Introduction

Tens of millions of trips occur each year. Many people love visiting both new and familiar locations. One of the key questions when planning travel is “*Where to next?*” When considering where to go next, my past travel history has been a strong factor. This can either be in a positive light: *"I’d like to visit someplace like Barbados”* or in a negative manner: *"I’d like to visit a country warmer than Sweden.”*

Each trip is different and can take various forms and use various transport and accommodation methods (airlines, train, cruises, hotels). This makes gathering a comprehensive view of travel data a problem. This is compounded by the general unavailability of official APIs from these providers from which we could fetch and then aggregate this information.

A second and larger issue is how do you determine similarity between destinations. Gathering geographic and demographic information such as weather, size and population can help but can we go beyond this to base recommendations on user generated data? Can social media help?

Data Collection and Processing

Overcoming User Travel History Limitations

While it can be hard to gather a user’s travel history from travel providers, the rise of social media sites such as Facebook, Foursquare and Instagram provides the ability to analyze user data from these sites as quite a bit of social media is now tagged with location information. Foursquare is specifically a location oriented social network, and given that I am a heavy user and my familiarity with their API on prior projects, I decided to extract countries that I had visited by analyzing my prior check-ins on the network. The result was a list of 9 countries extracted from 3,000 checkins at 1000 venues

Gathering Social Data for Recommendations

The was the more daunting task. Gathering social media data about countries around the world was challenging in order to ensure countries were fairly represented in the data set, I would need to gather a large amount of data, much more than would be present in the user travel history dataset.

The first question was which countries should be included in the list? In the interest of time, I decided to simply use the Top 100 Countries by GDP, in this case using GDP as a proxy for desirable locations. Once I had the countries identified, I used the GeoNames API to fetch information for each country including capital, population, land mass and most importantly ISO country code. The reason the ISO code was important was that it was a unique identifier and Foursquare also returned this information for venues within their API thus allowing me to match data up easily.

Now that I had this basic country information, I could begin to collect information from social media APIs. I focused my API searches on country names. For each country, I converted the country name into a hashtag of sorts (e.g. United Kingdom became unitedKingdom). For countries typically called by different names, I manually added other keywords. Each keyword was it’s own search. A sample country JSON record for the United Kingdom is below:

```
{
        "area_in_sq_miles" : 244820,
        "code" : "GB",
        "gdp_per_capita_rank" : 34,
        "keywords" : [
                "UNITED KINGDOM",
                "BRITAIN",
                "GREAT BRITAIN",
                "SCOTLAND",
                "ENGLAND",
                "NORTHERN IRELAND"
        ],
        "name" : "UNITED KINGDOM",
        "population" : 62348447
}
```

For Twitter in particular I also included the “travel” keyword to ensure that tweets were specific to travel. I did not use this for Instagram given that I believed that a hashtag in an Instagram post would already be more likely about a visit or travel somewhere compared to Twitter. For each post, I stored the matching keyword that it was found against for linking it to the country data. Please see below for a tabular description of my search criteria and post metrics.

Search Criteria

	Instagram	**Twitter**
Query Format	#<country>	<country> + “travel”
Example Query	#barbados	“barbados travel”

Post Information

Total Posts	**Min**	**Max**	**Average**
~34,000 Instagram + ~20,000 Twitter = **54,000+ posts**	98 posts (N. Mariana Islands)	2,975 posts (United Kingdom)	~500 posts per country

All of my data gathering was done with Python with the raw post data being stored in MongoDB for later processing. MongoDB was chosen due to it's ability to store JSON data naturally without defining a formal schema.

Processing Social Data

Dealing with text data presents it's fair share of challenges, especially when that data is coming from multiple locations in a variety of formats. The first necessary step was to standardize the raw post data coming from Instagram and Facebook into a single format. To accomplish this, the post text itself along with any comments and/or hashtags were extracted from the post object and concatenated into a single string.

Once this was complete, the Python NLTK library was the primary tool used given it's usefulness for natural language processing. The NLTK library was first used to tokenize the strings above into individual words. Once they were words, the library was used to lemmatize the dataset. Lemmatizing is the process by which different variations of a word are simplified into a single form. E.g. convert Travel, travels, traveled, traveler -> travel. The aim here was to make it easier to find groupings within the data.

At the end of the process there were over 1.2 Million non-unique tokens. It should be noted that the language of an individual post was not actively examined. Limiting the language of the post to English could have significantly reduced the data set on non English speaking countries. That being said, the lemmatization process above was set to English so it is possible that applying language filters could lead to a more refined result once analysis was completed.

As a final step, I used the SciKit Learn Python library to create a sparse matrix of the data set. To accomplish this I used the TfidfVectorizer module. This provided the perfect input format for Data Science algorithms needed in the Analysis step below.

Data Analysis

I initially planned on deploying a collaborative filtering approach to the dataset. In order to accomplish this, I wanted to use supervised learning. In this scenario, I as the user would manually group the countries I had been to and feed it into the Naive Bayes algorithm in order to build a model. I would then attempt to apply this model to other countries. My model could then have iteratively gotten better over time as I visit more countries and group them. Another option for the long term would be to identify similar users to myself and use their groupings as input data to my recommendations.

A second option was to stick with collaborative filtering but to attempt recommendations via unsupervised learning where instead of asking the user to group the countries themselves, I could feed a user's travel history into KMeans and have it calculate groupings that I could then apply to the full list of 100 countries.

I quickly abandoned both of these approaches as I felt that the limited training set provided for my situation (I have only been to 9 countries so *cold start problem)* would limit the model's ability to predict accurately. I could have enhanced the supervised learning model by manually grouping unvisited countries and then building a model based on a combination of actual travel history and potential destinations. This would have helped mitigate the problem of limited historical data but would require significantly more work and long term would be asking more of users. At the same time, including unvisted countries would enable more precise groupings as a user's current travel history is probably not authoritative to how they would group all countries. My history certainly was not.

To deal with these limitations, I decided instead to go with a content based filtering approach and unsupervised learning. Instead of relying solely on user provided data to group countries, I instead used the largest data set available. I used the posts I had acquired about each country. With this data, I identified a feature space composed of word tokens within the posts and completely ignoring whether I as a user had visited or not. My thought process around layering in user travel history was that once I had my clusters identified, recommending locations would be as simple as 'show me places like Barbados' where Barbados would be a country that I visited.

To do this, I used KMeans algorithm within SciKit Learn, generating models with varying parameters including number of clusters by which to group the data. I tested each variation of my model manually primarily based on my own idea of how I expected countries to be clustered. This was relatively simple but given that there were only 100 countries to cluster worked for this situation.

The biggest issue with this approach would be the potential for the model to be overfitted to my preferences for groupings. In a scenario with a larger number of users and more data, I could turn the problem back into a supervised model assuming a larger user travel data set on which to train the model.

Once I had my groupings identified, it occurred to me that in addition to identifying clusters, it could be useful to apply additional filter options as a decision on where to travel is rarely based only on similarity. There are other factors like country size, population, GDP, etc that could be useful to users. Given this, I added the ability to filter my clusters based on these factors. As part of this, I also added the ability to add negative comparison. E.g. "A country NOT LIKE Barbados."

I converted my result set into a simple web application built with Python and Flask, which you could query. It would return a display of countries with their flags depending on the query you entered into the search box. In addition, it would 'gray out' any countries you had already visited in the list. A list of sample queries are below and what they mean are below:

bb	LIKE Barbados
bb-	NOT LIKE Barbados
bb {size+}	LIKE Barbados BUT Bigger in Area
bb {people-}	LIKE Barbados BUT with Less People
bb; us	LIKE Barbados OR LIKE United States
bb {people+}; us {size-}	LIKE Barbados OR LIKE United States BUT with More People than Barbados and Smaller than the US

Conclusion and Next Steps

This project was a great chance for me to apply the course to an idea that I was very interested in. It enabled me to play with multiple components of of Data Science such as natural language processing and KMeans in order to build a recommendation system. On top of that building a user interface to the system was something that was extremely important to me as it depicted a usable product. The model itself can evolve over time as I gather more information and place it on a platform such as Heroku where others can access it.

Given that, in order to improve my model long term I would taking a closer look at the following:

- Would a collaborative filtering approach work better once I have more user data?
- What impact would taking into account the language have? Would it positively or negatively impact the model to lemmatize each post based on language?
- Could this be expanded to city level instead of just by country? Larger countries such as the USA and Australia can be very different depending on the city.
- Was GDP the best way to rank countries to be included in the project? China for example was missing given it’s lower GDP Per Capita ranking.
- What additional factors could be used the filter the list once clusters were identified? Things like languages spoken, average cost of travel, accommodation, etc.

Note: Sample Instagram and Twitter posts are included in the Appendices below.

Appendices

Sample Instagram Post

```
{
        "user_has_liked" : false,
        "attribution" : null,
        "tags" : [
                "taiwan",
                "america",
                "wanna",
                "gotothe",
                "usa"
        ],
        "comments" : {
                "count" : 0,
                "data" : [ ]
        },
        "filter" : "Toaster",
        "location" : null,
        "matching_tags" : [
                "America"
        ],
        "created_time" : "1384187068",
        "caption" : {
                "created_time" : "1384187071",
                "text" : "#Taiwan#Wanna#gotothe#USA#America",
                "from" : {
                        "id" : "222596262"
                },
                "id" : "586926154302340141"
        },
        "type" : "image",
        "id" : "586926126527659707_222596262",
        "likes" : {
                "count" : 0,
                "data" : [ ]
        }
}
```

Sample Twitter Post

```
{
        "created_at" : "Mon Nov 11 21:01:28 +0000 2013",
        "favorite_count" : 0,
        "favorited" : false,
        "id_str" : "400005377959677954",
        "lang" : "en",
        "matching_tags" : [
                "France"
        ],
        "place" : null,
        "possibly_sensitive" : false,
        "retweet_count" : 0,
        "retweeted" : false,
        "source" : "stream",
        "text" : "RT @france_images: Places to stay in Northern #France the village of Wierre Effroy #Travel http://t.co/WziJuk7YNm",
        "truncated" : false,
}
```