Data Analysis

Regions with lower broadband penetration rates In Google Search Trends

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.

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]
  - Variables: % Population of Internet Penetration, % Users Facebook 31-Dec-2017
- Dataset 2: India[3], Pakistan[4], Philippines[5]
  - Variables: ranking, search keyword, category, search volumes by subregion
- Tools: Selenium in Python - collecting the data, MongoDB - storing the data, R studio - analyzing the data

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 trends over time, it was not easy to filter these datasets out at a glance. Also, I got an interest to see the countries which might have lesser compatible search engine markets online, so I needed to research the dataset to combine with the Google Trend dataset.

To capture the region's broadband penetration rates, the internet world statistics dataset is available based on the research of 4,536,248,808 internet users on Jun 30, 2019. I was able to get the dataset below (Figure. 1) that contains each country's broadband penetration rates which are especially lower than the average penetration rates in the region. I added the country naming list of Google Trends that is able to collect regions' data from it. The ideal target countries came out with the name of India, Pakistan and Philippines and the rates of internet penetration.

[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/

```
under_mean_by_internet_penetration = [{'region' : 'Oceania', 'under_mean_country': {'American Samoa': '43.1%', 'Chrismas Island': '45.
                                      {'region' : 'Africa', 'under_mean_country': {'Angola': '22.3%', 'Benin': '32.2%', 'Burkina Faso'
                                      {'region' : 'Middle East', 'under_mean_country': {'Yemen':'26.6%', 'Iraq':'49.4%', 'Palestine (S
                                      {'region' : 'Asia', 'under_mean_country': {'Afganistan': '19.7%', 'Bhutan': '48.1%', 'Cambodia':

country_google = ['Argentina','Australia','Austria','Bangladesh','Belarus','Belgium','Brazil','Bulgaria','Canada',
                  'Chile','Colombia','Costa Rica','Croatia','Czechia','Denmark','Egypt','Estonia','Finland','France',
                  'Germany','Greece','Guatemala','Hong Kong','Hungary','India','Indonesia','Ireland','Israel','Italy',
                  'Japan','Kazakhstan','Kenya','Latvia','Lithuania','Malaysia','Mexico','Netherlands','New Zealand',
                  'Nigeria','Norway','Pakistan','Panama','Peru','Philippines','Poland','Portugal','Puerto Rico','Romania',
                  'Russia','Saudi Arabia','Serbia','Singapore','Slovakia','Slovenia','South Africa','South Korea','Spain',
                  'Sweden','Switzerland','Taiwan','Thailand','Turkey','Ukraine','United Arab Emirates','United Kingdom','United States
]

for region in under_mean_by_internet_penetration:
    lists = []
    for country in country_google:
        dicts ={}
        if country in region['under_mean_country'].keys():
            dicts['region'] = region['region']
            dicts['under_mean_country'] = country
            dicts['Penetration_Population'] = region['under_mean_country'][country]
            dicts['mean_of_the region'] = region['mean']
            lists.append(dicts)
print(lists)

lists = [{'region': 'Asia', 'under_mean_country': 'India', 'Penetration_Population': '40.9%', 'mean_of_the region': '54.2%'},
         {'region': 'Asia', 'under_mean_country': 'Pakistan', 'Penetration_Population': '35.0%', 'mean_of_the region': '54.2%'},
         {'region': 'Asia', 'under_mean_country': 'Philippines', 'Penetration_Population': '3.4%', 'mean_of_the region': '54.2%'}]
```

Figure. 1: Combining two dataset to get the target countries with lower broadband penetration rates.

I began to collect the data in Google Trned with each year's search keywords and ranking. It also required me to gather the keyword's href attribute as well, because I also need to see all of the search volumes by subregion that was located on each keyword's content page (Figure. 2).

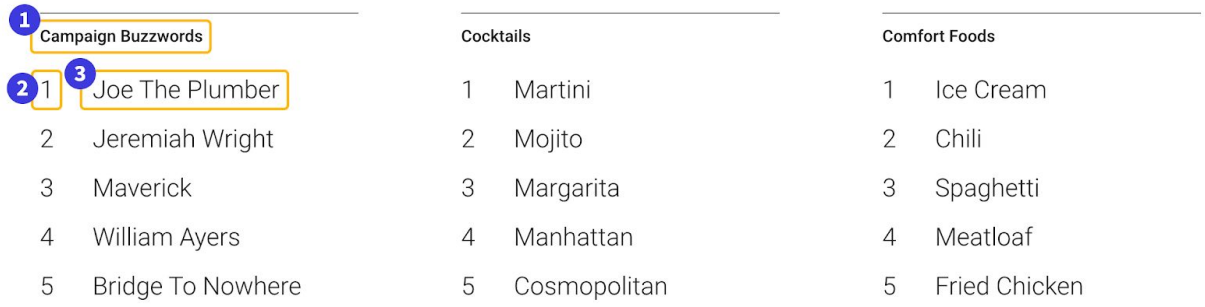

Figure. 2: The screen of Google Trend website display.
1 - category, 2 - ranking, 3 - keyword

The datasets relating to keyword were needed to be web scraping, so the Selenium library in Python was selected 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. 3), 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.

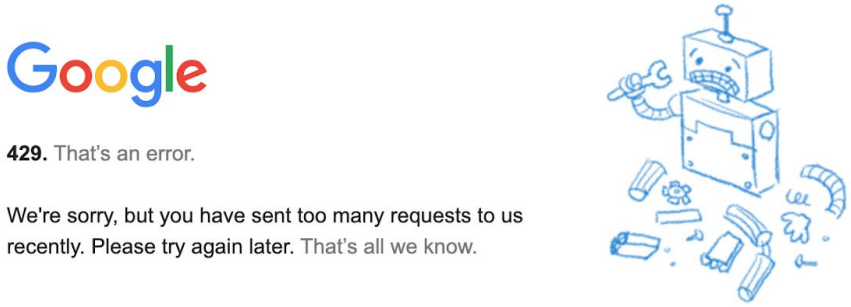

Figure. 3

So I will plan to find out the relationships and probabilities between each variable that I have after successfully storing the dataset.

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

Before I jump into analyzing the data, 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 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.

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/