Task 2 Overview

Overview

By now you should have already familiarized yourself with the Yelp data. In this task, you will work on mining this data set to discover knowledge about cuisines. In the Yelp data set, businesses are tagged with categories. For example, the category "restaurant" identifies all the restaurants. Specific restaurants are also tagged with cuisines (e.g., "Indian" or "Italian"). This provides an opportunity to aggregate all the information about a particular cuisine and obtain an enriched representation of a cuisine using, for example, review text for all the restaurants of a particular cuisine. Such a representation can then be exploited to assess the similarity between two cuisines, which further enables clustering of cuisines.

The goal of this task is to mine the data set to 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. You can see a sample set of reviews from all the restaurants for a cuisine, but you are strongly encouraged to experiment with your own set of cuisines if you have time.

Instructions

Some questions to consider when building the cuisine map are the following:

1. What's the best way of representing a cuisine? If all we know about a cuisine is just the name, then there is nothing we can do. However, if we can associate a cuisine with the restaurants offering the cuisine and all the information about the restaurants, particularly reviews, then we will have a basis to characterize cuisines and assess their similarity. Since review text contains a lot of useful

information about a cuisine, a natural question is: what's the best way to represent a cuisine with review text data? Are some words more important in representing a cuisine than others?

2. What's the best way of computing the similarity of two cuisines? Assuming that two cuisines can each be represented by their corresponding reviews, how should we compute their similarity?
3. What's the best way of clustering cuisines? Clustering of cuisines can help reveal major categories of cuisines. How would the number of clusters impact the utility of your results for understanding cuisine categories? How does a clustering algorithm affect the visualization of the cuisine map?
4. Is your cuisine map actually useful to at least some people? In what way? If it's not useful, how might you be able to improve it to make it more useful?

Note that most of these questions are open questions that nobody really has a good answer to, but they are practically important questions to address. Thus, by working on this task, you are really working on a frontier research topic in data mining. Your goal in this task is to do a preliminary exploration of these questions and help provide preliminary answers to them. You can address such questions by analyzing the visualization of the cuisine map and comparing the results of alternative ways of mining the data to assess which strategy seems to work better for what purpose. You are encouraged to think creatively about how to quantitatively evaluate clustering results. For example, you can consider separating all the reviews about one cuisine (e.g., Indian) into multiple disjoint subsets (e.g., Indian1, Indian2, and Indian3) and thus artificially create multiple separate cuisines that are known to be of the same category. You can then test your algorithm on such an artificial data set to see if it can really group these artificial subcategories of the same cuisine together or give them very high similarity values.

To receive credit, you must complete the following specific tasks. It would be wise to first attempt to finish these required minimum tasks before exploring other ideas that you might have, especially if you have only limited time.

Task 2.1: Visualization of the Cuisine Map

Use all the reviews of restaurants of each cuisine to represent that cuisine and compute the similarity of cuisines based on the similarity of their corresponding text representations. Visualize the similarities of the cuisines and describe your visualization.

The visualization shows the similarity matrix, with every cell corresponding to the similarity between two cuisines. The opacity of each cell is the similarity - with a higher opacity for a higher similarity.

Image 1 - Similarity Matrix from noIDF

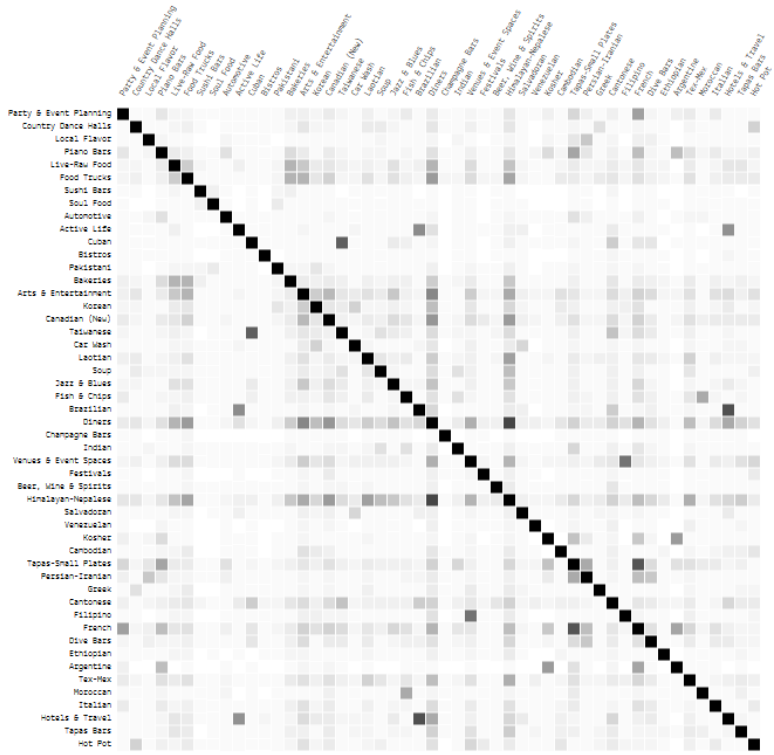
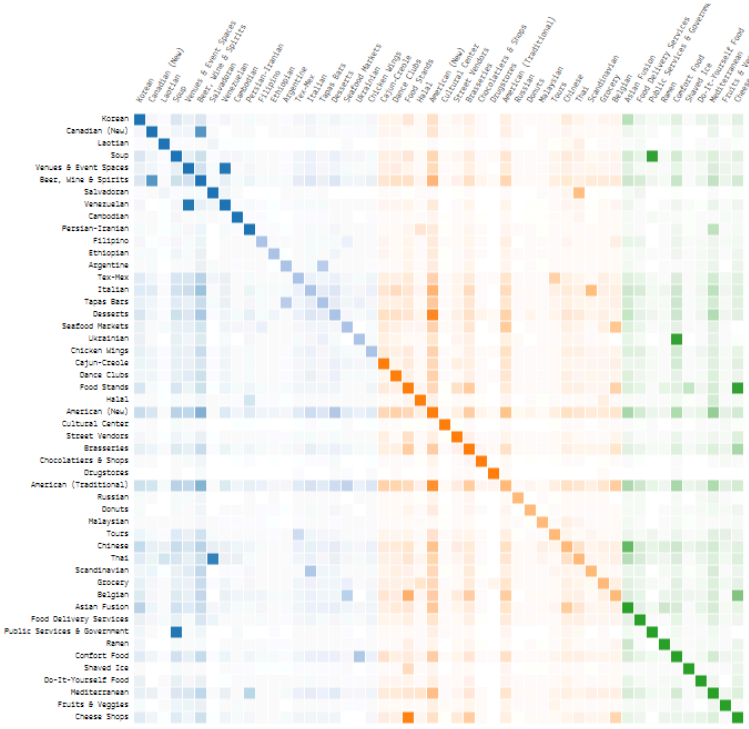


Image 2 - Similarity Matrix from IDF (Colors only for better distinction.)

Image 3 - Similarity Matrix from LDA (Colors only for better distinction.)



Task 2.3: Incorporating Clustering in Cuisine Map

Use any similarity results from Task 2.1 or Task 2.2 to do clustering. Visualize the clustering results to show the major categories of cuisines. Vary the number of clusters to try at least two very different numbers of clusters, and discuss how this affects the quality or usefulness of the map. Use multiple clustering algorithms for this task. Also note in that each color is a separate cluster in the sample images below.

Image 4 - Similarity Matrix from IDF

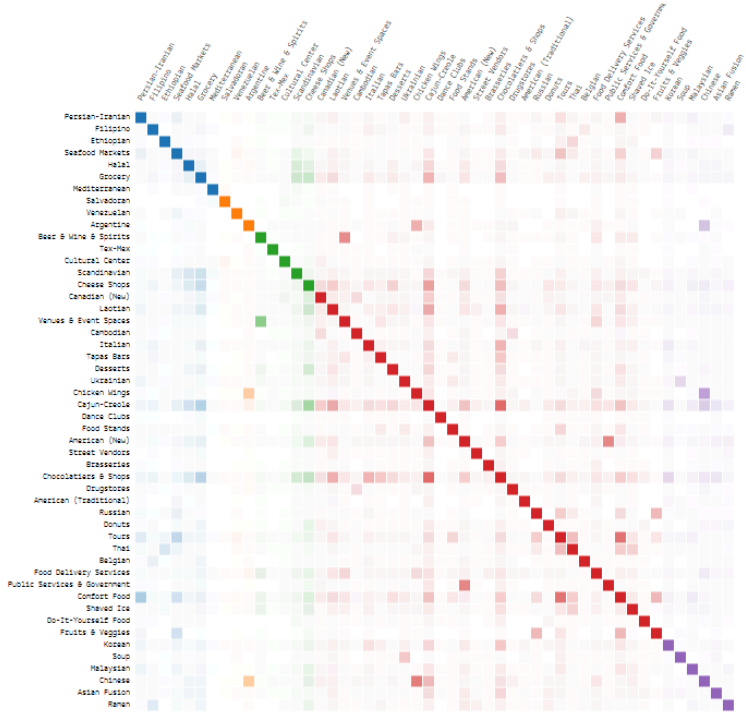


Image 5 - Similarity Matrix from noIDF

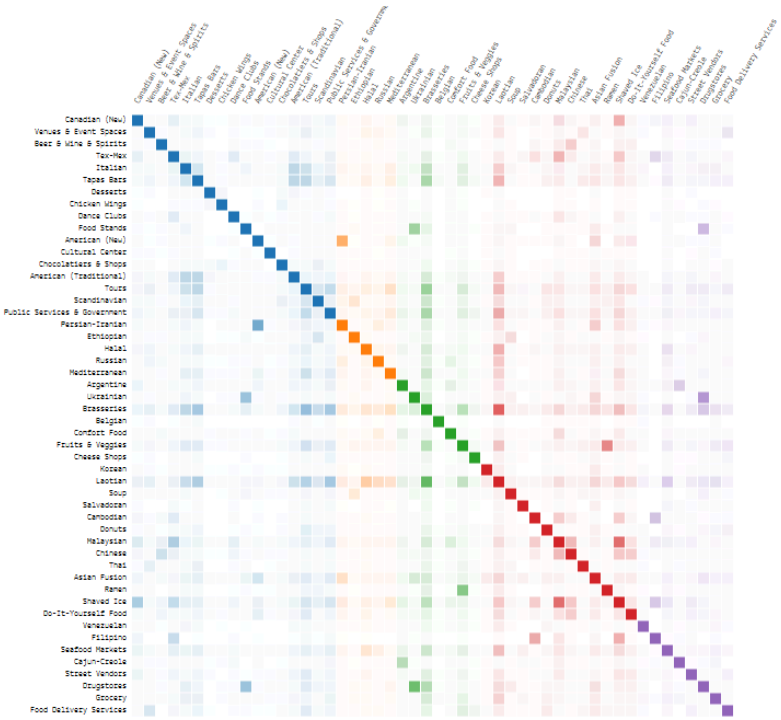


Image 6 - Similarity Matrix from LDA

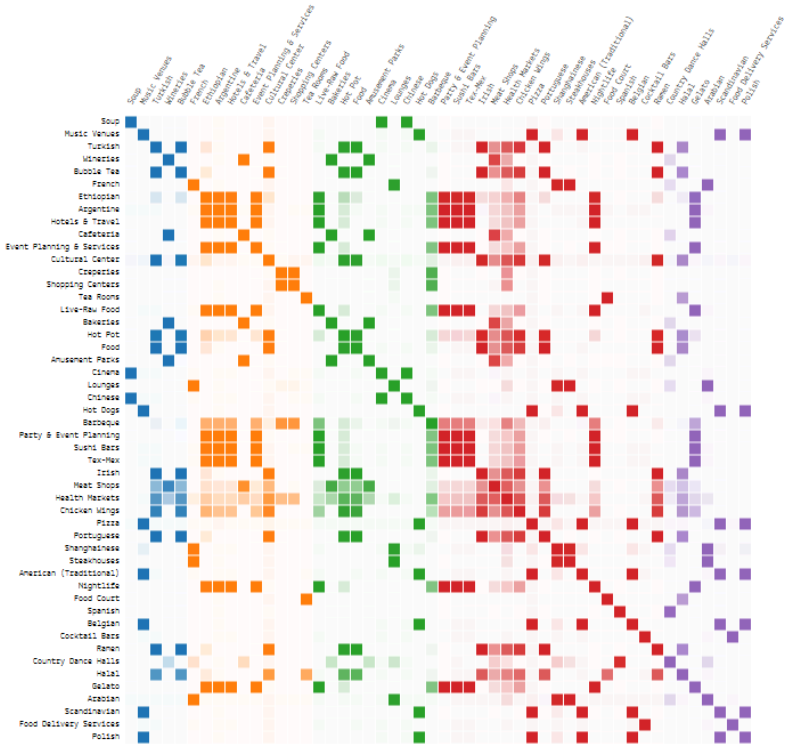


Image 7 - Similarity Matrix from LDA, with clustering and a Topic 1

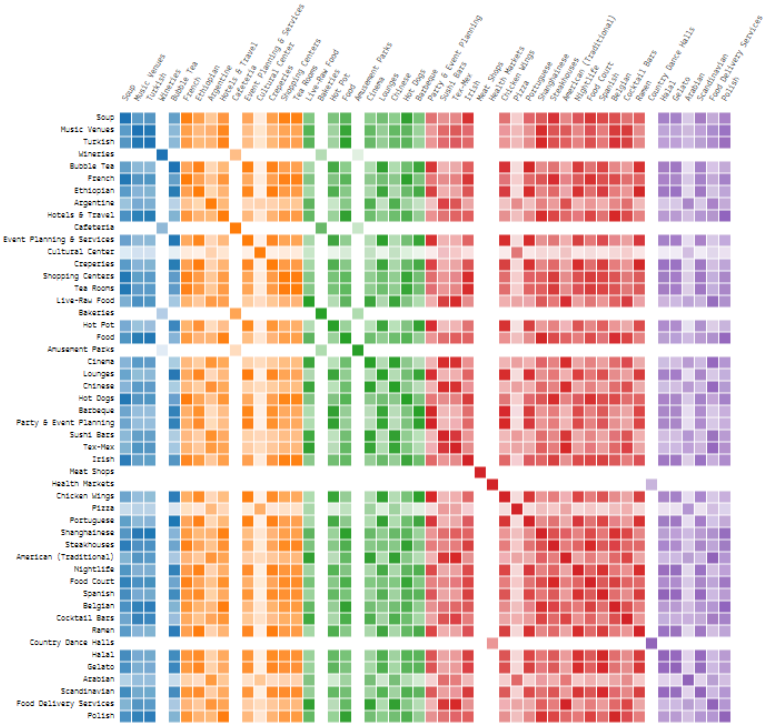
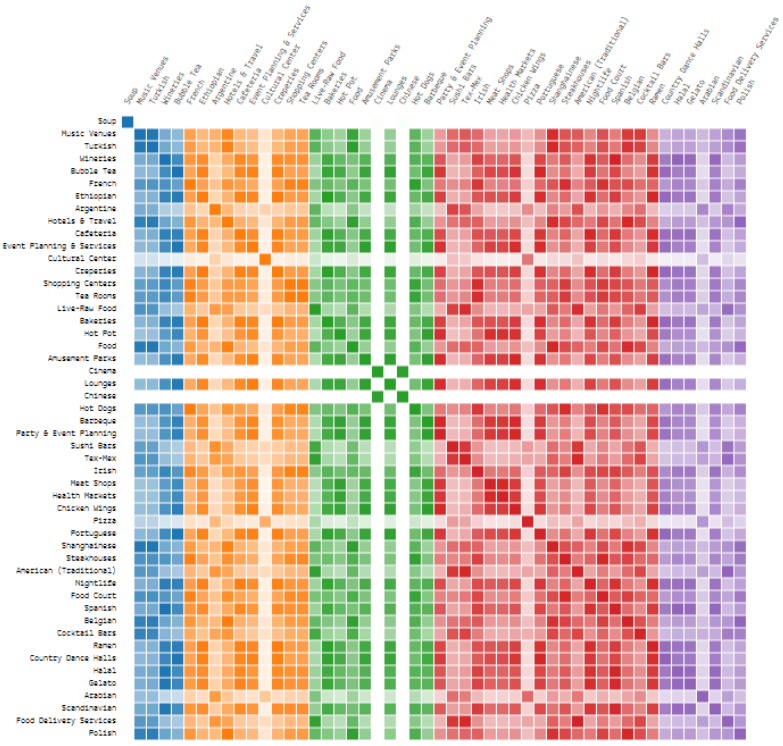


Image 8 - Similarity Matrix from LDA, with clustering and a Topic 2



Submission

You must submit a report in PDF format. We suggest that it be 2-3 pages long. Your report will need to include the following elements.

1. A written portion that provides sufficient detail for others to reproduce the process, including
- A description of the visualization used in Task 2.1
- An in-depth analysis/description about the visualization improvements used in Task 2.2
- A description of the multiple clustering algorithms used in Task 2.3 and how varying the number of clusters affects the quality or usefulness of the map
- The parameters used, how you applied the algorithms to the data set, how you evaluated the clustering results, what tool was used for visualization, etc.

- Your opinions about whether the results you generated make sense or are useful in any way. Are there any particular aspects of your visualization to which you would like to bring attention? What do you think the results and your visualization show?

2. Your visualization results, including:

- Visualization of cuisine map
- Improving the cuisine map – varying text representation and varying similarity function
- Incorporating clustering in cuisine map