



Phase 2: Selecting A Research Question & A Dataset

Team Report by Team OLAPPED CSCI 6401-01

Team Name :- OLAPPED

Course :- Data Mining

Instructor :- Prof. Shivanjali Khare

Course ID :- CSCI-6401-01

Session :- Fall 2023

Assignment :- Phase 2: Selecting A Research Question & A Dataset

1. Team Name :- OLAPPED

Team Member Names :-

- (1) Sean Vargas – svarg1@unh.newhaven.edu
- (2) Kaylie N Neal – kneal5@unh.newhaven.edu
- (3) Prashant Rana – prana4@unh.newhaven.edu
- (4) Rajdeep Bhattacharya – rbhat6@unh.newhaven.edu

2. Research Question :-

The Research Question: - How can the patterns and trends in the **Google Search Engine Queries Dataset** be used to identify better rules for finding the most promising **Keywords (Search Queries)** and **Topics (Categories)** for showing more relevant **Search Results** in the future and targeting the relevant **SERPs (Search Engine Result Pages)** for **Ad Suggestions** by **timings (Seasonality)** and **locations (Geography)** for **Customer/Viewer Satisfaction** and higher **Ad Revenues**.

3. Support for the merit of answering the research question using literature review or other valid references (min 2, max 5) :-

Reference 1. (From Google's side of view)

Heading :- How Google's \$150 billion advertising business works

Dated :- May 18, 2021

Link :-

<https://www.cnbc.com/2021/05/18/how-does-google-make-money-advertising-business-breakdown-.html>

An excerpt :-

Search is Google's most lucrative unit. In 2020, the company generated \$104 billion in "search and other" revenues, making up 71% of Google's ad revenue and 57% of Alphabet's total revenue.

That "search and other" figure includes revenue generated on Google's search properties, along with ads on other Google-owned properties like Gmail, Maps and the Google Play app store.

Advertisers using Google products can bid on search keywords — specific words and phrases that lead their ads to show up to **relevant users in search results**.

Reference 2. (From customers/viewers side of view)

Heading :- How using related searches on Google helps you boost your SEO

Dated :- December 21, 2017

Link :-

<https://www.smartinsights.com/search-engine-optimisation-seo/seo-strategy/using-related-searches-google-helps-boost-seo/>

An excerpt :-

However, understanding a user's intent sounds impossible at first. How are we supposed to know what people are *intending* to do?

This is where **Google's related searches** come in, which are the eight search results at the bottom of a results page. Using related searches, you can gain sharper customer insights so

that you can create relevant and valuable content that they and Google love, thus boosting your SEO.

Reference 3. (From Google's point of view)

Heading :- How our business works

Dated :- N.A.

Link :-

<https://about.google/how-our-business-works/#:~:text=Ultimately%2C%20we%20earn%20most%20of,we%20make%20money%20with%20advertising.>

An excerpt :-

So how does advertising at Google work? We make money selling ad space to businesses -- big and small, global and local -- in two key ways. First, businesses can reach potential customers by showing ads on a range of Google products such as Search, Maps, and YouTube.

Second, businesses can buy ad space that we show on sites and apps that partner with us, like news publications and blogs. In this case, most of the money goes to the partner and helps fund their content. So ads not only help support Google but also many other websites and creators.

Ultimately, we earn most of our money by showing ads alongside **relevant Search results on Google.com**. If you're interested, you can learn more about how we make money with advertising.

Important Key Words in the above-mentioned excerpts:-

From the excerpt of Reference 1 :- relevant users in search results

From the excerpt of Reference 2 :- Google's related searches

From the excerpt of Reference 3 :- relevant Search results on Google.com

Our Research Question takes hints from the idea of "Relevancy" presented in the above listed key words.

4. About our selected dataset :-

Description of the selected dataset that we want to work with –

The Data Set is named “trends” as it represents the Google Search Trends for a period of 20 years (2001 - 2020), pointing out the [Top 5 Google Searches (Search Queries)] by [Categories] with their [Global Ranks] and the ranks for the [Top Countries] by [Year].

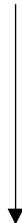
It has 5 Attributes (represented by 5 columns) namely **location, year, category, rank and query**. And 26956 Data Points (represented by 26956 rows).

4. I. The accessibility of the selected dataset :-

The Dataset is accessible online. We downloaded it.

A. We searched on Google Dataset Search.

<https://datasetsearch.research.google.com/>



B. We clicked one of the Search Results.


<https://datasetsearch.research.google.com/search?src=0&query=Google%20Search%20Data%20Kaggle&docid=L2cvMTFzc3E0a2QyZA%3D%3D>



C. We clicked that Search Result and navigated to the required Dataset Page.

<https://www.kaggle.com/datasets/dhruvildave/google-trends-dataset>

It was posted 3 years back by Dhruvil Dave.

 DHRUVIL DAVE · UPDATED 3 YEARS AGO

91New NotebookDownload (343 kB)

Google Trends Dataset

A dataset of all the Google Trends all over the world

Data CardCode (9)Discussion (0)

About Dataset

This is a curated dataset of Google Trends over the years. Every year, Google releases the trending search queries all over the world in various categories. It has trends from 2001 to 2020.

Image Credits: [Unsplash - lukecheeser](#)

Usability

10.00

License

ODC Attribution License (ODC-...)

Expected update frequency


Annually


Tags

EducationInternetTabularProgramming

4. II. Data collection methods :-

It is a curated dataset of Google Trends over the years, Google released as it releases the trending queries all over the world in various categories. This dataset has trends from 2001 to 2020. It also has the ODC (Open Data Commons) Attribution License.

 Open Knowledge Foundation

 Open Knowledge

Open Data Commons

LEGAL TOOLS FOR OPEN DATA

Open Data Commons Attribution License (ODC-By) Summary

This is a human-readable summary of the [ODC-BY 1.0 license](#). Please see the disclaimer below.

You are free:

- *To share*: To copy, distribute and use the database.
- *To create*: To produce works from the database.
- *To adapt*: To modify, transform and build upon the database.

As long as you:

- *Attribute*: You must attribute any public use of the database, or works produced from the database, in the manner specified in the license. For any use or redistribution of the database, or works produced from it, you must make clear to others the license of the database and keep intact any notices on the original database.

The License also reads :-

Disclaimer

This is not a license. It is simply a handy reference for understanding the [ODC-BY 1.0](#) — it is a human-readable expression of some of its key terms. This document has no legal value, and its contents do not appear in the actual license. Read the [full ODC-BY 1.0 license text](#) for the exact terms that apply.

4. III. Information about the data type :-

The Google Trends Dataset has 26955 Data Points, out of which 18431 are unique values. It has 5 fields/attributes represented in the 5 columns namely location, year, category, rank and query.

We would like to find the correlation between the unique queries, ranks, categories and locations over time.

Along with the Geography (denoted by the location) the time value (denoted by the year) is very important, it will give us a proper estimation of the top ranked searches by location and how they evolved over time.

We got the dataset in the **CSV Format**.

The screenshot displays the Google Trends Dataset interface. At the top, there's a header with 'Google Trends Dataset' and navigation options: 'Data Card', 'Code (9)', and 'Discussion (0)'. To the right, there are buttons for 'New Notebook' and 'Download (343 kB)'. Below the header, the main content area shows the file 'trends.csv (1.31 MB)' with a download icon. To the right of the file name, there's a 'Data Explorer' section showing 'Version 1 (1.31 MB)' and a list of files including 'trends.csv'. Below this, there's a table titled 'About this file' with the subtitle 'Database file'. The table has 5 columns: 'location', 'year', 'category', 'rank', and 'query'. The 'location' column lists 'United States' (8%), 'Global' (4%), and 'Other (23750)' (88%). The 'year' column shows '26955 total values'. The 'category' column lists 'People' (3%), 'Searches' (2%), and 'Other (25575)' (95%). The 'rank' column shows a range from 1 to 5. The 'query' column shows '18431 unique values'.

location	year	category	rank	query
Country	Year of trend	Category	Rank in that particular year, category and country	Trending term
United States 8%	26955 total values	People 3%	1 5	18431 unique values
Global 4%		Searches 2%		
Other (23750) 88%		Other (25575) 95%		

5. Plan [steps, data mining model/algorithm] for solving the research question using the selected dataset :-

We will use the **Apriori Data Mining Algorithm** that is available in the **Weka Tool**.

We will use the **Explorations Tab** in the **Weka Tool** to find suitable **Support** and **Confidence** values to arrive at an explainable **Correlation**. That will give us the **Rules** that we are searching for.

For example :-

Selecting suitable **Support** and **Confidence** with number of iterations for finding **Correlation** between **Location - Search Query**, **Location - Category**, **Category - Search Query**, **Location - Year**, etc. Here the **Rank** of the Search Query wouldn't be of much help individually as it is the first 5 ranks only and not the overall rank, so we will form a **Super Attribute** by joining it with Year and Category and then plotting it against the Search Query to get a Correlation.

These Correlations will help us formulate rules to find the best possible rules for finding the most promising **Keywords (Search Queries)** and **Topics (Categories)** for showing more relevant **Search Results** in the future and targeting the relevant **SERPs (Search Engine Result Pages)** for **Ad Suggestions** by **timings (Seasonality)** and **locations (Geography)** for **Customer/Viewer Satisfaction** and higher **Ad Revenues**.
