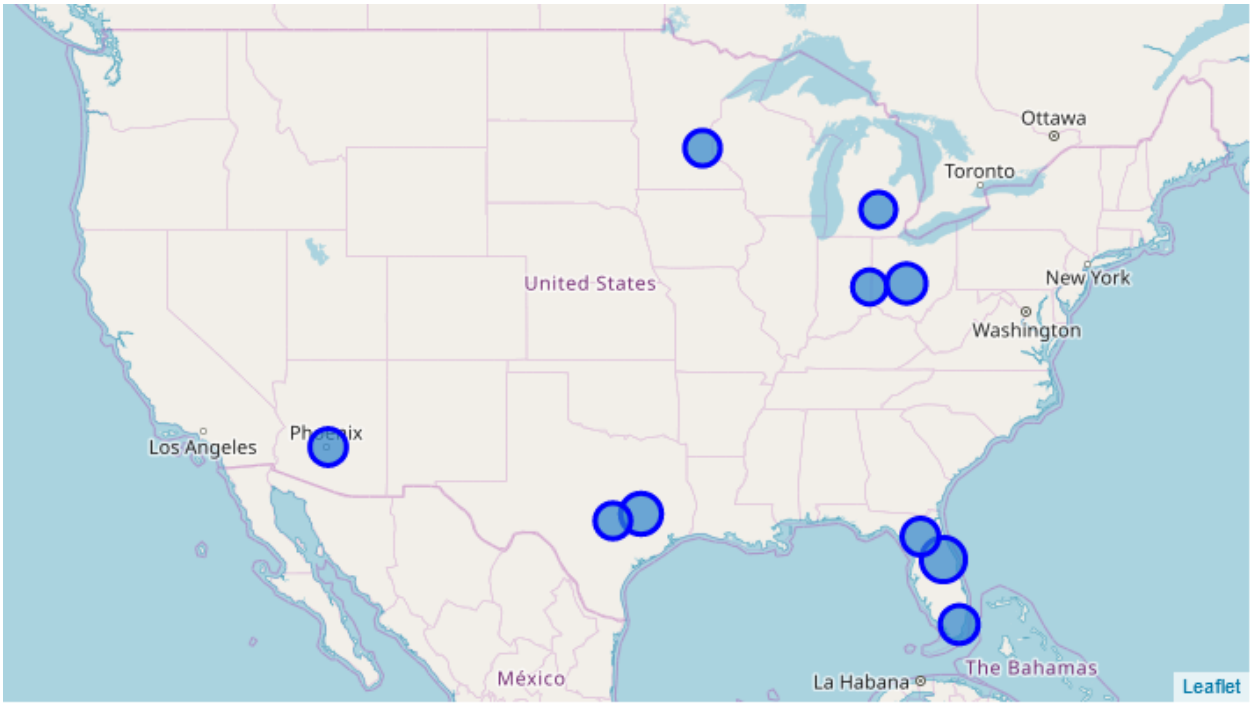
Applied Data Science Capstone

The Battle of Universities (Week 1)



Author: John Casavant

Coursera: Applied Data Science Capstone

Date: 13-January-2019

Table Contents

Introduction

This project is submitted for consideration of the Coursera Applied Data Science Capstone course. For this project we will analyze the Foursquare venue data associated with the top ten universities in the United States based on enrollment.

Problem: What makes universities in the United States similar? Can venue data be used to determine similarities?

We will try to determine if venue similarities can help us discriminate venue categories for the top ten universities in the United States and group these into clusters using K-Means classification.

Data

I have search and identified a world atlas webpage that identifies the enrollment by academic year 2015-16 for the top ten universities in the United States.

	Rank	University	Location	Enrollment
0	1	University of Central Florida	Orlando, Florida	63016
1	2	Texas A&M University	College Station, Texas	58515
2	3	Ohio State University	Columbus, Ohio	55508
3	4	Florida International University	Miami, Florida	54058
4	5	University of Florida	Gainesville, Florida	52519
5	6	Arizona State University	Tempe, Arizona	51984
6	7	University of Texas at Austin	Austin, Texas	50950
7	8	University of Minnesota	Minneapolis/Saint Paul, Minnesota	50678
8	9	Michigan State University	East Lansing, Michigan	50000
9	10	Indiana University	Bloomington, Indiana	48514

Source: https://www.worldatlas.com/articles/largest-universities-in-the-united-states.html

Using this source data on universities we will then capture and merge geolocation latitude and longitude data from the geopy geocoders Nominatim library.

Foursquare API will then be used to capture any venue and venue category data within one mile (1609 meters), defined as “walking distance”, from the center of each university location.

The data will be transformed into multiple dataframe objects required to analyze and map the results.

The university data will be displayed on a Folium map with the University enrollment size proportioned as the radius of each specific university location.

Methodology

Geocoder is used to generate the latitude and longitude of each university location.

K-means analysis is used to cluster the venue into 4 distinct clusters, and plotted onto a map of the United States

Each cluster will be displayed and analyzed to determine a discriminating venue category, if one exists or can be determined.

Results

To be added during week 5.

The results will be displayed using a Folium map showing the university clusters.

Discussion

To be added during week 5.

Conclusion

To be added during week 5.