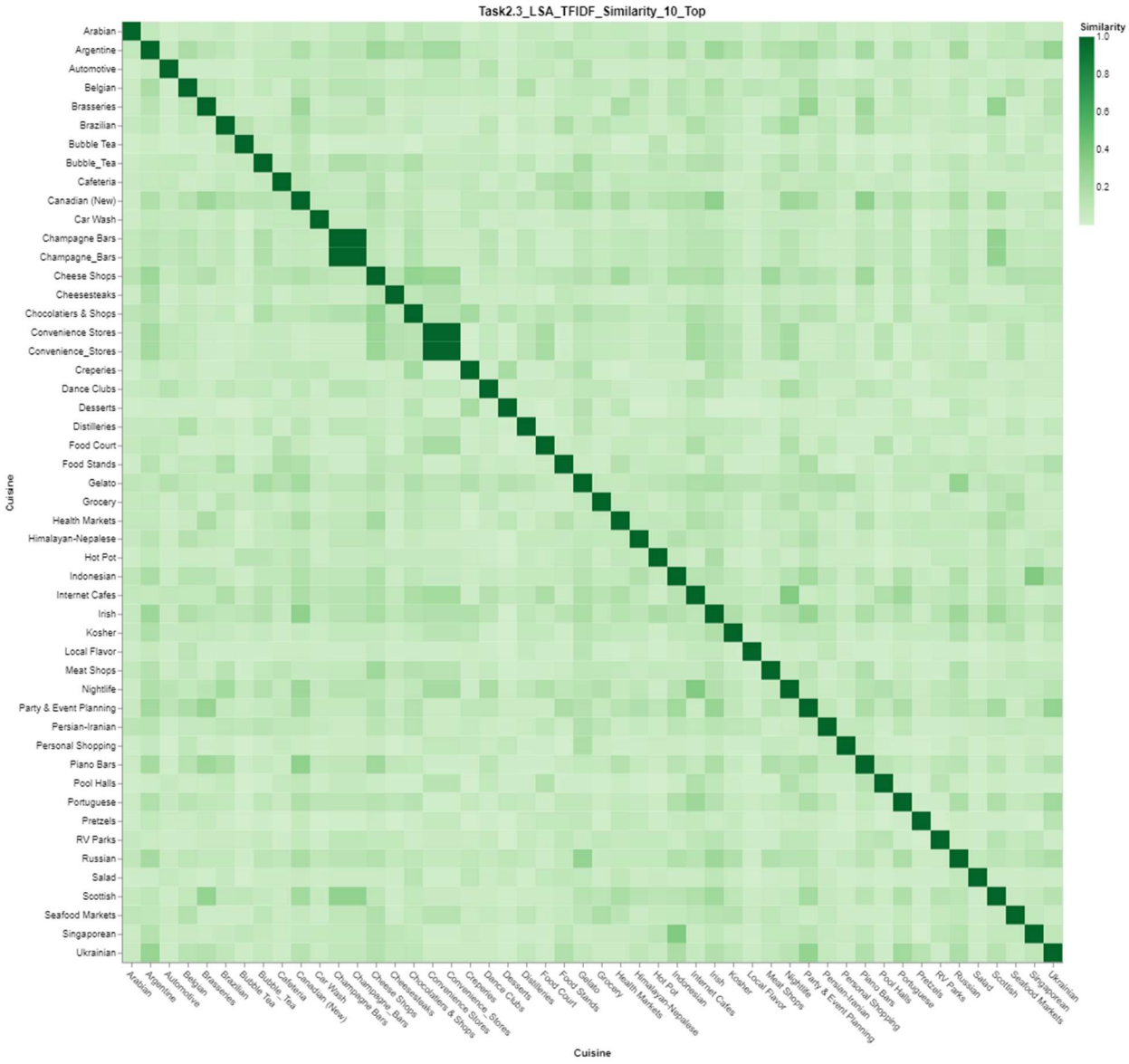
Task2.3: Incorporating Clustering in Cuisine Map

To compare functions and results of different clustering algorithms and different clustering numbers. I use LDA and LSA in this section. This section has two parts. In part 1, I use LDA and LSA to generate 10 topic clusters and visualize the similarities between all categories. In part 2, I generate 15 topic clusters.

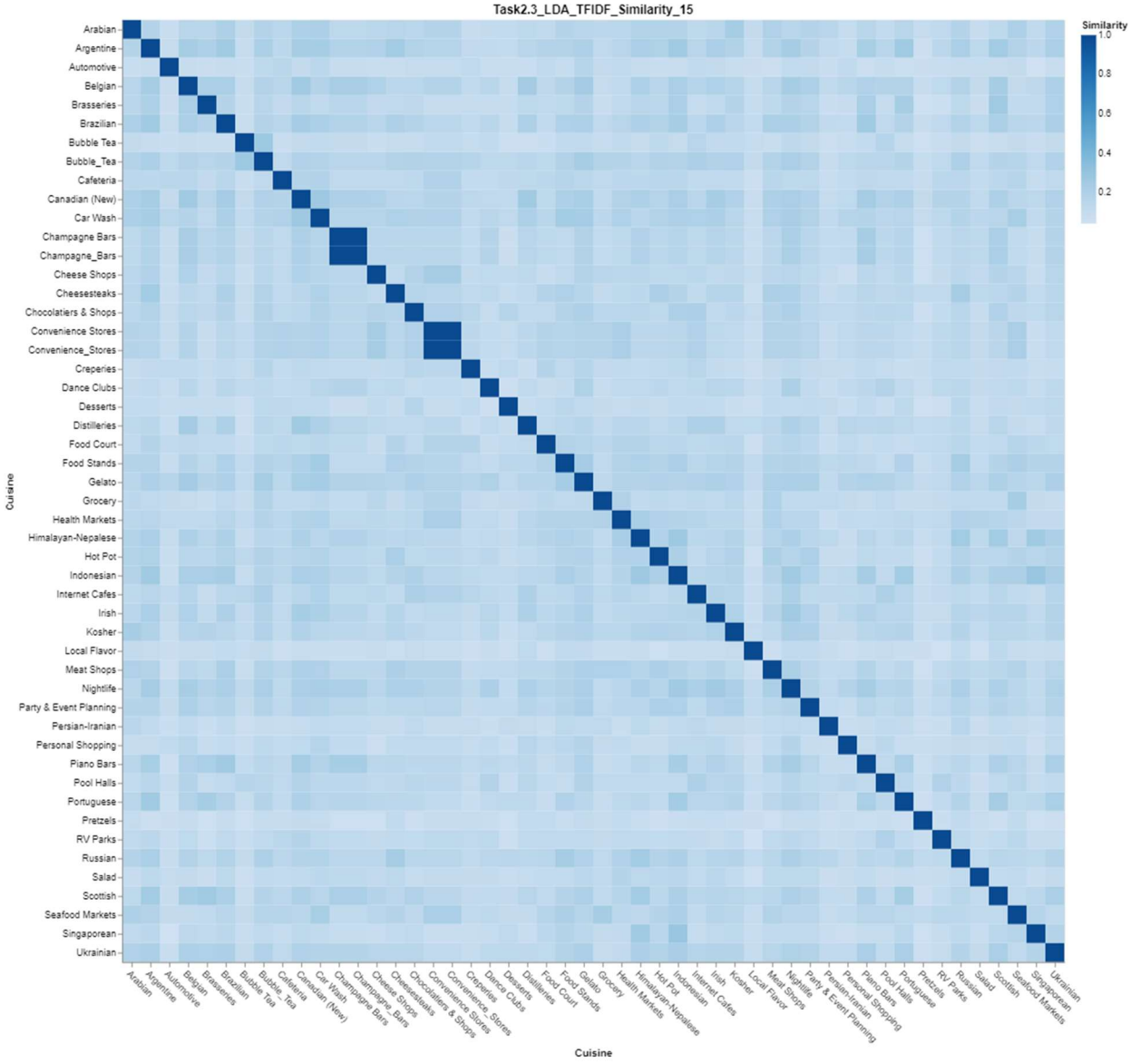
Part 1: 10 topics similarity using LDA



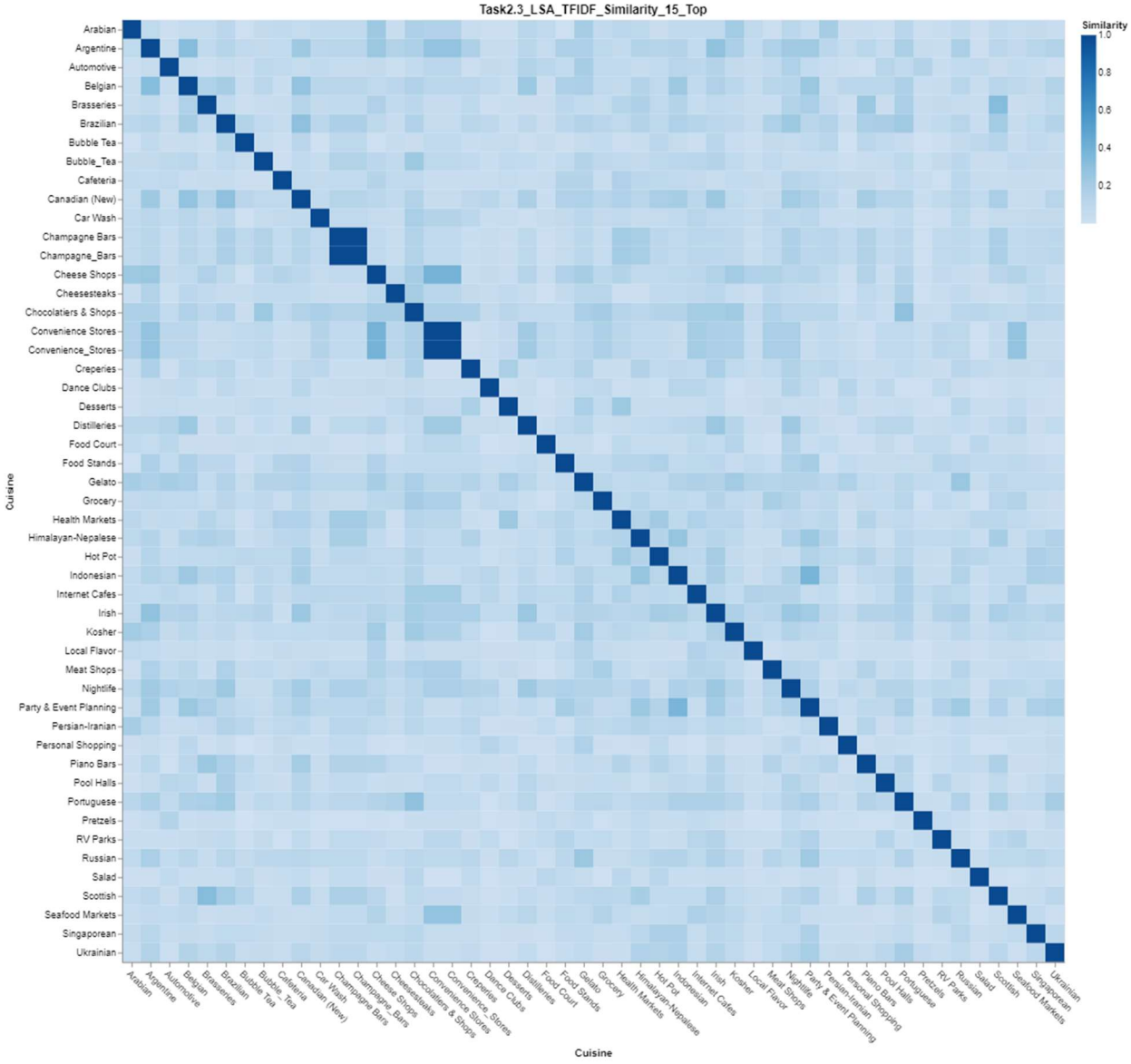
Part 1: 10 topics similarity using LSA



Part 1: 15 topics similarity using LDA



Part 1: 15 topics similarity using LSA



Conclusion

From the results and visualization from previous tasks. We can find that with the same text representation, The color of the visualization using LDA is darker than the visualization using LSA. LDA has a better performance than LSA in this cuisine topic clustering.