Vacation Trend Analysis

Chandan Gope
Daniel Lee
K.C. Tobin

Application Summary

Problem: What travel destinations are worth exploring at a certain time of year?

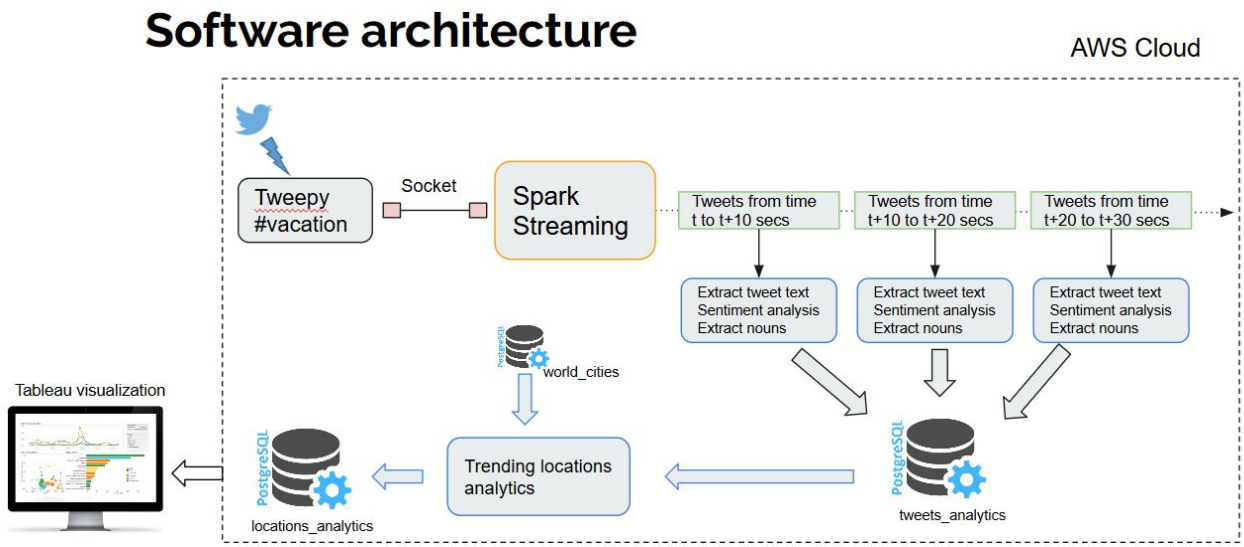
Solution: Vacation destination insight dashboard

The application is meant to provide insights and analysis to vacation goers and vacation planners by analyzing streaming information to categorize travel destinations. Utilizing the tweepy wrapper for the twitter API we extract streaming tweets with a filter for vacation. Utilizing this data we perform natural language processing techniques to extract relevant unigram features and perform sentiment analysis. We further impute the location subject from the text to identify the travel destination of interest.

With this information processed we are able to present a dashboard of travel locations across the world and visually communicate information regarding popularity of location spots as well as a general sentiment regarding those locations for users to peruse and help determine a future trip. We also present pre-built dashboards identifying the top locations by mentions and top locations by sentiment. While our data was accumulated over 3 days the logic could be built out to incorporate temporal components to identify the popularity of locations over time. For instance, we predict southern hemisphere destinations to be more popular during traditional North American winter months (December - February) as the Southern Hemisphere is in their traditional summer months. We also foresee being able to use this information to identify locations that see spikes in their interest by comparing their mentions between predefined time intervals.

Architecture

In order to build out a steel thread we utilized several technologies and hosted our storage and processing layers in an Amazon EC2 instance. Below you can see a visual representation of the steel thread architecture. We will now go into more depth on the individual components of the architecture.



Spark

Apache spark is an open source cluster computing framework. Spark serves as the stream processing layer in our steel thread taking incoming tweets and transforming them into the proper format to store into our storage layer.

PostgreSQL

PostgreSQL is a relational database management system and servers as the storage layer for our application. It maintains the incoming twitter data as well as the processed analytics data for future serving. Below you can see the primary databases involved and their schemas:

tweets_analytics schema:

Column	**Type**
ID	Varchar
Update_Time	Timestamp
Text	Varchar
Nouns	Varchar
Sentiment	Numeric

Sample data from tweets_analytics table

Abc tweets_analytics Id	Abc tweets_analytics Update Time	Abc tweets_analytics Text	Abc tweets_analytics Nouns	Abc tweets_analytics Sentiment
942173076938285061	2017-12-16 23:22:30	Last Year I traveled to #Bali for winter vacation. This year I may stay in with my twins.	Year,Bali,vacation,year	0.0
942173084139757568	2017-12-16 23:22:30	Hotels in Mexico	Mexico,-→,Mexico,Ho...	0.0
942173086379597825	2017-12-16 23:22:30	😂 https://t.co/37A1rhC7Fd	😂	0.0
942173125793468417	2017-12-16 23:22:40	@AprilCArmstrong They were talking about benefits, bc the dude just got a sweet job with paid vacation. The woman s... http...	benefits,,dude,job,va...	0.6808
942173149528981504	2017-12-16 23:22:40	@RogerBurtonwphs @WP_Athletics Yes sir we all need a nice vacation. I'll see you guys soon!!! Happy holidays to my @WP_A...	sir,vacation,I'll	0.918
942173198224896000	2017-12-16 23:23:00	Had a dream last night ,that I went on a vacation to Costa Rica cr	dream,night,Costa,Ri...	0.25
942173199370014720	2017-12-16 23:23:00	@tuttifruti92 @VAHNNNN it's not even a vacation, it's prob. just alone time before family time during Christmas, bu... https:/...	@VAHNNNN,it's,vaca...	-0.25
942173226649632769	2017-12-16 23:23:00	I wonder if their tourism industry will take a nose dive, or just shift to visitors with like-minded views to their... https://t.co/eL...	industry,dive,	0.0
942173246002262018	2017-12-16 23:23:10	Live 5 Investigates: Dozens of complaints filed against local vacation clubs - Live 5 News via @HDBoykinJr... https://t.co/8Szxt...	Investigates:,Dozens,...	-0.4019
942173289031651329	2017-12-16 23:23:20	@mmadamimadamm You totally deserve a "vacation" no need to explain. Thanks again for everything that you do!🖤🖤🖤	@mmadamimadamm,...	0.1848
942173355041583104	2017-12-16 23:23:30	https://t.co/yzm6G3QtoW. Join us next year on a trip to Nicaragua. Combine vacationing with meaningful cultural experience...	Join,year,trip,Nicarag...	0.5423
942173365820952576	2017-12-16 23:23:40	I could use some vacation! 😅😫#tired	vacation!,😅😫	0.0
942173385139806208	2017-12-16 23:23:40	Hotels in Bali	Bali,-→,Bali,Hotel,Tra...	0.0

location_analytics schema:

Column	**Type**
Text	Varchar
ID	Varchar
Update Time	Timestamp
Sentiment	Numeric
Place	Varchar
Latitude	Numeric
Longitude	Numeric

Sample data from location_analytics table

Text	Id	Update Time	Sentiment	Place	Latitude	Longitude
My dads talking about going on vacation in June and he's like where do ...	942145949945626624	2017-12-16 21:34:40	0.6486	Hawaii	20.837	-157.140
My dads talking about going on vacation in June and he's like where do ...	942145949945626624	2017-12-16 21:34:40	0.6486	Cancun	21.170	-86.830
Всем отличного дня из Мексики! #vacation #Mexico #TeAmoMexico ht...	942146012763906049	2017-12-16 21:34:50	0.0	Mexico	22.497	-101.126
BEACHIN取 | Summer Vacation on Florida Beaches Outdoors Allie #Tact...	942146073690288128	2017-12-16 21:35:10	0.0	Florida	26.505	-81.286
BEACHIN取 | Summer Vacation on Florida Beaches Outdoors Allie #Tact...	942146075070255105	2017-12-16 21:35:10	0.0	Florida	26.505	-81.286
#vacation - Wonderful places to go in Limerick, Republic of Ireland : htt...	942146078161371136	2017-12-16 21:35:10	0.5719	Limerick	52.665	-8.623
#vacation - Wonderful places to go in Limerick, Republic of Ireland : htt...	942146078161371136	2017-12-16 21:35:10	0.5719	Ireland	53.175	-7.960
Vacation in Florida hotel https://t.co/032xQuwrpm	942146087107756032	2017-12-16 21:35:10	0.0	Florida	26.505	-81.286
I don't know about you but I'd love to travel to Thailand. ?? #photo #lux...	942146100118646789	2017-12-16 21:35:10	0.7998	Thailand	13.939	100.856
#Colorado Springs ski shops hope for snow after slow start to season h...	942146304863547393	2017-12-16 21:36:00	0.4404	Colorado	39.087	-105.832
#Colorado Springs ski shops hope for snow after slow start to season h...	942146304863547393	2017-12-16 21:36:00	0.4404	Springs	-26.270	28.430
TIME FOR VACATION. (@ Singapore @ChangiAirport in Singapore) https...	942146326405500928	2017-12-16 21:36:10	0.0	Singapore	1.293	103.856
TIME FOR VACATION. (@ Singapore @ChangiAirport in Singapore) https...	942146326405500928	2017-12-16 21:36:10	0.0	Singapore	1.293	103.856
Желаем всем отличного дня из Мексики! #vacation #Mexico #TeAmo...	942146385406722059	2017-12-16 21:36:20	0.0	Mexico	22.497	-101.126
Battle Mountain tops Cheyenne Mountain in hockey https://t.co/rFL2dC...	942146390586687488	2017-12-16 21:36:20	0.1779	Cheyenne	41.140	-104.820

Tweepy/Twitter API

Tweepy is a python library that serves as a wrapper for the twitter API. This is the primary streaming datasource that is used for vacation analysis. We have a running streaming API that is filtered to the vacation track pulling down tweets.

Python

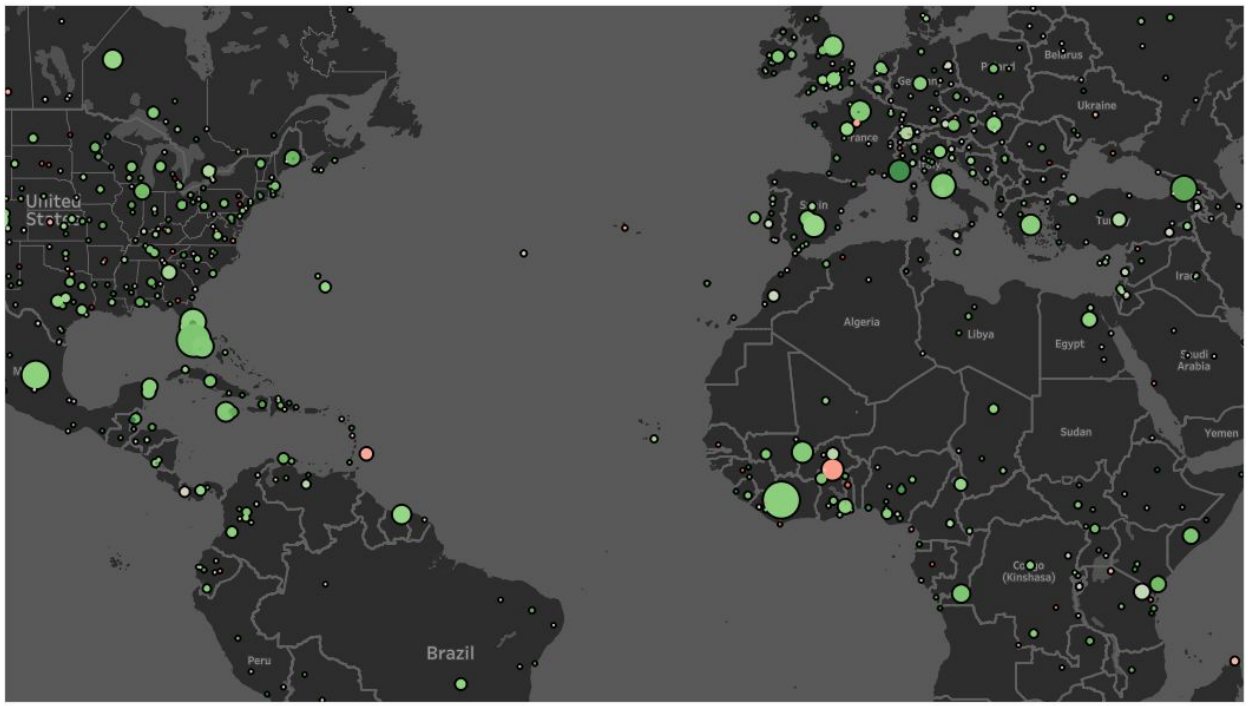
Python is a high level programming language that serves as the primary processing function for the application. In addition to the tweepy API calls, we use several Python libraries to transform the incoming data to a format that can be used for analysis. A few notable libraries:

- NLTK - natural language processing library used for sentiment analysis
- Fuzzywuzzy - natural language library for performing string matching using various metrics like Jaro-Winkler distance or Levenshtein distance
- Psycopg2 - library used for establishing connections and executing statements against postgres databases

Tableau

Tableau is a professional visualization software that provides an intuitive user interface for building data visualizations. We used tableau to provide a geographic symbol map for identifying tweet sentiments and frequency for various locations as well as prebuilt bar charts to identify vacation places of interest.

Trending Vacation Spots

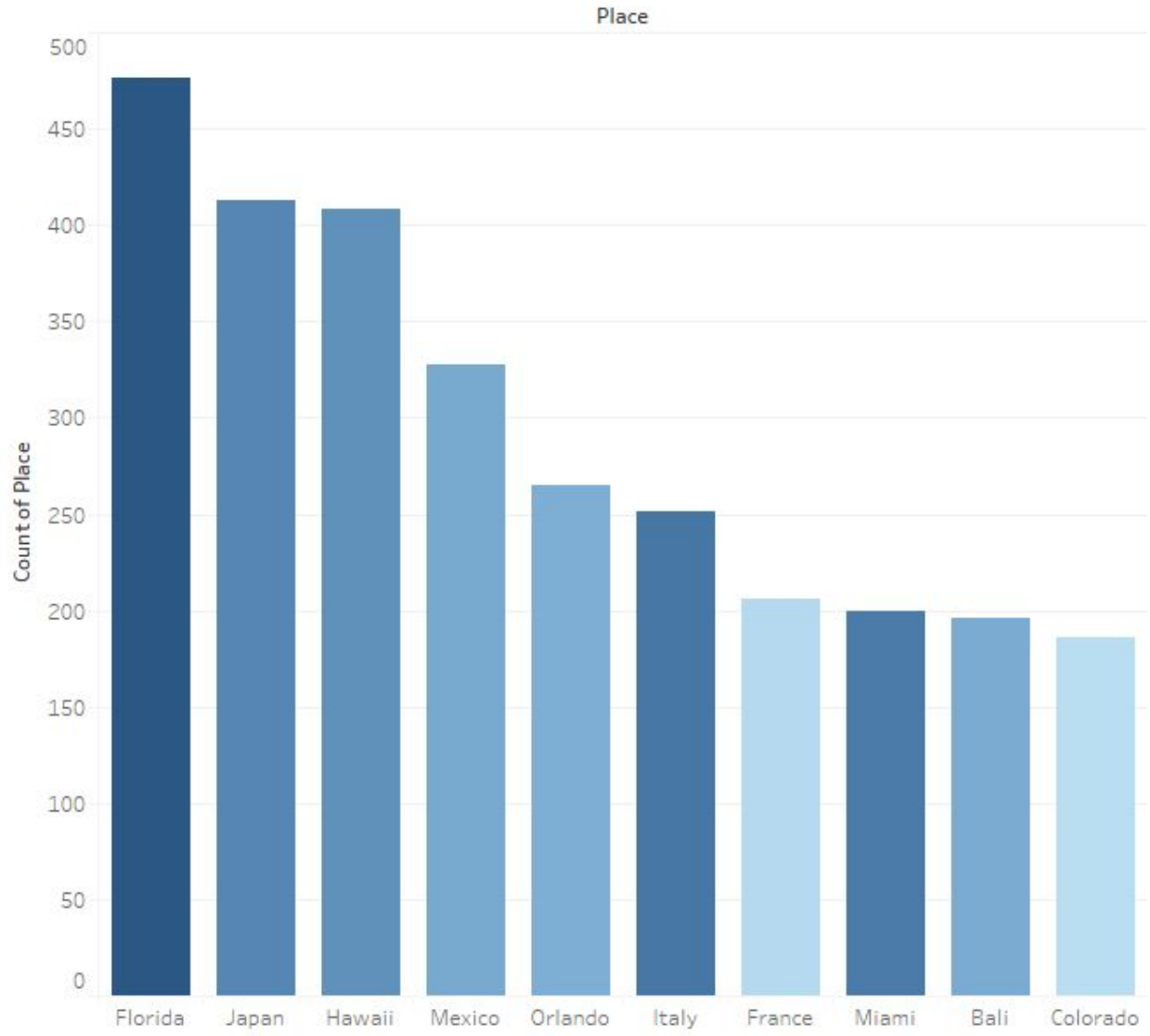


Map based on Longitude and Latitude. Color shows average of Sentiment. Size shows count of Place. Details are shown for Place. The view is filtered on Exclusions (Latitude,Longitude,Place), which keeps 1,023 members.

Results

After an analysis of over 1,000 tweets, the top three most tweeted about vacation destinations were Florida, Japan, and Hawaii. The top three locations with the highest average tweet sentiment were surprisingly Kissimmee (FL), Portland (OR), and Alaska. The bottom three locations with the lowest average tweet sentiment were Albuquerque (NM), Orleans (New Orleans, LA), and Barbados. Given the relatively small size of the final locations_analysis dataset, and given the ephemeral and rapidly changing nature of tweets, it is likely that these results would change over time.

Top 10 By Count

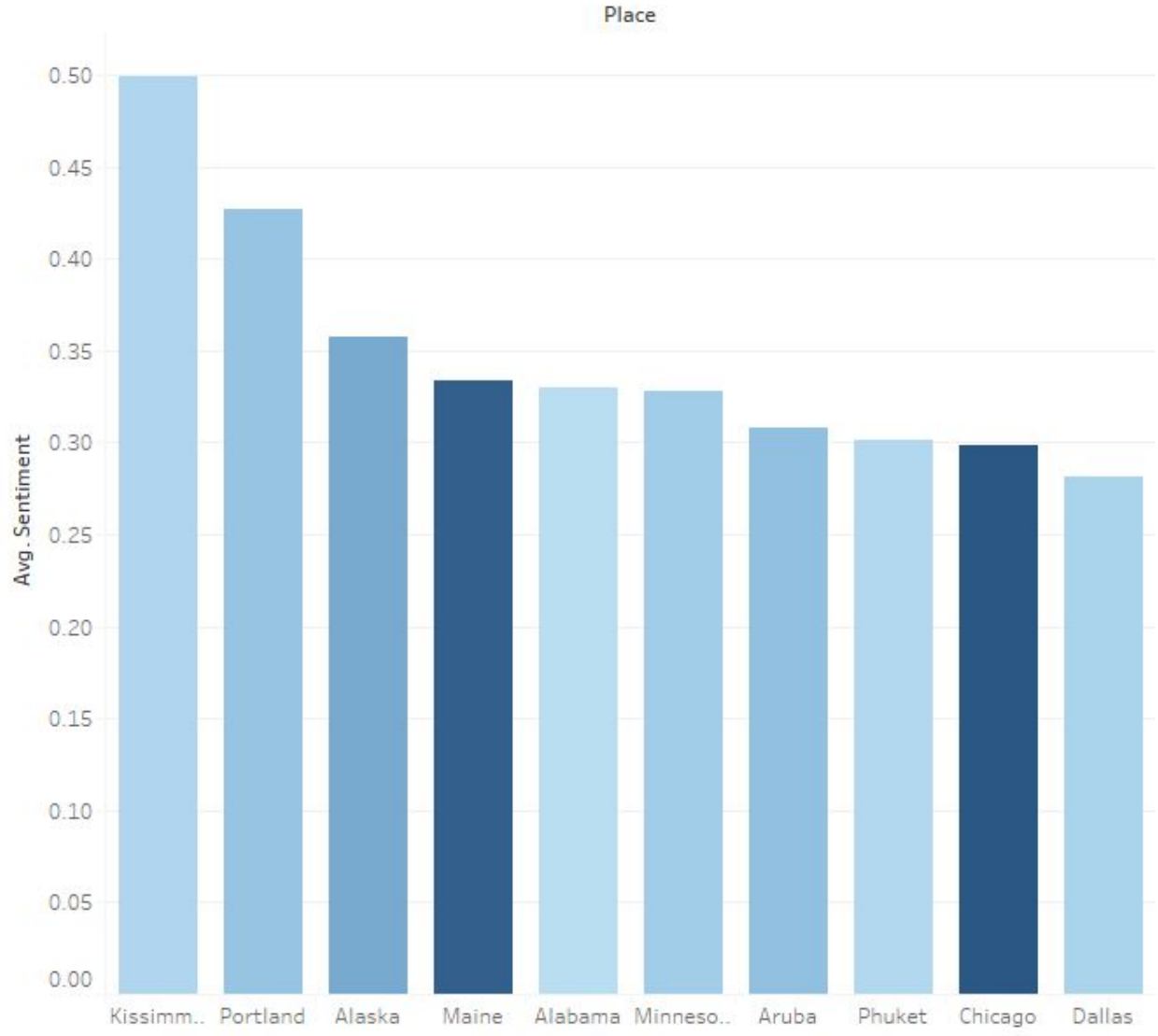


Count of Place for each Place. Color shows average of Sentiment. The view is filtered on Place, which keeps 10 of 1,023 members.

Avg. Sentiment

0.0902 0.2319

Top 10 By Sentiment

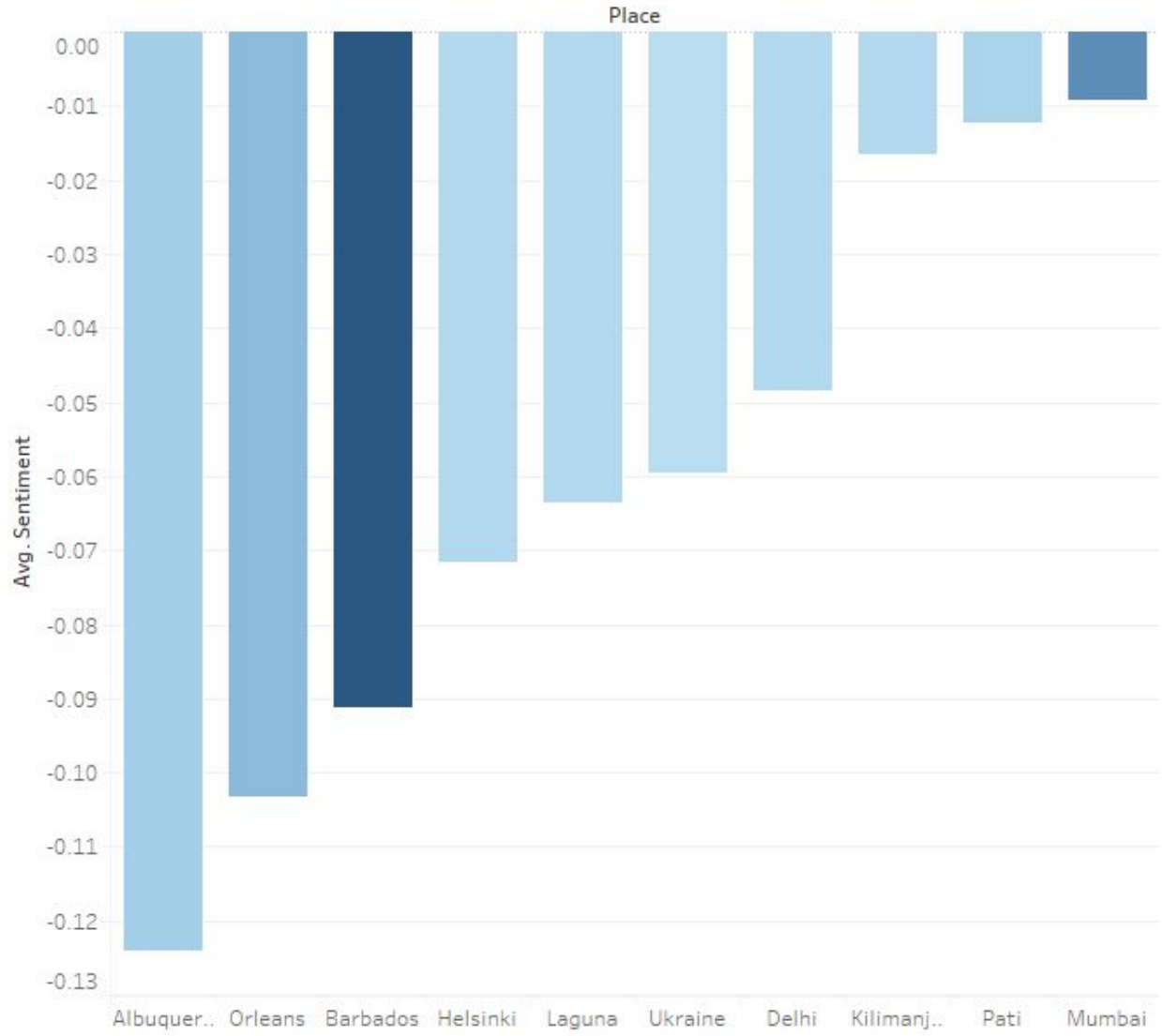


Average of Sentiment for each Place. Color shows sum of Number of Records. The view is filtered on average of Sentiment and Place. The average of Sentiment filter keeps all values. The Place filter keeps 10 of 1,023 members.

Number of Records

21 94

Bottom 10 By Sentiment



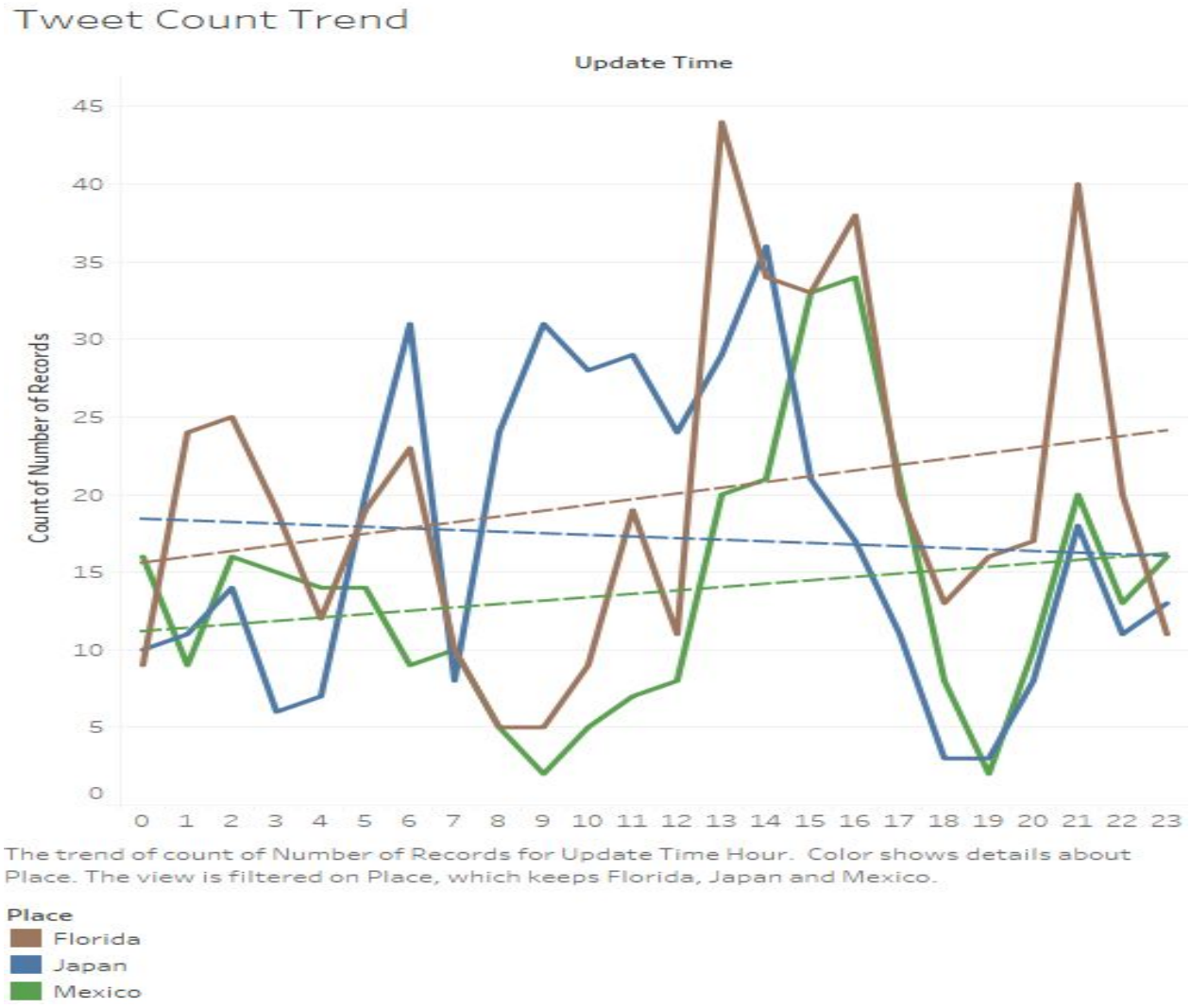
Average of Sentiment for each Place. Color shows sum of Number of Records. The view is filtered on Place, which keeps 10 of 1,023 members.

Number of Records

6 60

Serving/Visualization

Tableau offers professional services for a scale out architecture including its own server so we could continue to use that. Alternatively, scale out options include building our visualizations via various visualizations tools such as plotly, dash, d3, or even a shiny app.



How To Run

This project uses Twitter streaming, filters streams with #vacation, and analyzes them to compute trending locations with associated sentiments. Follow these steps to run the scripts -

1. Make sure the Linux instance has these installed - Spark, Postgres, Tweepy, psycopg2, nltk, git
2. Run the psql command lines provided in aws_code/psql_commands_createtables.txt - This will create the necessary tables required by the scripts
3. In a terminal run python aws_code/TweetRead.py - This will start the Twitter streaming on port 5555
4. In another terminal run spark-submit aws_code/SparkDemo2.py - This will start Spark streaming listening on port 5555
5. At this point, the table tweets_analytics will start getting populated every 10 seconds
6. On a remote computer(such as your desktop), run the python notebook provided in remote_code/tweetanalytics_v4.ipynb. You will need to adjust the hosturl param. This will get the tweets from the timewindow specified in the code, process those tweets, and add the information to location_analytics
7. Now start a Postgres connection in Tableau, connect to the hosturl where the above mentioned scripts are running, and connect to the table location_analytics
8. You can now visualize the data in table location_analytics

File Dependencies

- Python
- Git
- Psycopg2
- Tweepy
- Hadoop
- Postgres
- NLTK
- FuzzyWuzzy
- Vader SentimentIntensityAnalyzer

Scale Out

In order to scale out our application we would need to switch from storing our data in PostgreSQL to Hive or Red Shift. Areas for future consideration also include Incorporating other data sources, (e.g. TripAdvisor, Facebook posts, or other social media sites, etc.)

In order to speed up the analytics of processing tweets, we also need to move the location extraction processing from batch processing to stream-processing. The stream processing takes about 0.5 seconds to find words matching locations against a table of 7300 world cities. Currently we are performing the processing serially, after items are added to the postgres table, which significantly adds to the processing time.

Future Work

Due to time constraints there were several features that were initially in scope that we were unable to get to that we would look to incorporate a full scale build out with the scale out architecture. You can see the list and brief description below:

- Personalized recommendations, creating user profiles, preferences, mentions, nearest neighbor analysis based off prior travel mentions
- Incorporate additional data sources, TripAdvisor, other social media sites
- Incorporate trending metrics over time
- Improving the logic of our algorithms to better match locations and assess sentiment of tweets.