Final report Team – O

Omer Gazit 318254026 and Ido Bouhnik 206586794

Repository link: https://github.com/omergazit2/searching_for_science

Introduction:

The research question we will discuss in this project is whether the quality of Google search results of scientific terms is differs between different languages. And if so, is there a language barrier that slow down the arrival of newer and more contemporary concepts?

Previous studies have shown that there is a gap between the quality and quantity of search results between different languages. These studies used English as a benchmark to compare the quality of other language content to it.

In this project, we are interested in reexamining this question with respect to basic scientific terms and concepts (e.g. Atom) and compare them with newer and more contemporary concepts (e.g. climate change).

We have collected Search Results pages of selected basic scientific terms in four languages (English, Russian, Hebrew, and Arabic). And would like to understand the quality of content offered in each of those languages.

Motivation - Better understanding the impact of language on the quality and relevance of the content we consume and access to. Contemporary issues, such as COVID-19 and climate change, may be of special interest due to their importance for policymaking and individual decision-making.

The **biggest obstacle** we encountered during the project is the question of how do we decide what is quality content? Is it by the form the content is presented? its quantity? how easy is it to be found?

On top of that, we encountered another problem when trying to collect data. The difference between websites is large and larger between languages, performing scraping on hundreds of different websites accurately is very complicated.

Considering these issues, we decided to focus our research on two fronts:

First, we explored the difference in quality of the results of different languages by manual ranking the results. In order to do this, we collected for each term and language the first three websites in Google searches and ranked them according to a pre-defined ranking method.

Second, we decided to focus on Wikipedia, the largest web-based open encyclopedia covering more than 300 languages to make a comparison between similar sites where the content should be similar.

Data overview:

First, we used SerpAPI tool to gather for each term its Google result first page in every one of the four languages. This tool has also allowed us to set a location from which the search will be made, for each language we have chosen a country where it is the national language to ensure the most quality results.

From this primary data collecting we have created **“result.csv”** that contain information of the search as: term ID language, and country of the search.

We also collected information about the results: number of results, position in the page order, and the link for the website.

For each search, we ranked the first three results according to a ranking method we built to determine site quality without going into the depth of site content analysis, which we had a hard time doing so without different language speakers. **“top3_edited.csv”**

We used the next petameters for the rank:

- Presence of a clear definition.
- information – quantity of related information the site contains.
- visualization – illustration related to the term.
- Site quality – website look professional.
- Adds – site contains advertisement.

Further details on the raking method at the **read me** file in the GitHub repository.

From Wikipedia we created “Wikipedia_DF” we collected for each term the next values:

- Number of visualizations per page.
- Number of sections (sub-topics) the content is divided for.
- Number of references (link for external pages).
- Text length (number of words).

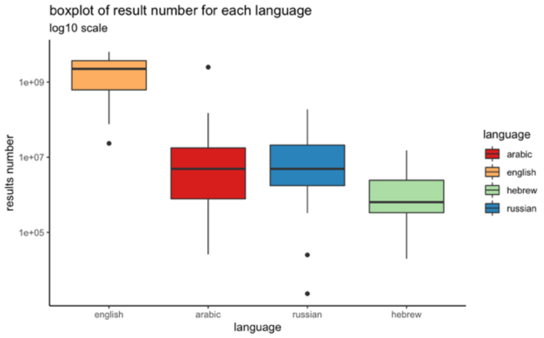
Those parameters, we believe, indicate about the **quantity of information**.

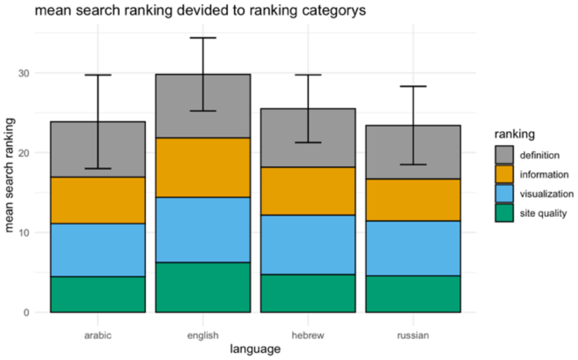
Methods and results:

First, we have noticed the big gaps between the **result number** from google search. We preformed multiple T-tests to Determine statistical difference between them.

Then, we divided the results into two groups, contemporary and non-contemporary.

We normalized them in relation to the number of results in English (on the assumption that this is the language with the richest amount of information on the web) and preformed Wilcoxon test to determine that there is no statistical significance for the difference between the groups' averages.





Second,

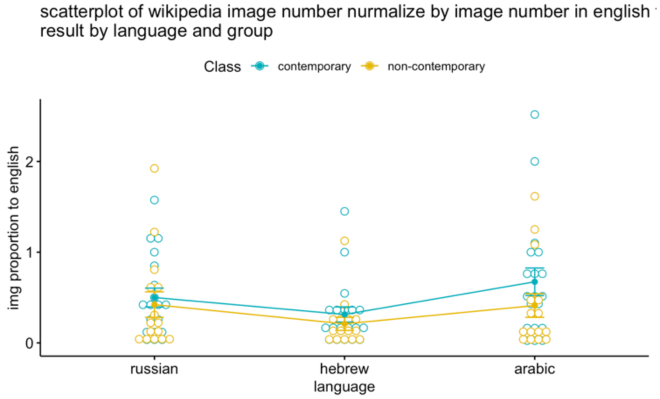
we defined the **manual ranking quality** rating of each term as the sum of each page rank (top 3 pages) in Google search results. We used one-way ANOVA test and Tukey multiple pairwise-comparisons in to determine statistical significance between the English’s mean rank to the other languages. The model met with its assumptions.

Third, we analyzed the data we mined from Wikipedia to compare the quantity of content available in different languages.

After testing the results, we decided to make a comparison between languages using the Kruskal-Wallis test which does not assume that the results are normally distributed. In those tests, we examined statistical difference between languages for each parameter. The results indicated a significant difference in each of the parameters measured:

- Number of images – in English the number of photos is higher than all other languages.
- Number of sections – Hebrew have less sub-topics then all other languages.
- Text length – English is the language with the longest text, then Russian and Arabic and Hebrew with the shortest.
- Number of references - English is the have the most references, then Russian and Arabic and Hebrew least.

After re-establishing the difference between the languages we divided the observations into groups of contemporary and non-contemporary terms, normalized them in relation to English results. For each petameter we performed a Wilcoxon rank sum test to show that there is no significant difference between them.



In conclusion, in each of our analysis the result shows a gap between the languages, in matters of quantity and quality of the content between languages. however, Contrary to our initial assumption we found no evidence of a significant gap for contemporary and non-contemporary in the ratio of search results to ideal content (English search results).

Limitations and future work:

The main limitation we had in the project was the inability to give a reliable rating to the information without relying on various translation tools. Consequently, most of the metrics we used are of quantity rather than of quality which limits our analytical capabilities. In addition, our manual ranking focused on a small sample of results for each term, future research on this topic would increase the number of samples and give more reliable results.

Because the content rating method is primarily subjective, further research in the field will require rankings of several language speakers to obtain a credible content rating.

In addition, it is possible to train an NLP model in different languages which will be able to replace manual ratings and collect a large sample in a flash of results.