City Genres

Group B
Jintong Yu jy3246@columbia.edu
Xianfeng Jiang xj2292@columbia.edu

I. Project Proposal

1. Description

Based on a list of 180 typical travel destinations suggested by Priceline.com, we are interested in finding out commonalities of these destinations and what attributes make some of them more similar than others from a traveler's perspective. To provide evidence of attributes for these destinations, we will collect random facts in terms of travel activities and other descriptive features using various online sources, and scape documents from reliable websites. Then, we will use hierarchical clustering to find subgroups within these destinations. Making use of the collected data, we expect to explore the question from the perspective of destinations' attributes recorded in a tabular format, destinations' general description derived from scraped documents, and combination of the former two perspectives.

- Tabular Attributes: What are the most common attributes and travel activities among the selected travel destinations? Are there more metropolis or destinations with more natural scenes? How would you divide these destinations into some "marketable" groups?

- General Description: What attributes extract from the documents are the most prominent among the selected travel destinations? How many representative attributes are there for defining these destinations into meaningful groups?

- Combination: Overall, what representative attributes (genres) would you choose for dividing these destinations into some "marketable" groups? Describe each group.

2. Data Source

1) TripAdvisor: We will web scrape the "Brief Introduction" section on the main page and the tags of the top 30 attractions by popularity on the "Things to do" page for each destination.
2) Wikipedia: We will also web scrape all paragraphs from the Wikipedia page for each destination.
3) Random facts manually recorded in a table based on information provided by TripAdvisor and Wikipedia.
4) World Tourism Organization: number of visitors

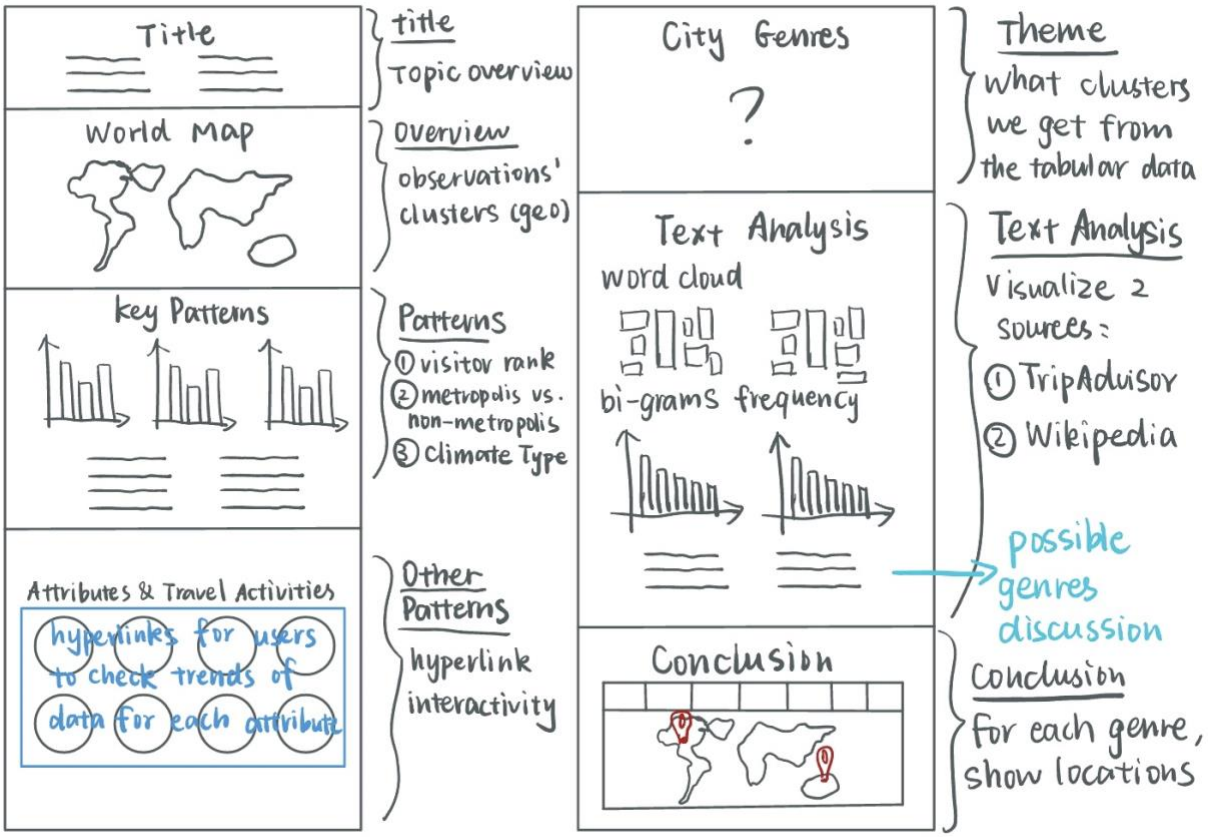
3. Visualizations

- Map: Visualize the geographic locations of the 180 travel destinations on a world map using leaflet. By adding clusters to the map, users could easily identify the geographical similarity among destinations. We will also notify the representative attributes for each destination in its pop-up.
- Scatterplot: Display clusters of data points based on the tabular data and text data.

- Word Cloud: Plot the most common words used to describe these destinations.
- Bar Chart:
  a) Plot the most common bi-grams used to describe these destinations.
  b) Plot the frequency of each defined attribute and travel activity based on the tabular data.
- Interaction: By clicking the button representing a specific group, users could see the corresponding word cloud or bar chart that demonstrates the representative attributes for that group.

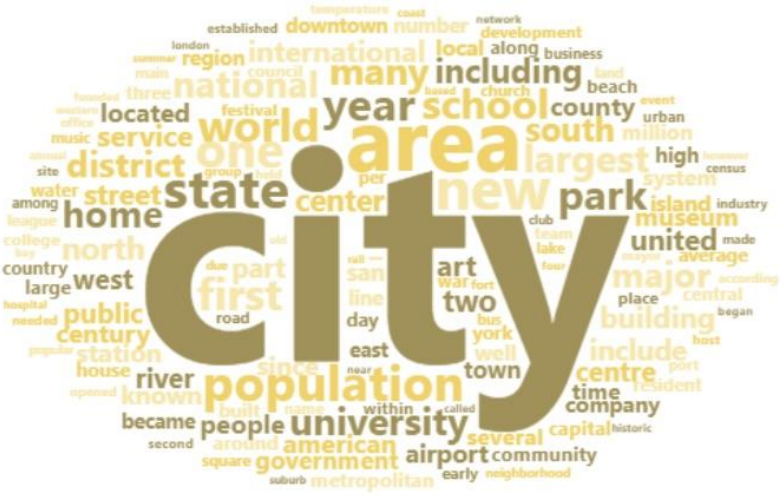
II. Process Documentation

1. Sketches of Final Design



2. Visualization Progress

1) We only keep the TripAdvisor text for creating word cloud.
   For the word cloud section, when we calculate the frequencies for the terms within the scrapped Wikipedia data, the most frequent words do not reveal anything interesting, so we decided to drop the Wikipedia text. The original word cloud of Wikipedia text is shown below:



2) Replace the scatterplot with a density graph

We want to explore the patterns of temperature distribution among these destinations, so we first plot a scatter plot to explore whether there are clusters of values. However, as the scatterplot shown below, the values are just scattering around the graph and the concentration is not obvious. Therefore, we turned to a density graph, aiming to figure out where values are dense. Luckily, we found concentration of values in the range of [25, 50] as shown below.