Final Report

Using Google Searches to Predict Restaurant Guest Counts

Anthony Lusardi

Fall 2020

Overview and Motivation

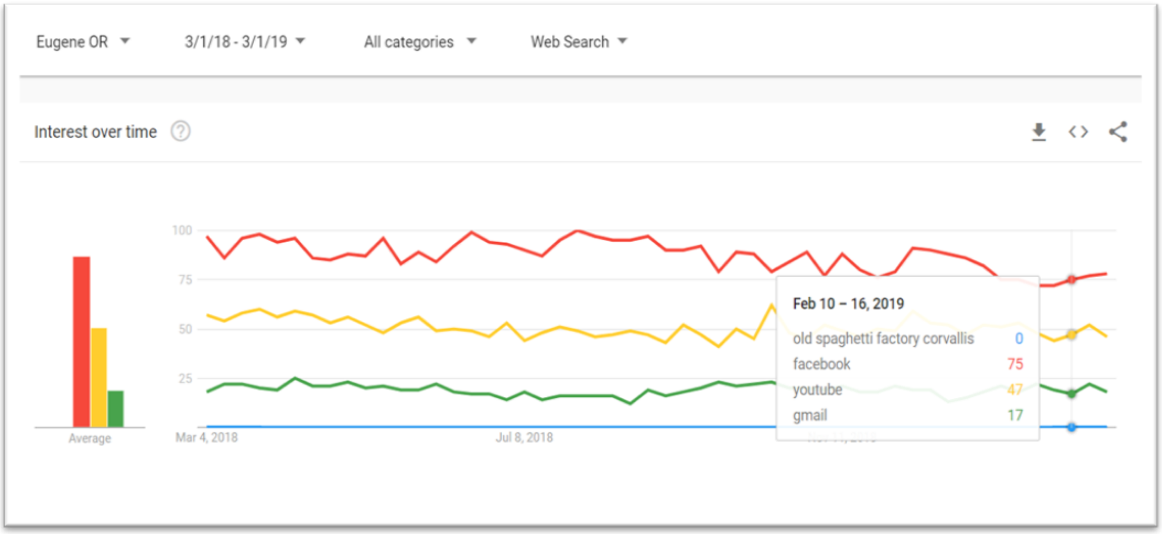
What I hoped to complete with this project is was use available data collected on google searches for a restaurant title to then generate a linear regression model to predict the average amount of guests expected to come in on a given day.

Initially it was a far-fetched idea however, I had found In a previous study in 2019, a researcher had been able to predict the forecasted visitors on opening night for a movie theater based on the google searches for the movie title in a geographic area. It was a simple idea that proved to be very successful for consulting a movie theater company in Germany.

Challenges

Collecting data from a Data Monopoly

The single most important element of this project is scaling search activity. Google only shares data that is scaled from 0 to 100 and relative to search activity within the area with no indication as to how many searches there actually are.(Fig 1) Therefore, I needed to create a scaling system between the search term and the constant search traffic terms called "anchor terms" eg:("facebook", "gmail").



Eugene OR ▾ 3/1/18 - 3/1/19 ▾ All categories ▾ Web Search ▾

Interest over time

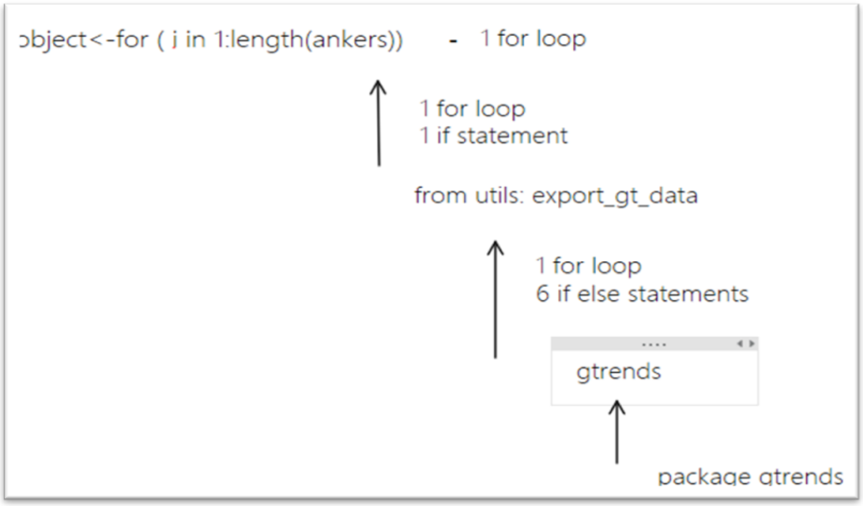
Feb 10 – 16, 2019

old spaghetti factory corvallis	0
facebook	75
youtube	47
gmail	17

Figure 1

Working with another person's work

At first look, I was very impressed by the amount of documentation available for this repository. In addition, it was all written in R which was a great advantage when working with expediting the code-writing process. However, after taking a look under the hood especially on the `utils` folder, I found that although it was well-documented, there were quite a few inefficiencies and some unneeded functions such as adjusting for a premiere movie night schedules in Germany and multiple nested for loops . An example of a nested `for` loop provided in (Fig 2)



Data gathering -preprocessing and data analysis

The dataset used I had complied with conjunction with the restaurant. It contains key data for recorded guest activity at the Old Spaghetti Factory (OSF) in Corvallis OR, between 01/01/2018 and 12/31/2019. (Fig 5).

Results - Searches and Guest Counts

After preprocessing and scaling the two highest weekly search terms "old spaghetti factory" and "spaghetti factory" from March 2018 to March 2019, we can saw that the relationship of guest counts and search activity in the area is not significant when fitted into a linear model.

Term	Coefficient	Std Error	t Ratio	P Value	Conf High	Conf Low
(Intercept)	1715.66	34.26	50.063	0.0000000	1782.83664	1648.49829
old spaghetti factory	24.21	43.52	0.556	0.5804951	109.51652	-61.08351
spaghetti factory	-20.557	43.514	-0.472	0.6387571	64.72984	-105.84490

In conjunction with the summary for this model, even with the most direct keywords to predict guest counts, there is no significant evidence to say that keywords predict guest behavior. However I did find that I found that seasonality and time dependency was an enormous factor in predicting guest counts by correlating weekdays with guest counts this would explain the non-independence of data which nullifies linear regression.

Learning Opportunities and Future Research

Coding Efficiency

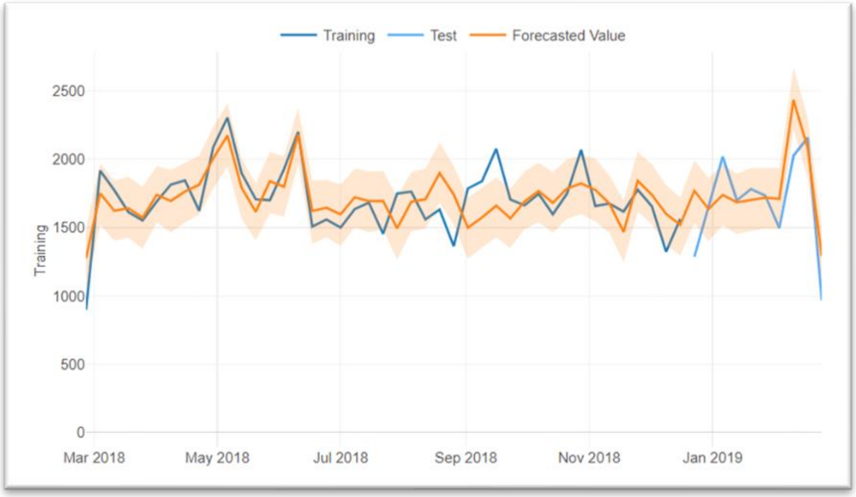
In order to make the code more efficient, I would need more time to really pick apart the nested for loops. Much of this could be improved with a total reconstruction of the repository. At the end of it I found that the only necessary functions could be summarized into 25% of the space. Fortunately, I had spent most of my time understanding repository with notes, which led to a greater understanding of the process.

Geographical Issues

In the future I would focus more time on finding better geographic areas to isolate search terms. The anchor terms that are used were taken from the entire state of OR as designated with the gtrends function. I could not find a way to isolate searches within the Eugene Area as indicated by google trends interface in (Fig1). An alternative could be to download the scaled title with every search term in .csv's to use the scale function to then isolate a more accurate geographic area.

Prediction

Consistent with the results, I found that seasonality and time dependency was an enormous factor in predicting guest counts. A GAM time series model was also used in the box office prediction. Initially, I had the idea to do a linear regression model because there was a high correlation of variables and that it was the simplest version of predicting data. However, I had fitted a preliminary GAM time series model named "Prophet" from Facebook and this proved to be quite accurate with a MAPE of 12.9% or 86.1% accuracy on predicting weekly guest counts based on previous data.



"There are no mistakes just learning opportunities" - Dad

References

Schmitz, M. (2019, December 6). *Using Google Trends data to leverage your predictive model*. Medium. https://towardsdatascience.com/using-google-trends-data-to-leverage-your-predictive-model-a56635355e3d