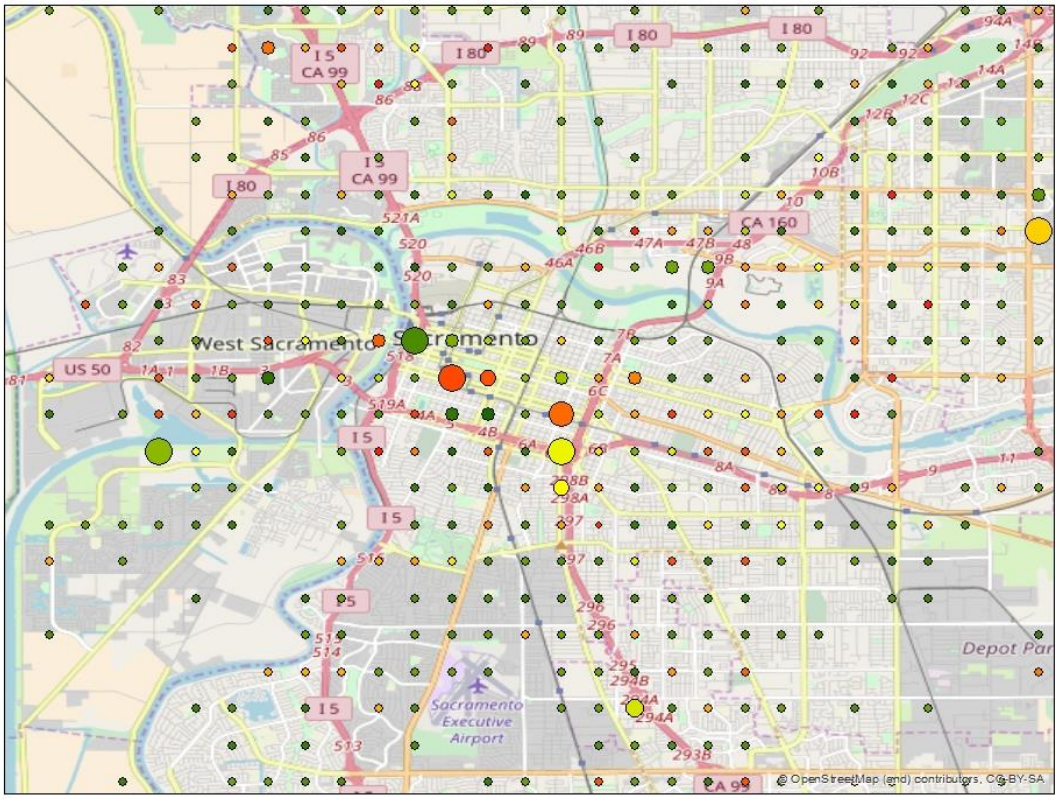
GEOG 375

Twitter Based GeoSentiment Analysis

M.Hernandez, PhD

Sacramento Tweet Sentiment

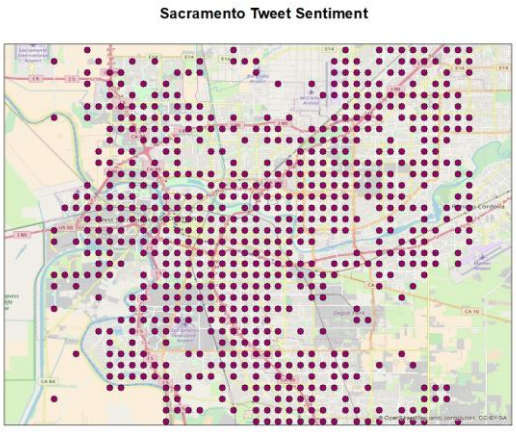


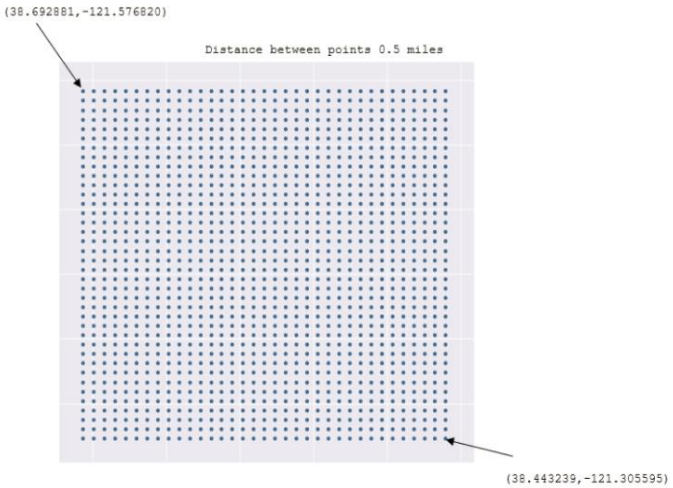
Executive Summary

This project details process and outcomes of implementing a geospatially dependent sentiment analysis of twitter data for the Sacramento region during the month of April 2018. The process presented outlines one approach to implementing ArcGIS's arcpy python module for automating the generation and management for a point feature class rendered onto three dynamically updated ArcMap documents. In addition to these geoprocessing techniques, this project outlines data pulling of tweets and uploading images to Twitter. The data pull outlined here details how to create and maintain an independent and persistent data pull from Twitter that returns geospatially dependent information. This geospatially dependent twitter information is then encoded with a sentiment value of positive, negative, or neutral that is later rendered onto the ArcMap document. While the data uploading mechanism is detailed and applied to the generated images of ArcMap documents for a specified Twitter feed.

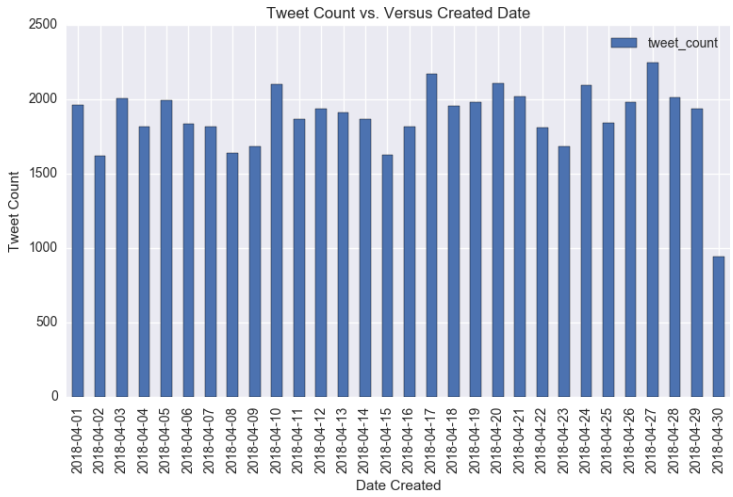
Summary

This project highlights how one may leverage ArcGIS's arcpy python module for automating geospatially dependent information. The geospatial dependent information presented here encompasses mathematically determined sentiment of Twitter tweets captured in the Sacramento region between 04-01-2018 to 04-30-2018. The term sentiment of a tweet is defined here as a view or attitude of an individual tweet taking on values of positive (+1), negative (-1), or neutral (0). Due to the way that this project captures tweets, the sentiment value can be encoded to a specific geographic region enabling the rendering of sentiment in a cartographic representation. Geographic techniques specific to ArcGIS such as map generation, feature class generation, feature class data management, spatial referencing, and layer management are presented here. This project generates a grid of geographic data points with a separation distance of 0.5 miles apart, covering a geographic region that includes downtown sacramento. (The figures below are visual representations of the geographic data points that were monitored throughout this study.)

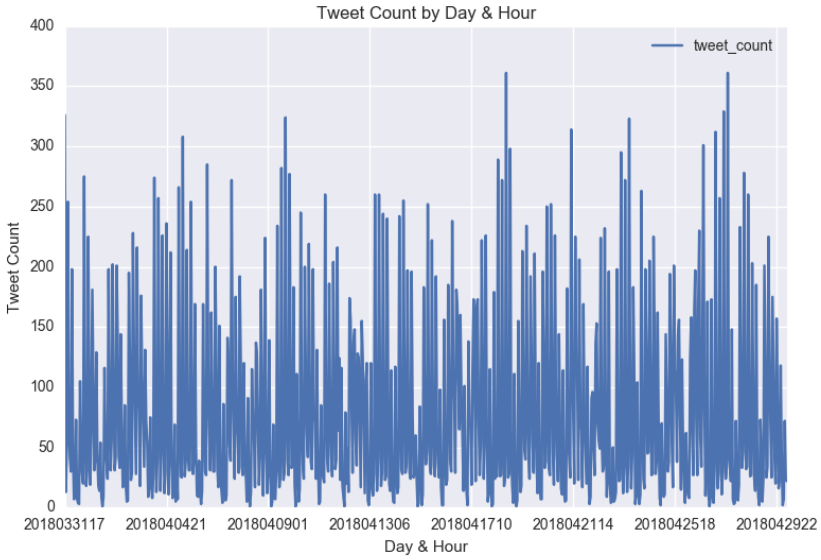




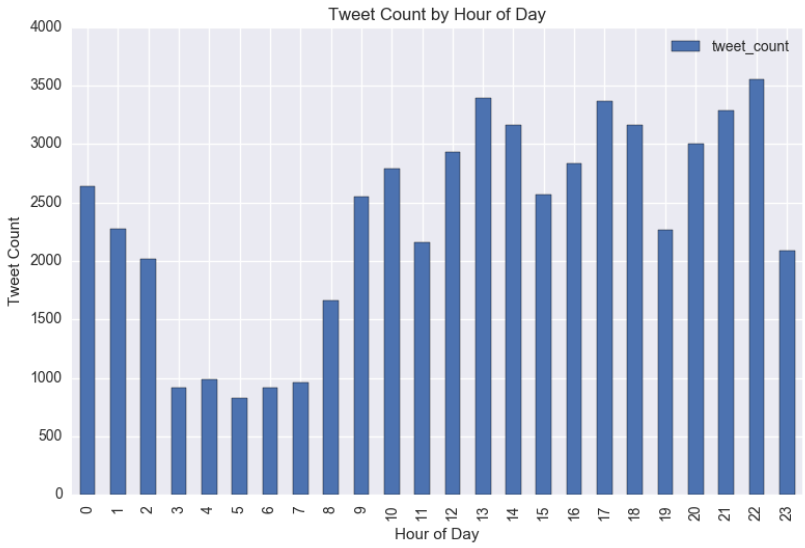
This project captured over 65,000 tweets for the specified geographic extent across 30 days. The figures below show the observed tweet capture by day, day-hour, day of week, and hour of day.



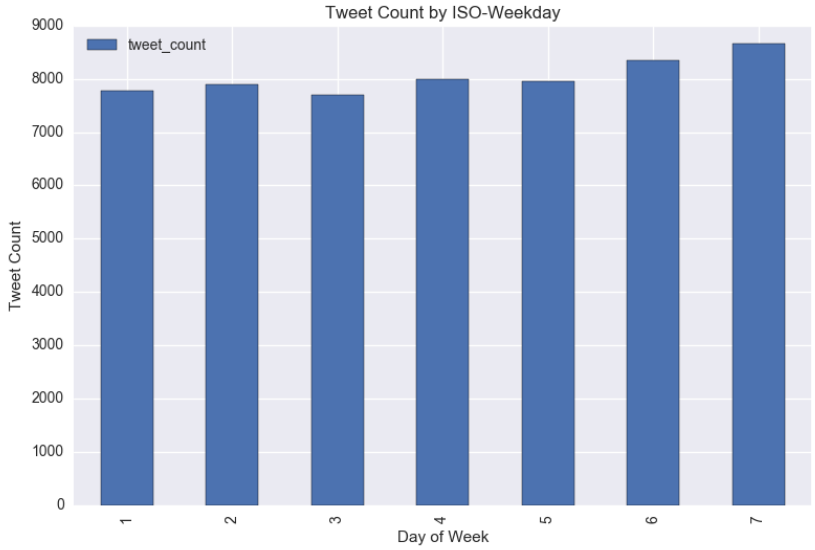
The figure above shows the captured tweets over the projects date range. We see that the average daily tweet capture 1,877 per day.



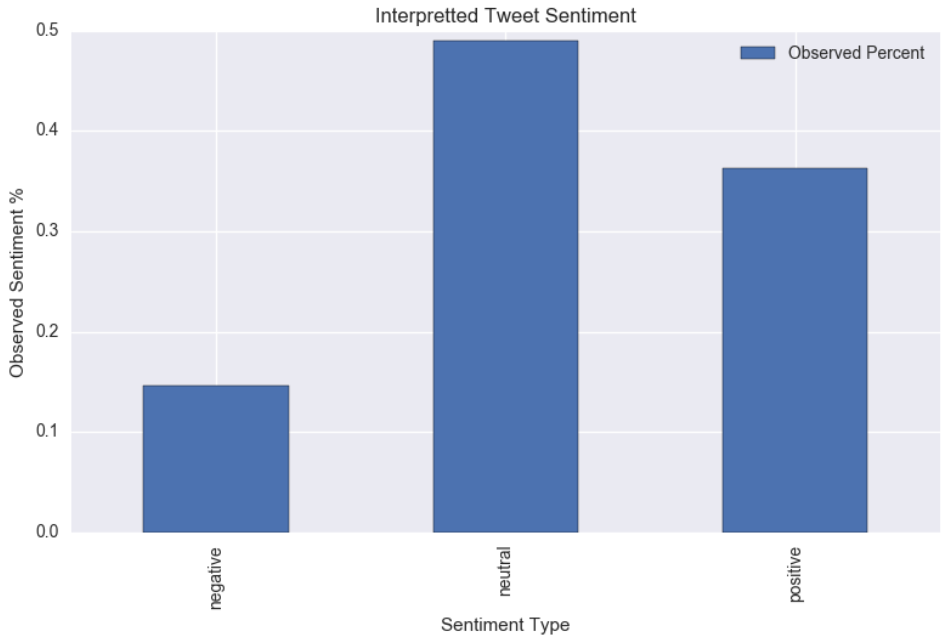
The image above shows the captured tweet count by of each day with an average capture of 78 tweets per hour.



The figure above shows the tweet capture by hour of the day. We see a sharp decline of activity with an average tweet capture of 921 between the hours of 0300 to 0700 that gradually increases to an hourly average of 2,346 between the hours of 0900 to 0200. The twitter activity appears to be greatest between the hours of 2000 - 2200.

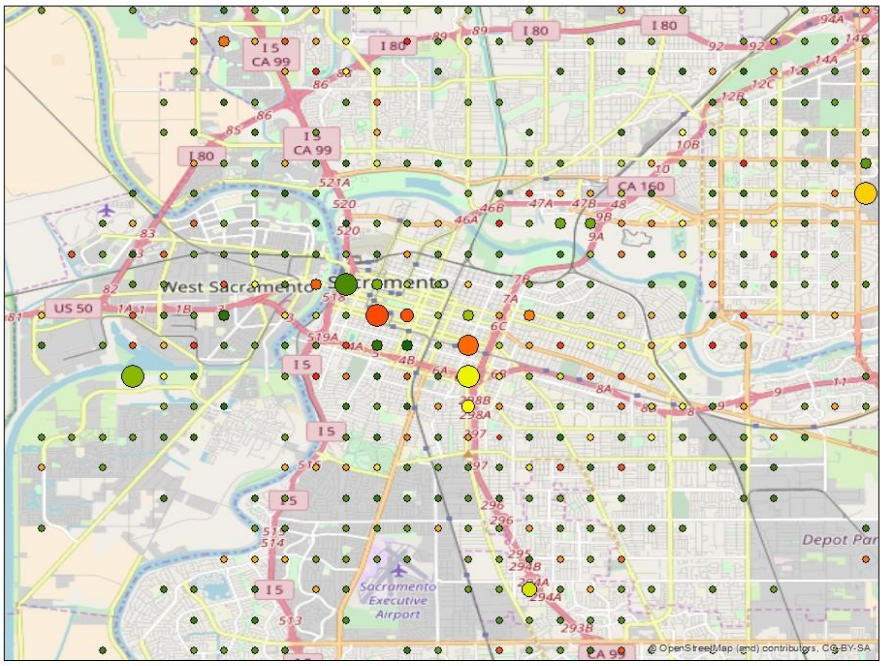


The figure above shows that the captured tweets by day of week are greatest on Saturday (6) and Sunday (7). One thing to note about this figure is that, for the month of April there are two additional days of Sunday and Monday in the calendar for 2018, and were not corrected or removed from the dataset.



The above figures shows the percent of interpreted sentiment by captured tweet. This figure details that the sentiment analysis could not definitively identify 49% of the captured tweets. The analysis did identify that 36% of the tweets were positive, and 15% were negative.

Sacramento Tweet Sentiment



The above figure shows the geographic area of focus with the sentiment values shown. Here the color scheme of red to green represents the negative to positive sentiment for the entire time range of study. Here we see the density of non-neutral sentiment tweets are centered in downtown Sacramento.

Purpose

The purpose of this project is to highlight a use case of ArcGIS's arcpy python module for geographic rendering of geospatially dependent information. While this project leverages an open source form of data, the approach to automating the rendering of geographic information to a map remains unchanged when applying to private datasets. As such, any individual may take the workflow detailed here (in the code) and reimplement the approach as they see fit. This project information is centered around how the use of geospatially dependent sentiment analysis elucidates publicized personal commentary across a geographic extent.

Geoprocessing

Specific areas of geoprocessing concerning ArcGIS's arcpy module are found within the arcpy_create_feature_class.py script. This script carries out a variety of tasks such as

- Environment declarations
- Feature class declarations
    - Field datatypes | Shape | Spatial referencing
- File/Feature management of existing items
- Inserting new records into the feature class
    - Nominal information determined during write process
- ArcMap document referencing for templates
    - There are three specific templates that are leveraged during this process.
    - All sentiment values: tweet_tamplate.mxd
    - Positive sentiment values: positive_template.mxd
    - Negative sentiment values: negative_template.mxd
- Definition Queries
    - These are leveraged for the positive and negative tweet templates
        - Implementation of a definition query in each template allows for both templates to reference the same feature class in the geodatabase and limit the types of features to be rendered in the template
- Symbology management
    - Due to the dynamic nature of the cumulative sentiment value for a given geographic extent the symbology has to be updated prior to exporting such that daily variations in sentiment value do not skew the visual representation
- Title management
    - As the overall process is intended to run on a daily basis and the information that is rendered onto the map is for a given time period, the title on the map needs to be updated through the management of LayoutElements

Project Challenges

One of the biggest project challenges was ensuring that the spatial referencing for the outside data source was correctly identified prior to feature class creation and management. There is little documentation for the geographic coordinate system that twitter allows developers to engage with. Another challenge was coming up with a workflow that maintained a compartmentalized data analysis pipeline. Specifically in reducing the number of feature classes to maintain during the map generation. This project went through multiple iterations of maintaining a feature class for each map, until the final approach of having three different map templates with varying definition queries was chosen. The approach of ensuring that the symbology was dynamic based upon the extremum values that were present within the feature class was as an unforeseen issue that had to be determined and corrected for. In addition to the aforementioned challenges, there were challenges in the use of custom geospatial analysis models located in personal toolboxes that would run in the interface but would not run when trying to engage the models programmatically. This challenge remains existent to the project at the time of the transcription of this summary, and is the root cause for the lack of an implementation of a kernel density raster layer.