**Capstone Final Report: Recommending Universities Based on Nearby Venues**

by Isaiah Kaufman

**Introduction:**

For my final project, I will be making a dashboard tool that allows you to find universities similar to any town you want to search. Many sites allow you to research schools based on academics and college life and many other factors, but not many tools exist that allow you to find universities similar to any city you want to search. For example, if you live in Norfolk, Nebraska, and you enjoyed growing up there, then you may want to find a university that has similar venues compared to Norfolk. Although there is much more to a town than just its venues, this tool could still give potential students a good idea of places they make feel more content.

This tool could potentially be useful for websites like Niche that have various surveys that help students find the right school. Not feeling comfortable or "at home" at a university can cause students to transfer colleges or drop out. This added feature could improve student satisfaction in their choices, which in turn could bring more students into Niche to use their surveys and searches, which would bring in more advertising revenue. This type of tool could help universities better target their ads when they advertise on websites like Niche. Additionally, universities could be potential investors in this method, since they have a financial incentive to reduce transfers and dropouts.

**Data:**

A total of three data sources were used:

**1. UniRank (https://www.4icu.org/us/)**

This site offers a full ranked list of accredited colleges and universities in the United States. I can use Requests and BeautifulSoup to extract this data and get a full list of colleges and universities in the United States, which will provide the universities for the clustering.

**2. Bing Maps (https://www.bing.com/maps)**

I can use the Bing Maps API through geopy.geocoders to find the coordinates of the universities to use for the FourSquare API. Bing Maps will also provide the coordinates for the city search later. Inputting the name of a city or university searches for the location on Bing Maps and returns the coordinates according to Bing's data.

**3. Foursquare (https://developer.foursquare.com/)**

I will be using Foursquare to collect lists of venues near each university, which is the data that can then be used to cluster the universities using K Means. This will also be used whenever a location is inputted in the dashboard to collect a list of venues for any given city to then compare to the universities.

**Methodology**

**Data Collection and Cleaning:**

The list of universities was retrieved from https://www.4icu.org/us/ using the Requests library. BeautifulSoup was then used to extract the names of the universities and then converted the top 300 universities on the list to a Pandas dataframe.

From there, each university was run through the Bing API using Bing().geocode() from the geopy.geocoders library to retriever the coordinates, which were then added to the dataframe. Universities for which Bing did not return coordinates were deleted from the dataframe. Folium was then used to map the locations of all the universities on a map of the United States.

The coordinates were inputted into the Foursquare API to retrieve the top 100 venues near each university. The Foursquare API returned json files, from which the venue names and categories were extracted. Pandas.get\_dummies() was then used to get dummies for the venue category in order to turn the category into a numerical value, and then grouped the dummies by university. The final dataframe contained the university name and then each venue category with a number proportional to the number of that venue type near the university.

**K Means Clustering:**

The KMeans library from sklearn.cluster was used to cluster the universities into 8 clusters based on the venue categories data. Folium was used to visualize the univerisites, color-coded by cluster, on a map of the United States.

**Making Recommendation Dashboard:**

The JupyterDash library was used to make a dashboard in which you can input a location, and then the dashboard will list the top 25 recommended universities most similar to the location that was inputed. This is done using a function that uses the same methods listed above to add the searched location to the university dataframe, and then the k means predict function is used to determine to what cluster of universities the searched location belongs. From there, the top 50 universities that belong to that cluster are displayed. The dashboard is live updated, so when a new location is entered, the list refreshes.

**Results**

Since the algorithm uses data that cannot be easily visualized, it is challenging to determine exactly what the similarities are within a cluster, but based on the top venues at each university within each cluster, the universities generally fell into the following clusters:

**Cluster 1:** Parks, scenic spaces, and general natural places.

**Cluster 2:** High concentration of coffee shops and cafes.

**Cluster 3:** Breweries, bars, and pizza places.

**Cluster 6:** Many American restaurants as well as various other restaurant styles.

**Cluster 7:** Lots of pizza and sandwich placeText

Description automatically generated with low confidenceTable

Description automatically generated with low confidence.

Map

Description automatically generatedKey: Red: Cluster 1, Purple: Cluster 2, Blue: Cluster 3, Green: Cluster 6, Yellow: Cluster 7.

A few clusters resulted with only one university. Cluster 4 contained only the United States Military Academy, with top venues of lakes and campgrounds. Cluster 5 contained only Vanderbilt University, with American restaurants and art galleries as its top venues. Cluster 8 contained only Northern Arizona University, with top venues of theater and home services.

The dashboard worked as described in the methodology section, and was able to successfully display the top 25 recommended universities whenever a location is entered. Several locations were tested, and the list changed depending on the location entered, as demonstrated in the above images showing the lists.

**Discussion**

The objective of this project was the create a method to recommend universities that would be similar to a location that is searched by the user, and although there are some issues with the methodology, the dashboard created was successful in completing the recommendation task. Whenever a location is inputted, a list is generated with the top 25 universities that fall into the same cluster as the location.

The K means clustering was supposed to generate eight clusters. Cluster 1 resulted in universities with natural spaces. Cluster 2 contained universities with more coffee shops. Cluster 3 resulted in universities with bars and pizza. Cluster 6 had lots of American restaurants, and cluster 7 had many pizza and sandwich places. These categories are very loosely defined, however, and many universities had many venues similar to universities in other clusters. For example, the most common venue at University of Michigan is parks, but University of Michigan is in cluster 2.

One problem was that three of the clusters only had one university, with cluster 4 containing only the United States Military Academy, cluster 5 containing only Vanderbilt University, and cluster 8 containing only Northern Arizona University. With only one university each, these clusters are not useful for recommendations, and no location I entered recommended these clusters.

Although the program worked to recommend universities based on clustering, this may not be the best method for a few reasons. One reason is the aforementioned problem that several of the clusters ended up with only one university. This means that although I intended for there to be eight clusters, but effectively only ended up with five clusters that could be recommended. The clustering method is also problematic since it always recommends only the top universities within a cluster, and not the ones that are most similar within a cluster. This means the same 25 universities will be recommended from each cluster any time that cluster is selected, even if those 25 universities are not actually the most similar to the inputted location.

There are several methods that could optimize this recommendation system. One method would be to optimize which venues are used to cluster the universities. For example, the algorithm I used included venues like "Intersection" and "ATM", and these are venues that probably do not influence people's sense of home at a given location. Likewise, venues like restaurants and parks could be weighted so that venues that influence the quality of life of a location influence the model more heavily. Another approach would be not to use clustering at all, and rather recommend universities by plotting the venue category values, then calculating distance from the venue categories of the inputted location to the venue categories of the universities, and then recommending the universities closest to the location.

**Conclusion**

People transferring out of universities is a common problem that can hurt a university financially and can hurt their reputation if transfer rates are too high, and so universities are likely to invest to reduce their retention rates. One common reason that people transfer is because of the location and environment, so I developed an algorithm to recommend universities based on nearby venues. This would help students find a university most similar to their hometown or any other location they like so they can be more confident that they will like the area.

The algorithm was successful in clustering the universities and then recommending universities based on any inputted location and to which cluster the location belongs. Although the program was designed to use eight cluster, only 5 clusters were ultimately helpful, since three of the clusters had only one university.

This program is a good start on improving university recommendations based on nearby venues, but could be improved through removing unimportant venues like intersections and weighting more important venues like parks and restaurants more heavily. Also, the clustering method could be adjusted to a distance method that finds distance from universities to the inputted location rather than forming clusters. With more optimization, the deployment of this method on sites like Niche could potentially be used to help students make more informed decisions about what college they attend and reduce transfer rates.