

Comparison of Beaches

Analysıs and clusterıng of beaches

usıng

data scıence tools and technıques

GOKHAN EKICI | IBM DATA SCIENCE CAPSTONE | 06/07/2020

# Definition of The Problem

When you move to a new country or when you would like visit a new country for sun, sea and beaches, you have many websites that provide information about you options. Just like moving to a new neighborhood you want to learn the details, but some details are confusing and some of the content you have to read biased. When you search for “best beaches in US” on Google for example, you receive over 400.000 results. And although you know Google sorts them in a relevant order, when you check those results you are confused, you see that the top lists do not overlap, and you might get confused.

On the other hand, when you are deciding to buy a tool, a furniture a holiday package you mostly rely on people with similar tastes like you. But if you are new to a country, if you will visit it for the first time then the chances that you will find someone to understand what you want and respond to it with a good recommendation gets lower.

Today as we have so much information why would not I make my own decision, based on available data out there? At least a similarity analysis, a clustering algorithm might help us all choose places to see given the similarity between the places we like our favorite beach and the surrounding venues and neighborhood. This could be a self-service model which could lead to an affiliate model for tourism agencies, or could be directly used by tourism agencies.

A choice of a beach holiday for me has 3 main elements.

1. The beach, its rating and the neighborhood restaurants and cafes and bars etc.
2. The climate
3. The accommodation choices

During this analysis I will focus on the first 2 as most beautiful beaches already have many accommodation choices and there is plenty of reliable price and feature data on booking.com, hotels.com and Airbnb. I think the hard part is finding the right beach.

# Data

I will use 4 main data sets available online, the ones available for free for the sake of this Capstone Project.

1. A list of best beaches in US derived from USNEWS Travel site, Tripadvisor and some others at first.
2. Foursquare Ratings for these beaches
3. Climate data from visualcrossing.com
4. Foursquare Venue Data available through

The Beach based rating, important climate information (temperature, humidity etc.), venues and classes of venues in the neighborhood will be normalized and clustered so that we can compare one or more favorite beaches of ours to the new ones we have no idea of and want to see how they fit into our tastes.

I intend to use K-Means and DBSCAN for clustering a short list of beaches and create a model that can be used for analysis for other beaches etc. as well.