A Proposal for Visualization of Networks Based on Topic Models Using Restaurant Reviews

Basic Info

- Project title: Visualization of Networks Based on Topic Models Using Restaurant Reviews
- Team members:
  - Sooraj Subrahmannian (smsubrahmannian@dons.usfca.edu)
  - Kunal Kotian (kkotian@dons.usfca.edu)
- Github repo: https://github.com/kunal-kotian/visualizing_review_networks

Background and Motivation

We are interested in networks, and hence decided at the outset that we would choose a problem that lends itself to network analysis. Additionally, both teammates have some experience working with Latent Dirichlet Allocation (LDA) for extracting ‘topics’ from a text corpus. Hence, we decided that visualizing networks constructed from the outcomes of topic models will be a great way to merge our two interests.

Furthermore, we (both team members) have also previously collaborated on a project involving building a recommendation system using restaurant reviews from Yelp in the past. Hence, being familiar with the Yelp reviews’ dataset, we decided to use it for this data visualization project as well. This should lower the risk spending excessive amounts of time handling unforeseen issues with the data.

Project Objectives

We are going to work with Yelp reviews for Las Vegas restaurants.
We aim to address the following questions:

- Objective 1: Visualize a network of restaurants
  For every restaurant, we can extract ‘topics’ that emerge from its reviews. What does a network of restaurants constructed using similarities between its ‘topics’ look like?
- Objective 2: Visualize a network of topics
  Using the topics that emerge from all restaurant reviews, we can construct a network of topics using a similarity measure. What does this network of topics look like?
- Objective 3: (Optional) Enable interactive exploration of network clusters
  The structure of the networks of restaurants and topics is dependent on the way the nodes (restaurants or topics) are linked to each other. This node linkage is controlled by a ‘threshold’ for similarity that can be tuned. We will add an interactive feature to the plots allowing users to tune the linkage similarity threshold and visualize changes in the networks’ configuration.

Data

We are planning to use Yelp reviews dataset. This dataset is readily available on the Yelp website. Please find the dataset here: https://www.yelp.com/dataset

Data Processing

The Yelp review dataset has the following characteristics such as

- 6 GB of curated data
- JSON format
- 7 million ratings & tips of 150,000 businesses

Data Cleanup

We do not expect to require substantial data cleanup. Keeping the size of the dataset in mind, we have filtered the data down to restaurants in Las Vegas.

Quantities to be Derived from the Data

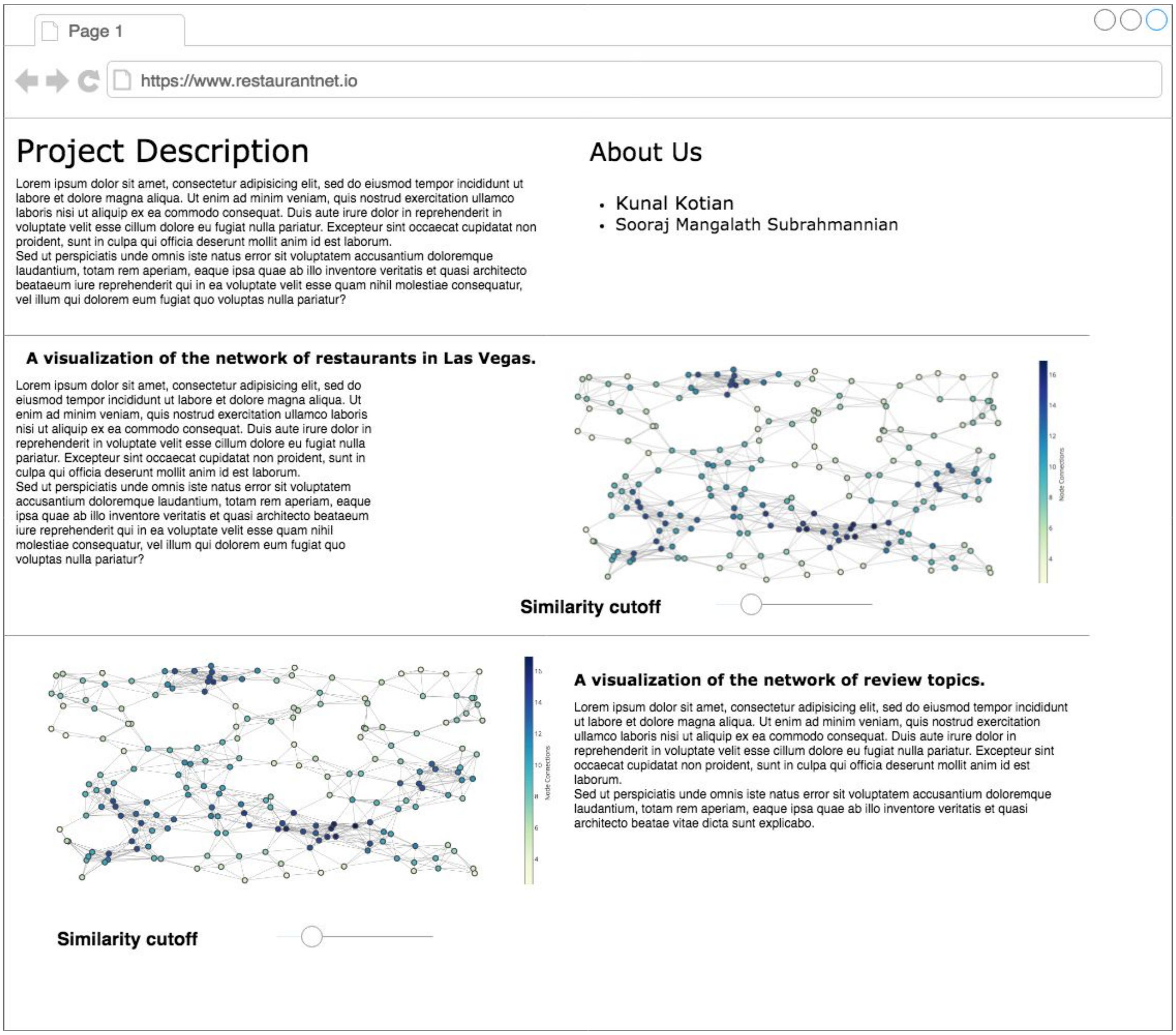
1. A trained LDA model
2. A topic-term matrix and its Jensen-Shannon distance
3. A document topic matrix and its Jensen-Shannon distance

Implementation of Data Processing

We started preprocessing the dataset a while back for a separate project. Each review in the filtered dataset was processed using Python packages SpaCy and regex to get a final of tokens. The final tokenized documents i.e, are currently being monitored for further processing to account for our new objectives.

Visualization Design

Design 1: Single page, compact format (less vertical scrolling)



Project Description

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beataeum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

About Us

- Kunal Kotian
- Sooraj Mangalath Subrahmannian

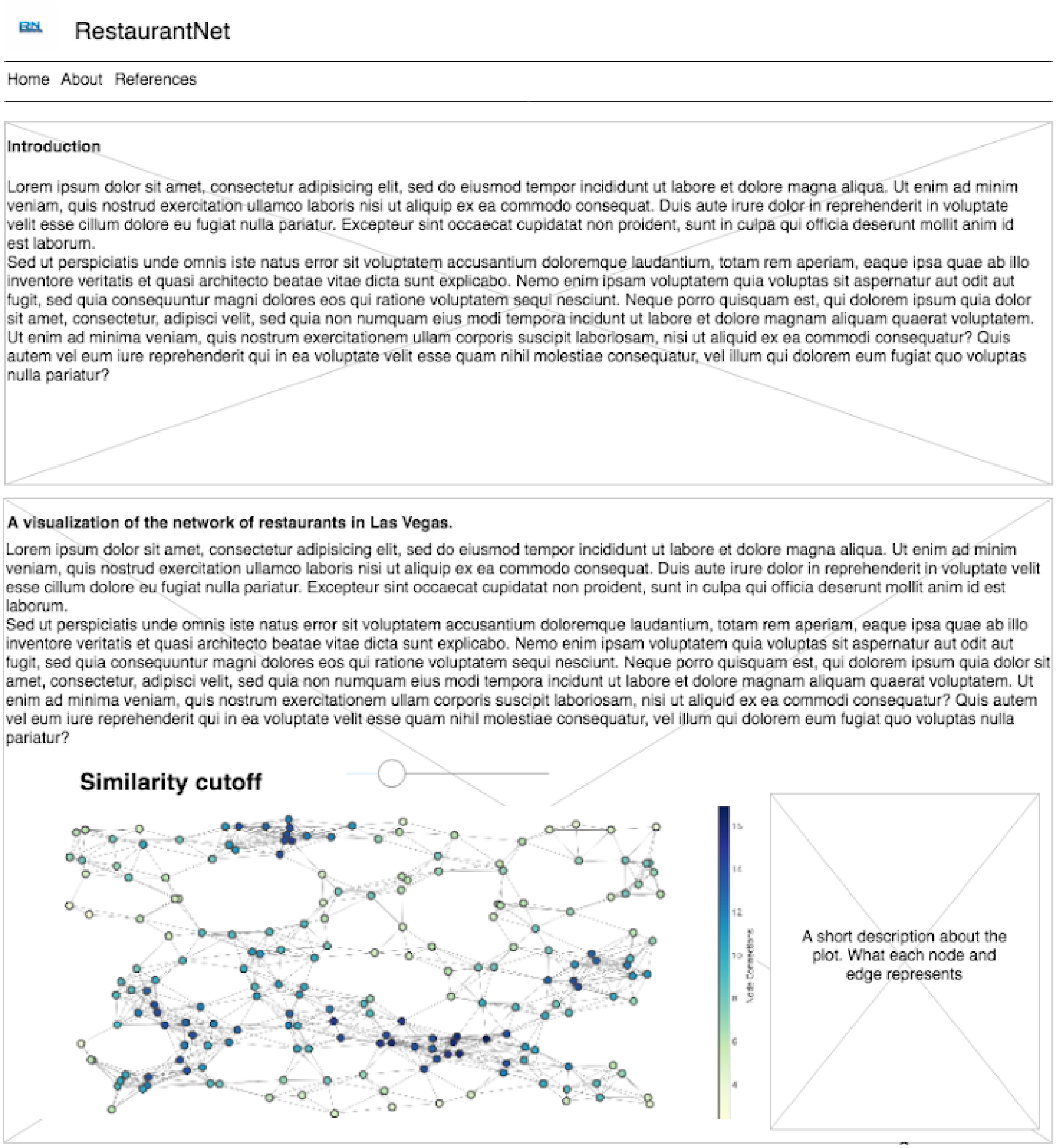
A visualization of the network of restaurants in Las Vegas.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beataeum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

A visualization of the network of review topics.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo.

Design 2: Single page, story format (significant vertical scrolling) - Part 1



Introduction

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

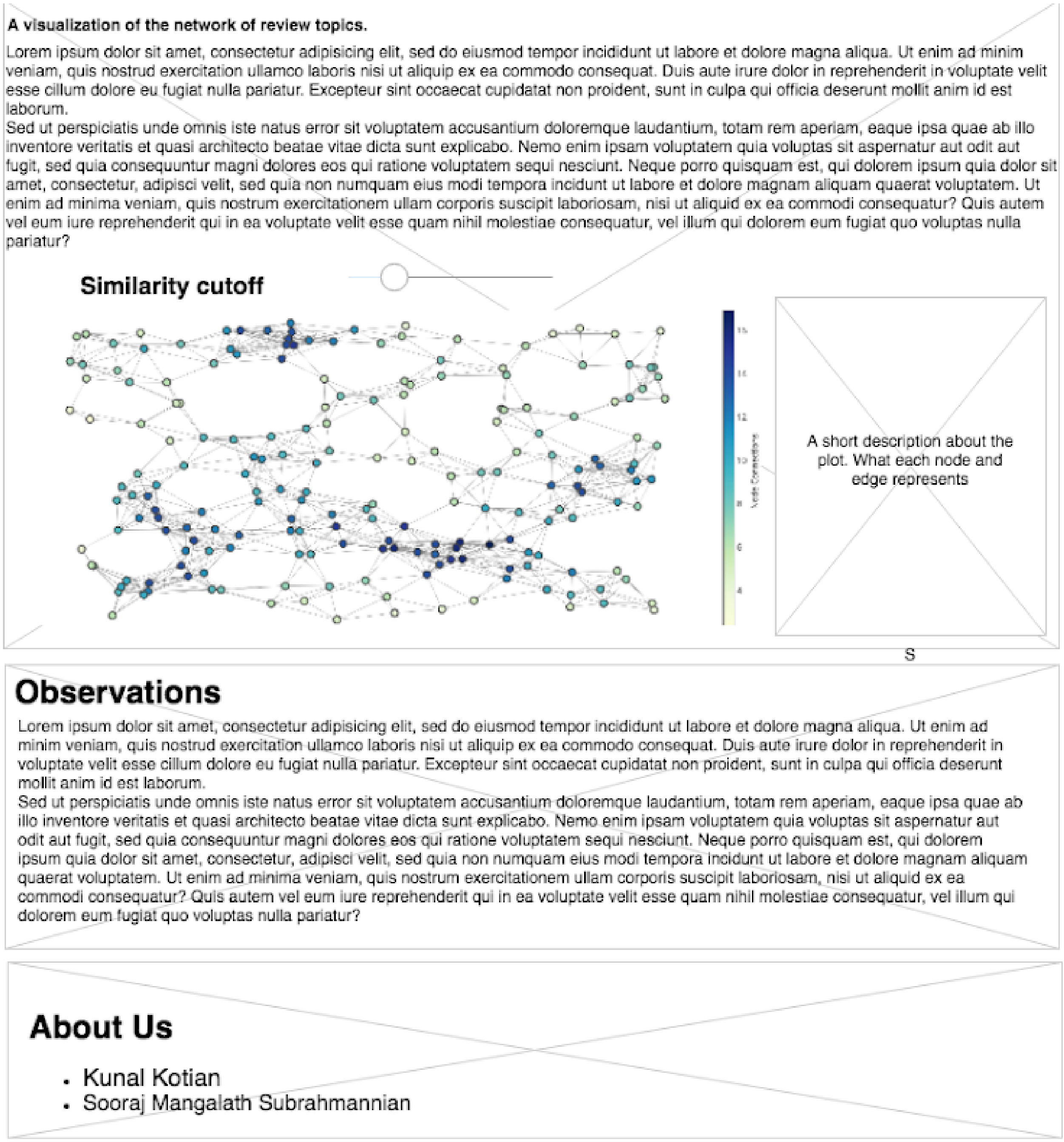
A visualization of the network of restaurants in Las Vegas.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

A short description about the plot. What each node and edge represents

Design 2: Single page, story format (significant vertical scrolling) - Part 2



A visualization of the network of review topics.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

A short description about the plot. What each node and edge represents

S

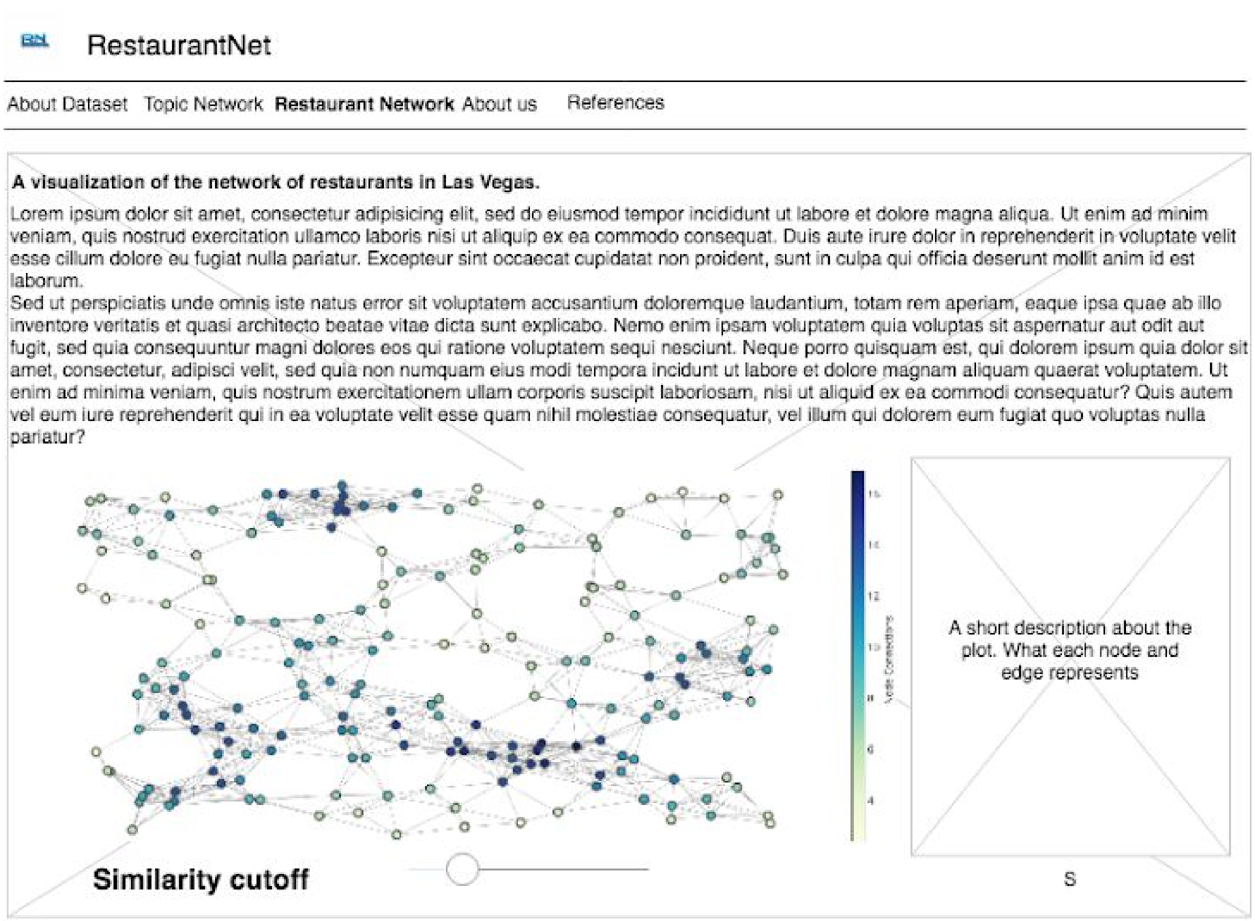
Observations

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

About Us

- Kunal Kotian
- Sooraj Mangalath Subrahmannian

Design 3: Multiple webpages, one major element per webpage



RestaurantNet

About Dataset Topic Network **Restaurant Network** About us References

A visualization of the network of restaurants in Las Vegas.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

A short description about the plot. What each node and edge represents

Out of the 3 designs, we plan to choose the **Design 2**. This is because we suspect that having all the content on a single page such that it can be accessed easily by simply scrolling in the vertical direction, would make it easier for a reader to digest the information shown.

Must-Have Features

- Feature 1: A visualization of the network of restaurants in Las Vegas.
  We will train an LDA model and use it to determine the ‘topic vector’ for each restaurant. Using this information, we will calculate the Jensen-Shannon distance between each pair of restaurants, and then apply a similarity threshold/cutoff to form linkages between the restaurant nodes. For the minimum viable feature scope, we will choose a pre-tuned value for the similarity threshold and display the resulting network on our website.
- Feature 2: A visualization of the network of review topics.
  Using the trained LDA model mentioned above, we will also obtain a collection of topic vectors for the entire Las Vegas restaurants review text corpus. Similar to Feature 1, we will again calculate pairwise Jensen-Shannon distances between topics and then apply a pre-tuned similarity cutoff to form linkages between nodes and display the resulting network on our website.

Optional Features

- Feature 3:
  Features 1 and 2 refer to network plots constructed using pre-tuned similarity thresholds. For our optional feature, we would like to add a slider that allows users to change the value of the similarity threshold and view the reconfigured network structure on the fly. We would restrict the slider to only take on fixed threshold values at regular intervals. We would have to precompute the networks for every possible value of the threshold that can be set using the slider, and then display the corresponding network in response to movement of the slider.

Project Schedule

Week	**Goal**
Apr 15 - Apr 21	Complete the proposal, create a basic webpage, train the LDA model, finish the data processing tasks, and make one network plot
Apr 22 - Apr 28	Fine-tune visualizations, work on adding the optional interactive slider, write supporting content for the website
Apr 29 - May 5	Prepare a project presentation
May 6 - May 12	Make changes based on feedback received after presentation and submit the final project