Project – team 25

1. Introduction

research question: Is there variation in the ranking for scientific terms on Google across countries over time?

Specifically, the research seeks to address the following sub-questions:

- is there any overlap between result searches in similar languages for scientific terms on Google without considering the ranking?
- if there were an overlap, is there an overlap between the specific ranking? If not, what is the variation?

this research is important as it sheds light on the consistency and accessibility of searches of scientific terms across different countries over time. Understanding the variations in the ranking can provide insights into the factors influencing search rankings.

In general, ensuring consistent search for scientific terms over time is of paramount importance. The expectation is that users conducting searches on Google for scientific topics should receive reliable and relevant information consistently, regardless of the geographical location or the specific time of the search.
link to the Rmd - https://drive.google.com/drive/u/1/folders/1gezBw2ElIG_LhPjarVOiIjlwwrTsi9lK

2. Data explority

link to the dataset - https://docs.google.com/spreadsheets/d/1EXv_HItjN6xKb32XF-0uK-w0D3qAbtzH/edit?usp=drive_link&ouid=101402361537178560805&rtpof=true&sd=true
Our research focused on analyzing changes in the ranking of searches of Google over time. We used data collected by searching for scientific concepts in various countries, languages, and on different dates. Among the 61 features in our dataset, we identified several key ones for our research. The date of collection (8) allowed us to organize the data chronologically. The link (10) represented the specific webpage we analyzed, while the result number (9) indicated its ranking on Google page.
additional explanation is in the proposal appendix.

3. Methods and Results

STEP 1: Data Preprocessing:

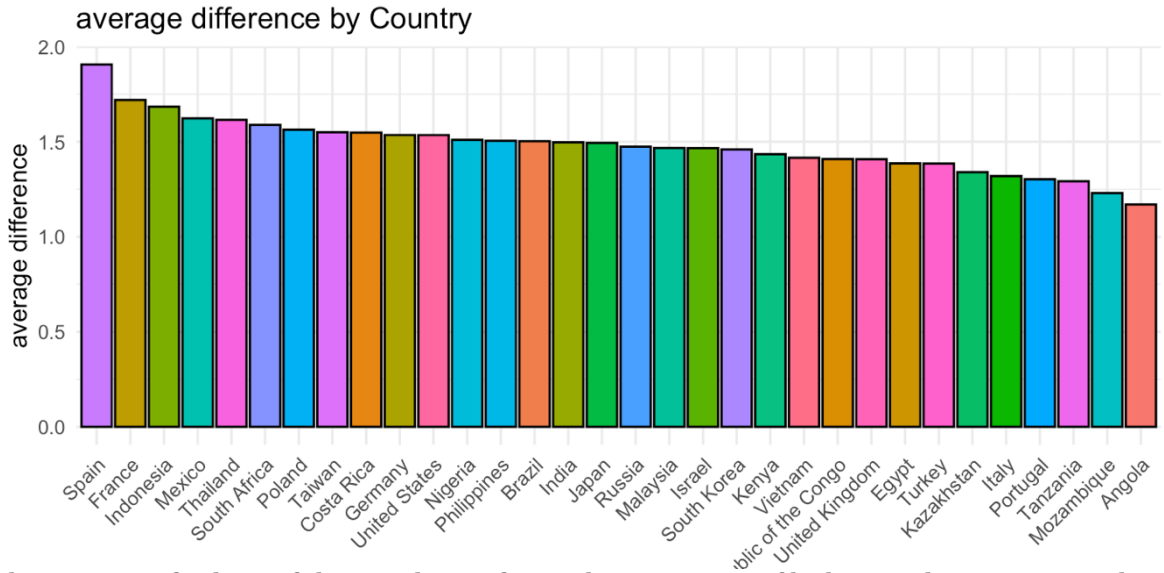
first, we expanded the dataset by incorporating data extracted from HTML using a Python code. This allowed us to gather more comprehensive information for our analysis. that's because we found that the given dataset was insufficient to provide a comprehensive analysis of the variations. Additionally, to enhance dataset readability and understandably, we opted to replace complex variable names with more generic terms, improving data interpretation and analysis.

STEP 2: Overlap Analysis:

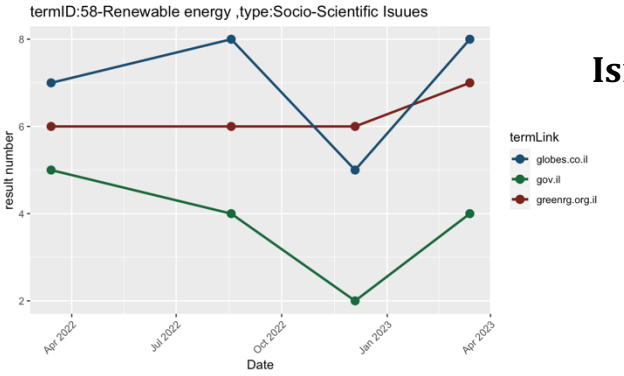
To enhance the dataset and focus on relevant information, we employed a Python script to create the "matching links" table. This table was generated by including only those links that demonstrated an overlap in search results over time. By filtering out irrelevant links, we aimed to ensure that our analysis focused on the most pertinent data points for investigating the variations in search results for scientific terms across countries.

STEP 3: Variation Analysis:

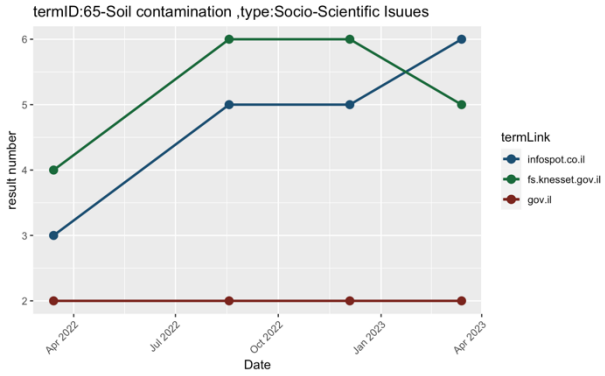
We focused on the result number feature (9) and their variations. We calculated the average difference between consecutive result numbers for each country. This involved sorting the data by date, splitting it based on unique search queries and calculates the average differences in the ranking of Google search results over time for each country, indicating the relative change in position. The following plot summarize this relative change:

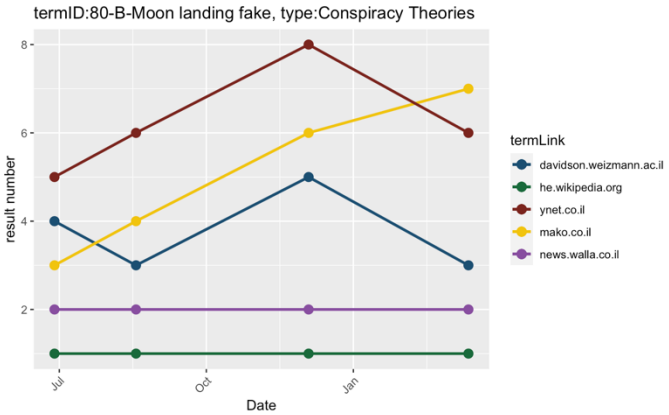


The primary finding of this study confirms the presence of links overlap over time, but to answer the question is there a noticeable variations in the changes observed across different countries we'll need to use a robust statistical methodologies to ascertain the presence of significant differences. trying to perform a T- test, the standard error estimates returned nan. we can see the variance of the average difference is 0.0213, when there is insufficient variability, it becomes challenging for the estimate the standard error accurately.



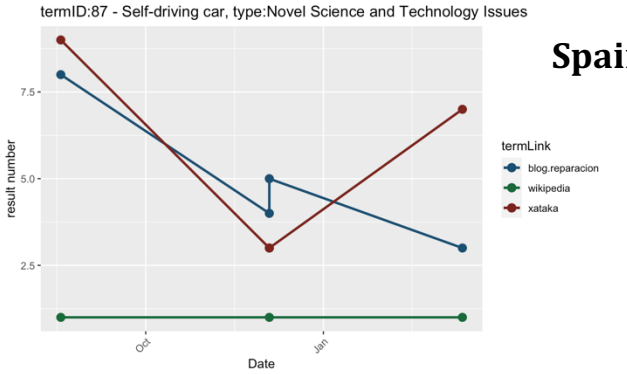
Israel

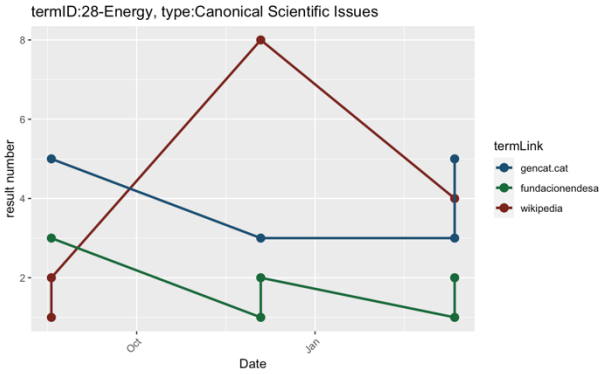


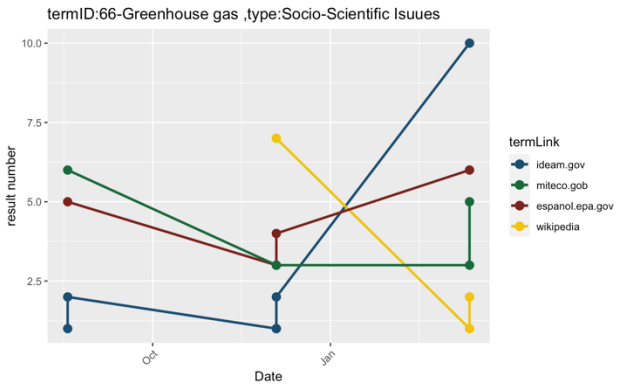


This plots shows the change in the result search over time. Each plot represents a chosen term. We can see which sites are more stable, like Wikipedia and which are not. within more time we could categorize the site types and see if the is a non-circumstantial relationship between the change in the result search and the site type.

Spain







In comparison to the Israel , Spain has significant changes in the plotted data. They indicate sudden increases or decreases in the result search change, what reflecting a more pronounced and noticeable shift in the trend. notice that this observed pattern in the data, aligns with the calculated average difference results. Notably, Spain exhibits the most substantial change among the countries analyzed, but it is still hard to determine the relationship between the country and the result search. Note that this plots represent individual cases and do not represent a trend in the entire data.

4.Limitations and Future Work

In our approach, we aimed to investigate the consistency of search results quality and accessibility for scientific terms across different countries over time. Based on the available time and data, we concluded that by examining the variation in the number of search results from different websites over time and across countries, we could gain insights into the consistency of information.

However, there are limitations to our approach. For one, it is important to note that if a high-quality and accessible website moves down in the search results, it does not necessarily indicate a failure in providing reliable information. This is because a new site appearing in that search result position could also be a reputable source of information.

Given additional time for our project, we would address the aforementioned limitations. In addition, we could involve scraping more internet pages and develop automation that will fill the rest of the features and developing models to predict the quality and accessibility of a given site. By employing these models, we could gather a more extensive dataset, enabling us to perform a more precise statistical analysis. Additionally, our objective would be to delve into the factors that underlie the disparities witnessed in the average difference index across diverse countries. Specifically, we aim to gain a comprehensive understanding of the factors responsible for extreme variations, exemplified by the contrasting results observed between Israel and Spain.

examining the variation in search results for scientific terms across different countries and over time can provide insights into how the Google algorithm works. It can shed light on the factors that influence the ranking and presentation of search results, such as language preferences, location-based relevance, user behavior, and temporal dynamics of information indexing. By studying these variations, researchers can gain a better understanding of how Google's algorithm processes and delivers search results.

Appendix

plots choosing -

Although anecdotal we wanted to see the differences overtime of different links of the same term.
for example the result number of a wiki page stays consistent on a certain topic while a different site on the same topic might be all over the place.
In the code I first got sub data frames.
Each sub data frame is made out of rows that have a certain unique term_id.
Next for each sub data frame
We again get sub data frames based on the links.
For Each sub dataframe of the current sub dataframe we plot with the date collection on the x axis and the result number on the y axis.
We only plot dataframe that we thought had enough observations.
We could then go through the different plots like a slide show and pick the ones that were relevant.

scraping -

getting more data from the html -
the code is very messy and unclear because it is a single use code. meaning once I ran it it was not relevant. (plus I did it at 1:00 - 4:00 in the morning)
so here is an explanation for how it was done:
in the original data the amount of similar links that were googled in different times was limited.
This meant that we couldn't analyze the changes over time.
so we needed more data which we had in the form of html google searches.

we concluded that we had 8 relevant features to get

"feature0": "term_id", -- Id of the term

"feature1": "langs", -- language

"feature2": "country",

"feature3": "term", -- query

"feature4": "type", -- what subject does the query relate to

"feature8": "Collection date",

"feature9": "Result number",

"feature10": "Link".

we had 3 directories containing a large amount of google search results as html files.

And now for the code.

The idea is to take a directory and iterate over the html files.

For each html file we iterate over the Google result and take the sites where the link isn't corrupted.

If the link isn't corrupted we add it as a "row", with where we saw it in the search result. Or in other words the iteration number.

If it's corrupted we move on to the next site in the result.

The rest of the features were obtained from the html file name. (Other than the date which only changes depending on the directory)

The naming format of the html files contained the term Id, location and language.

Through which I was able to get some of the rest of the features by creating dictionaries that connect the "term Id" to the the "type"

And the "term Id" + "langs" to the "term".

I'm saying some because the dictionaries were created based on the existing data we had. And some terms that were searched in the html weren't included in the main dataset.

Note that I could have gotten the "term" through scraping the query itself and I honestly don't remember why I decided to not do it.

But it doesn't matter because the term_id was more relevant.

After finishing going through the directory I transformed the list of tuples into a data frame and made a CSV file out of it.

This was repeated on all 3 directories.

I then read all the CSV files and added them to the original dataset.

Cleaned duplicates if there were any and thus we created the new dataset.

This proved essential to our project because we could immediately see multiple links showing up at different times or even the same exact link showing up at the same time but with a different result number.

This was weird at first but we verified that it was possible that the same exact link will show up twice or more on the same google search result.

overlap links:

This code snippet reads data from a CSV file and performs comparisons on links based on certain criteria. Here's a short explanation of the code and its purpose:

Imports: The code imports the necessary libraries, including csv for reading CSV files, tldextract for extracting domain information from links, and pandas for data manipulation.

Function: The compare_links function compares two links based on their domain and subdomain, returning True if they match.

Reading the CSV: The code reads the relevant data from a CSV file named 'relevent_data.csv' and stores it in a dictionary called data. It extracts information such as value names, links, dates, and countries for each row in the CSV.

Comparing Links: The code iterates over the collected data and compares links. If multiple dates are associated with a link, a result is generated, including information about the link, countries, dates, value names, and comparisons between different value names.

Exporting Results: The code creates a DataFrame using the collected results and exports it to a CSV file named 'results3.csv' using pd.DataFrame and the to_csv function.

The purpose of this code is to analyze the provided data, identify links with multiple dates, and compare them based on their value names. The code generates a result table that can be used for further analysis or presentation purposes.