

Visualisation of the Cuisine Map

AUTHOR : HAMORA HADI

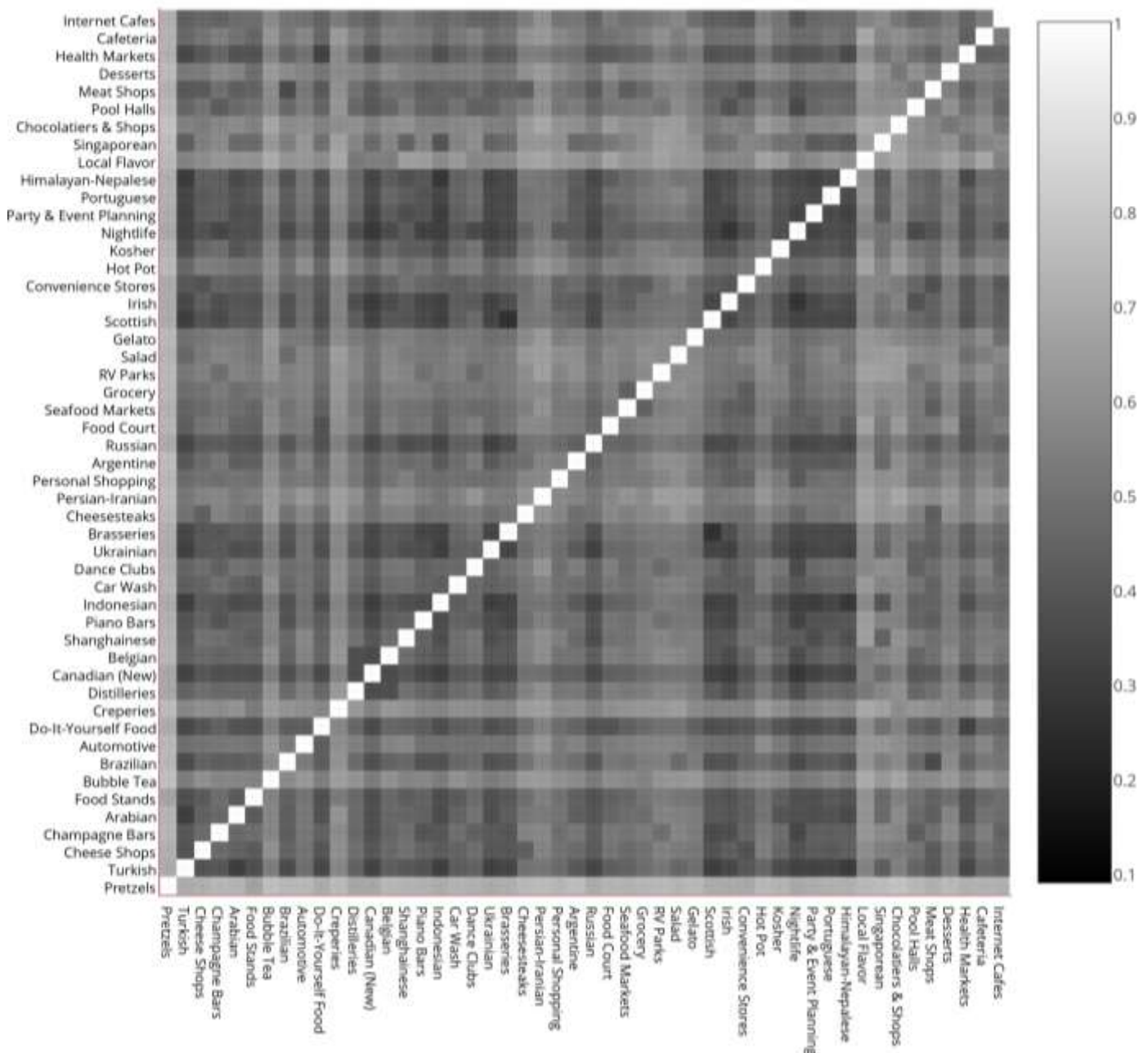
For this task, I randomly picked 50 categories from the given category dataset

1. Preprocess each representation as below:

1. Tokenizing representation with `nltk.tokenize`
2. Removing stop words with `nltk.corpus.stopwords`
3. Removing punctuation such as “, !@#\$~”, etc
4. Removing low frequency (count == 1) words
5. Stemming using `nltk.stem.LancasterStemmer`

2. Calculate the word count of each category representation from step 1

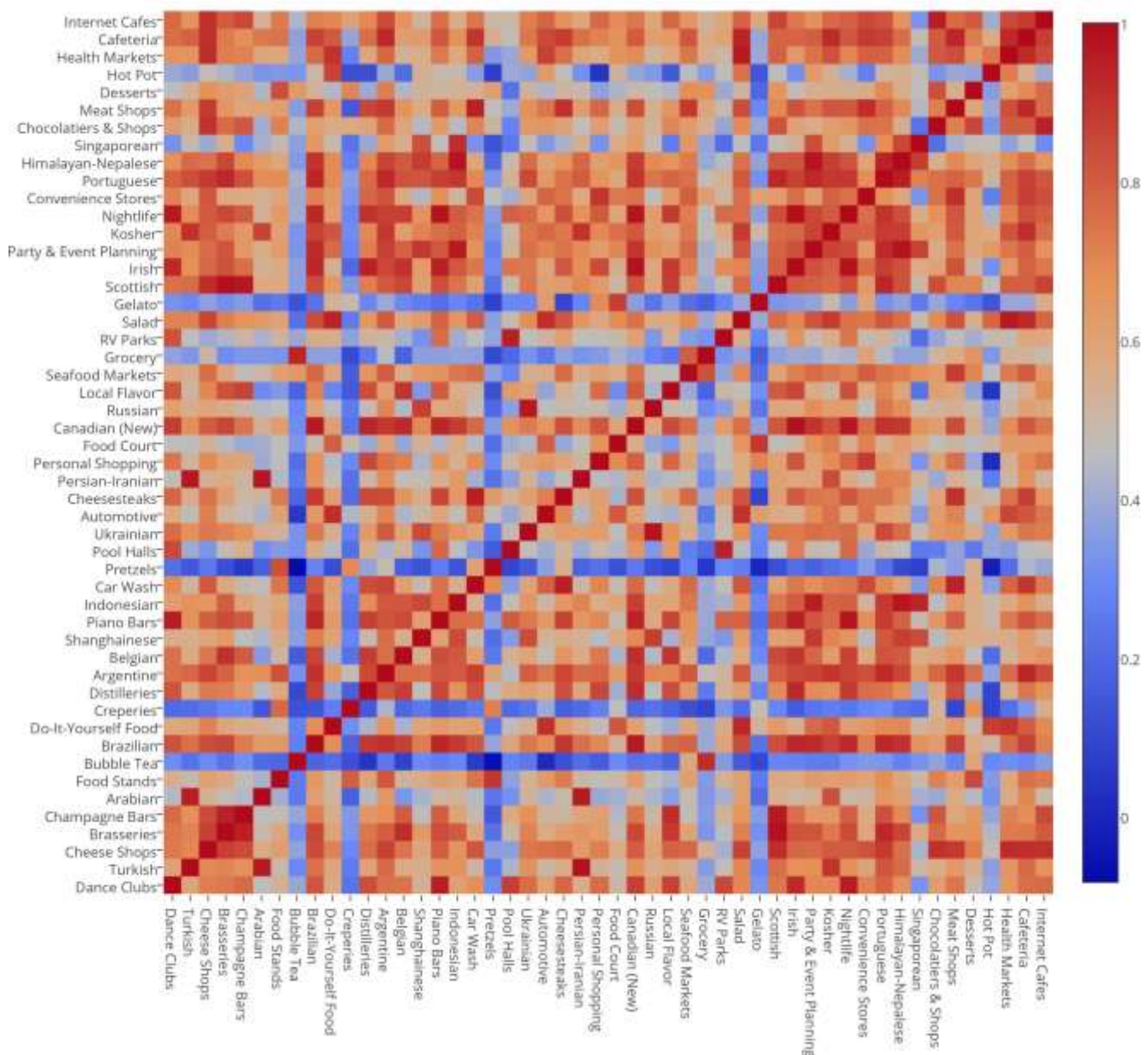
3. Calculate the cosine distance between each two of them without tfidf.

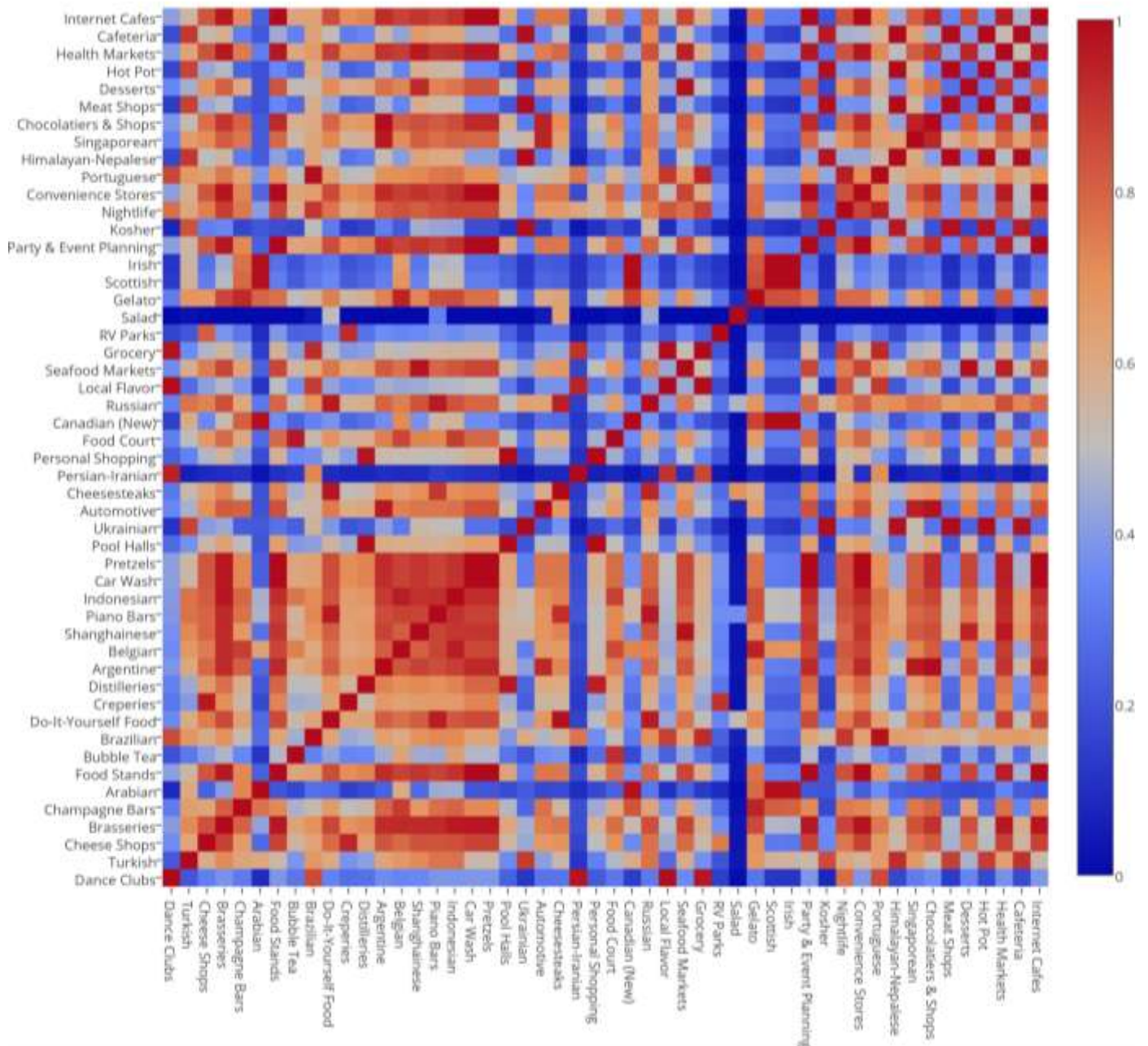


Improving the Cuisine Map

For the second task, I used gensim and nltk with below steps:

1. Get all reviews related to specific category as text representation of that category
2. The same preprocess step as task 2.1.1
3. Train the preprocessed reviews of step2 with LDA and LSI model respectively.
 1. Firstly, reviews of a specific category is processed by TFIDF
 2. By training with LDA/LSI model with 10 topics to get the dictionary, MatrixSimilarity index and model instance
 3. Calculate the similarity between each two of the cuisines based on the result in previous step.

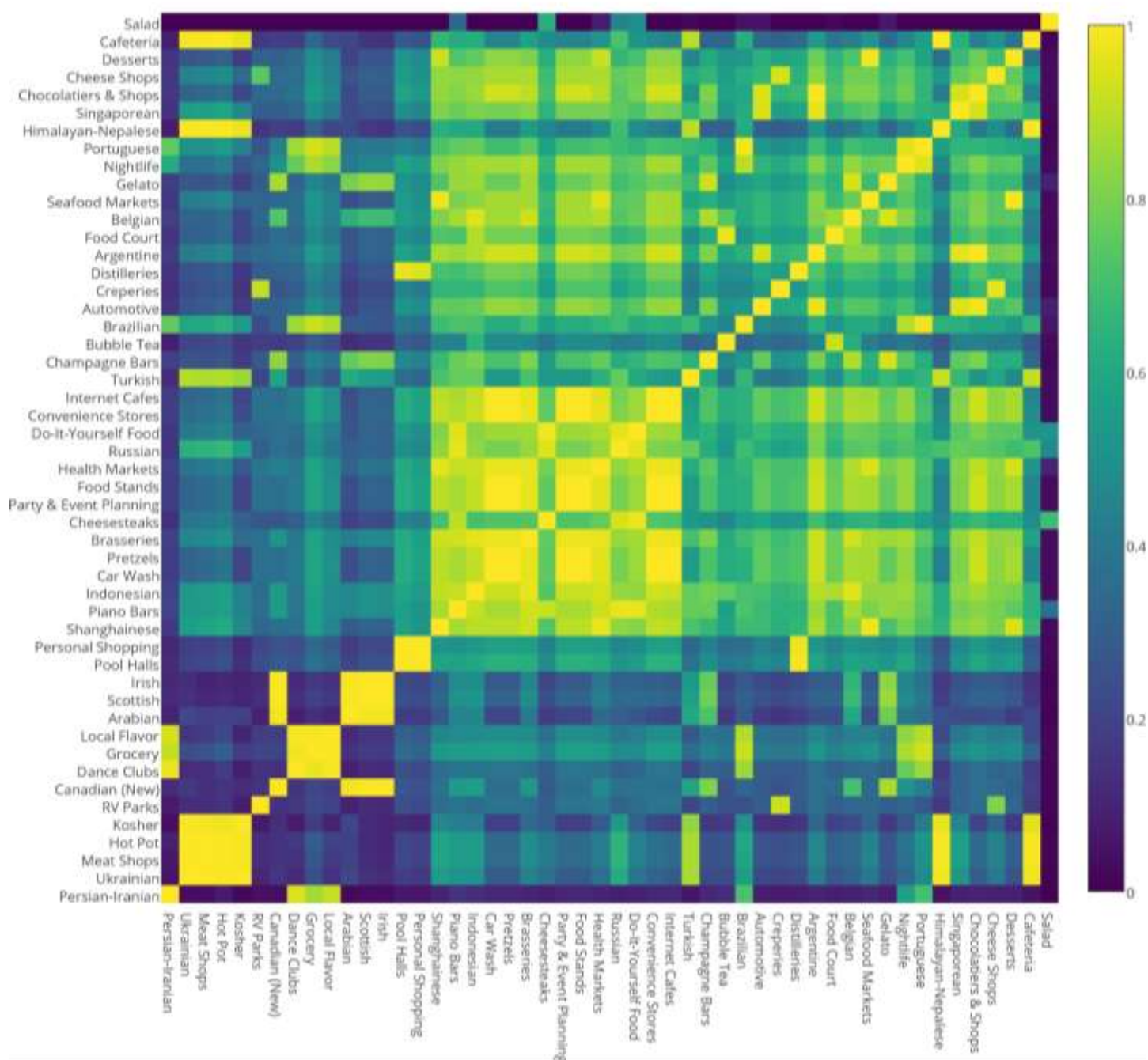




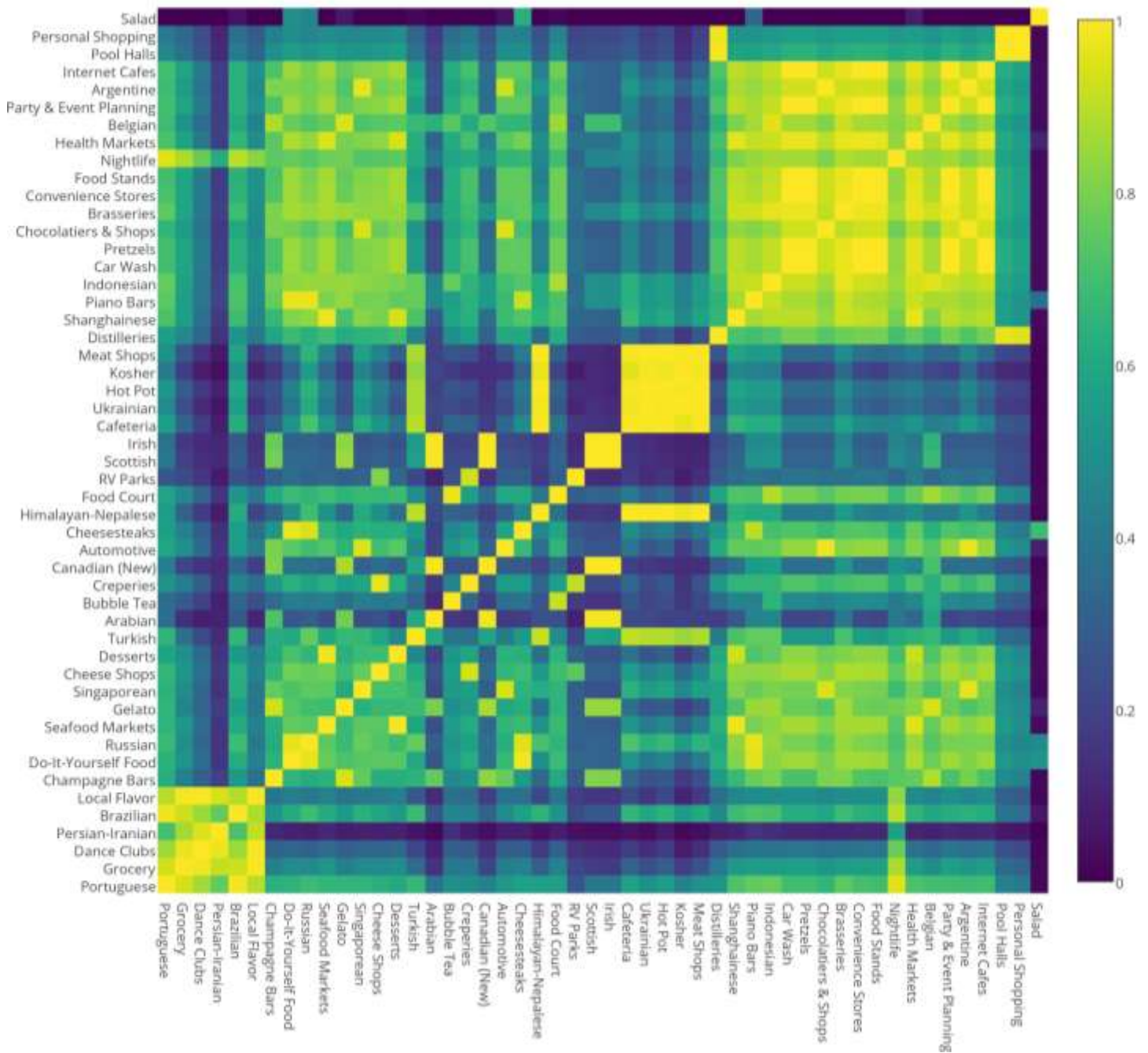
From the graphs shown above we can see two trends:

1. LSI model similarity is close to cosine distance similarity in task1, whereas the LDA model shows a somewhat different similarity.
2. LDA model results in a more clustered style where a bunch of cuisines can be seen similar, LSI model is not so obvious.

I applied the same processing step as task 2.2 for text representation of categories. Since the similarity of each two categories are available, I grouped them according to mutual similarity by moving similar categories closer to each other.



Then I tried to visualise cluster difference based on LDA model with different cluster numbers.



From the above results we can see that with the same text representation, LDA provides a better result than LSI in terms of similarity cluster, the cluster is more identifiable using LDA model.

Just this, Thank you 😊