Yelp Review Sentiment Breakdown

Introduction

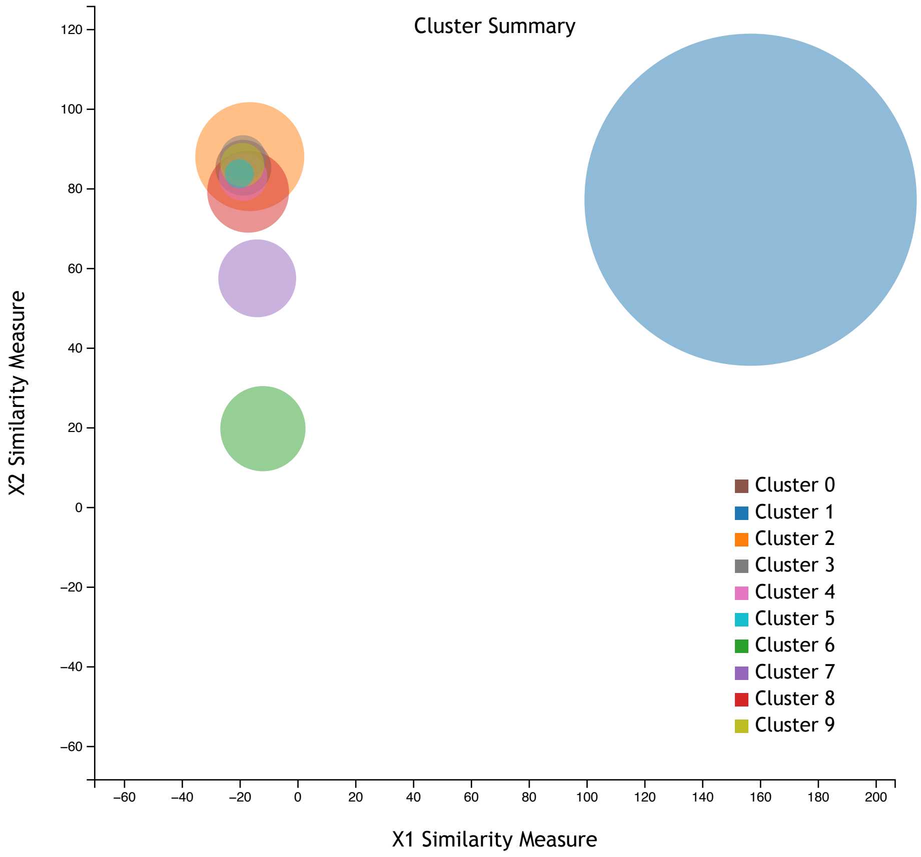
Yelp is an online platform where users can find and share reviews for businesses like restaurants, hotels, and more. It helps people make informed decisions by providing ratings and opinions from others. For this project I used the Yelp Academic Dataset, a large collection of data made available by Yelp for academic research purposes. It contains a wealth of information, including user reviews, business details, user profiles, check-ins, and more. In this project, I seek to break down the general sentiment in a business' reviews. To do this, I vectoried the review data by creating a TF-IDF matrix and used K-Means Clustering to derive groups from those vectors. I also generated a similarity matrix for the vectors using cosine similarity. In the graphs below I will walk you through the results.

Summary of Review Clusters - Table Based

The visual below was created to summarize the data in each review cluster. Each circle represents one cluster, there are 10 total for each business. The radius of the circles corresponds with the relative number of reviews in that cluster -- a bigger circle is a cluster with more reviews. The other main aspect of this visusual is position. I used singular value decomposition to reduce the centroid vectors for each cluster from thousands of dimensions down to just two. The resulting 2D centroids give us an idea for how similar each cluster is relative to its neighbors. Clusters that are overlapping are likely much more similar than clusters that are distant. To get an idea for the sentiment in each cluster, hover your mouse over the corresponding circle and it will show the top 5 words for that cluster. The final aspect of this visual that's worth mentioning is color: used to distinguish between clusters. Opacity was turned down such that each overlapping cluster is visible. Note: Pay little attention to the values on the axes I left them in to give an idea for the difference in scale between the x and y axes. Remember the important factor is not the number itself, but the distance you see between clusters.

Business:

EQ-TZ2eeD_E0BHuvoaeG5Q



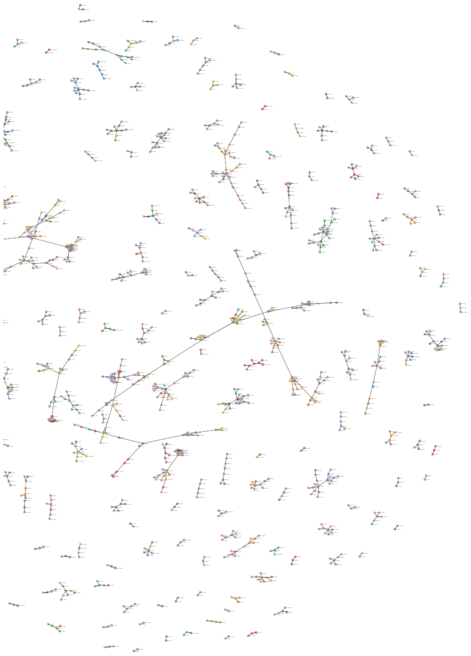
Graph of Reviews - Network Based

This node link graph is generated to visualize a different type of review grouping. Each node is a review, each link is generated between a node and another node with the highest cosine similarity to it. In this way, we've generated a somewhat different form of clusters. The stroke width of each link corresponds to the magnitude of similarity. Each node was colored based on its k-means cluster in order to see if there is continuity between groups generated with k-means vs cosine similarity. I would recommend looking at this visual on a desktop, and to use a mouse when navigating.

Business:

EQ-TZ2eeD_E0BHuvoaeG5Q

Review Similarity Graph



Map of Businesses - Geometry Based

Here, I have included a map to locate each of the yelp businesses. Each business is marked with a red dot. Hover your cursor over the dot to reveal the business name. Note: The locations have been chosen arbitrarily for the time being.

Business Locations