Casey Poon, Leanna Hue
Professor Hu
CS 396: Introduction to the Data Science Pipeline
19 January 2020

# Assignment 1 Summary

In order to clean the dataset for cities in Arizona, we limited the dataset to only reviews in Arizona by filtering the dataset to only have “state=‘AZ’” entries. After manual inspection on the modified dataset, we noticed that some cities had inconsistent capitalization, residual punctuation, unicode encoding errors, and vague city names such as ‘az’ or ‘arizona’. In order to reduce the total types of variation due to capitalization, we first temporarily converted all the city names to lowercase. We then removed all punctuation and encoding errors from every city entry. If a specific variation of ‘arizona’ was preceded by another word (eg: phoenix, arizona) we changed the city name to everything preceding the variation of ‘arizona’. In the case where nothing preceded ‘arizona’ or a variation of it, we temporarily set the city name to an empty string. After all the string modifications, we converted all cities from all lowercase to proper nouns. To validate that a city in the dataset was actually a real city, we went to Arizona’s state website and created a list of all the incorporated city names. We then iterated through each entry again; if the city name was an empty string or not in the list of city names, we got the latitude and longitude of that entry and use geopy’s reverse search function to get the city name. If geopy was unable to get a city name from the latitude and longitude, we omitted the entry from the dataset.