Data Analysis

Regions with lower broadband penetration rates In Google Search from 2008 to 2018

Hyerim Hwang

Table of Contents

Introduction

Google is the most commonly used search engine in the world across all platforms, “with 92.62% market share as of June 2019 handling more than 5.4 billion searches each day and 78,881 searches in 1 second" (Google Search Statistics). Since Google offers localized search engines to more than 188 regions, the Google search results were able to capture each region’s interests or big events over time. I’d like to see people searching keywords from each region in the world through the Google Trends website, it would be an efficient indicator for gaining their interests or trends over time.

However, China has the most Internet users in all countries, but Google Search Trends does not support Chinese data. In addition, Asia has the most Internet users of all continents[1], but also Japan uses the Yahoo Japan service the most and so does South Korea consuming the Naver the most, and China utilizing Baidu the most (Internet Stats & Facts for 2019). Since none of them use Google as a primary search engine, Google search users in those countries might apply other search engines more often.

I decided to rather look over the regions that have a lower broadband penetration rate than look overall regions that have other options to search for something online. The lower broadband penetration might be an indicator to have a lesser compatible market for the search engine, so it will lead me to more prominent results by using the Google search dataset to see those region’s overall trends.

[1] Asia has the most Internet users of all continents — accounting for 49 percent of all Internet users (down from about 50 percent in 2017 and up from about 48 percent mid-2018). Europe is a runner up with 16.8 percent of all Internet users.

Methodology

- Dataset 1: Internet Usage Statistics[2]
- Dataset 2: See what was trending in 2018 - India[3], See what was trending in 2018 - Pakistan[4], See what was trending in 2018 - Philippines[5]
- Tools: Selenium library in Python for collecting the data, csv file for storing the data, R studio for analyzing the data
- Variables: ranking with search keyword, grouping by each category, search volumes by subregion

To capture each region's dataset through the Google Trend website, I began to collect the data with rankings and keywords on each year's search keywords list gathering the keyword's href attribute as well. It is because I also need to see all of the search volumes by region that was located on each keyword's content page, so the proportion of search volumes and the name of regions also looked over and stored hopefully to answer my hypothesis, "Each region will have a different category of interest by its search keywords in Google, and also, neighboring countries might share similar interests in trends over time".

The Google Trend website has been stored the data related to Google search results through each country and analyzed the general trends by classifications per keywords with rankings since 2001. When I tried to figure each region's own indicators or trends over time, it was impossible to filter these datasets out at a glance. I also got more interests to see if there are common interests or events between the neighboring regions.

[2] https://www.internetworldstats.com/stats.htm
[3] https://trends.google.com/trends/yis/2018/IN/
[4] https://trends.google.com/trends/yis/2018/PK/
[5] https://trends.google.com/trends/yis/2018/PH/

The keyword datasets from 2008 to 2018 were needed to be web scraping so I use the Selenium library in Python which allows me to click on “show 5 more” buttons by an automatic bot based on my code. To analyze the data based on the robust basement, I built the MongoDB cluster and database to store each year’s keyword and other information on each collection. Inside of the collection, I stored the data of each keyword including the data of ranking in each year, search keyword grouping by category, search volumes by region.

Right now, I’m having an issue in web scraping (Figure. 1), so I couldn’t store the data to csv file yet, but I think I collected datasets successfully related to all variables that I need.

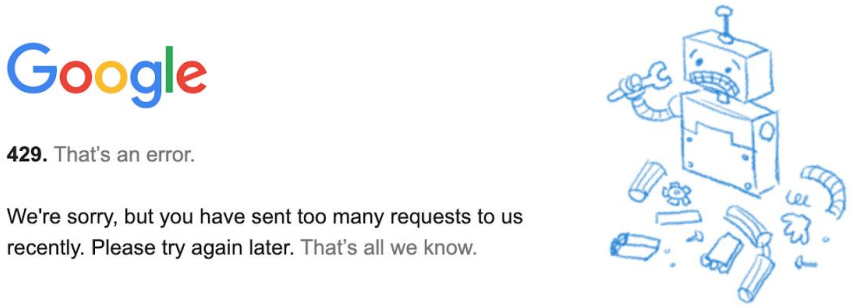

Figure. 1

I will grab and store the dataset with this structure below (Figure. 2 and Figure. 3). I will make sure all keywords do not overwrap with each other, so I will stem the keywords with transforming into lowercase. Then, I will group by each region within the same amount of year and aggregate them with keyword ordering by search volumes. Since search volumes already measured by the proportion of total searches in the location, I don’t need to find the number of population of the region to make the frequency of the searches normalization.

Global Trends in 2008 ~ 2018

Google Trends over Time

YEAR	CATEGORY 1	RANKING 2	KEYWORD 3	INTEREST BY REGION	SEARCH VOLUMES
2008	Campaign Buzzwords	1	Joe The Plumber	United States	100
2008	Campaign Buzzwords	2	Jeremiah Wright	United States	100
2008	Campaign Buzzwords	3	Maverick	United States	100
2008	Campaign Buzzwords	3	Maverick	Brazil	94
2008	Campaign Buzzwords	3	Maverick	Philippines	69
2008	Campaign Buzzwords	3	Maverick	South Africa	64
2008	Campaign Buzzwords	3	Maverick	Canada	53
2008	Campaign Buzzwords	4	William Ayers	United States	100
2008	Campaign Buzzwords	5	Bridge To Nowhere	New Zealand	100
2008	Campaign Buzzwords	5	Bridge To Nowhere	United States	69
2008	Campaign Buzzwords	5	Bridge To Nowhere	Canada	51

Figure. 2: The estimated dataset that will be stored in a csv file collecting by Selenium.

Figure. 3: The original Google Trend website display

So my plan is to find out the variables' relationship and probability to see the general trends.

1. To find out the relationship between the ranking and keyword over time, or each category's number or contents over time

2. To find out which keyword shows up the most over time
3. To find out which region reveals to the list the most over time
4. To find out which category lists up the most over time
5. To build a new list that contains each region's keywords by higher search volumes, grouping the keywords data by each region and year
6. To find out the probability of each region's trends in 2019 based on what I found

Findings

Conclusion

Appendix

Include your code in the appendix.

References

Google searches in 1 second, Retrieved from:
https://www.internetlivestats.com/one-second/#google-band

Internet Usage Statistics, Retrieved from:
https://www.internetworldstats.com/stats.htm

HostingFacts Team. Dec 17, 2018, Internet Stats & Facts for 2019. Retrieved from:
https://hostingfacts.com/internet-facts-stats/