Project proposal - searching for science

Team O - Omer Gazit 318254026 and Ido Bouhnik 206586794

1. Introduction

In our project "searching for Science" we are focusing on the question "Does the search language affect the quality of search results?"

Does a kid growing up in Bangladesh get the same quality of information online as a kid growing up in Israel? Online information is often thought of as universally accessible.

However, numerous studies have shown that language barriers and quality of information differ substantially across the globe.

In this project, we are interested in reexamining this question with respect to basic scientific terms and concepts (e.g. Atom), which should not differ as much from one language to another.

We have collected the Search Results pages for 26 carefully-selected basic scientific terms in four languages (English, Russian, Hebrew, and Arabic).

We would like to understand the quality of content offered in each of these languages, how the different components on the page differ, and ultimately develop better measures for the quality of scientific search results.

2. Data

The data is gathered from Google searches html files.

For each term we scrap the first Google search page with the code in the scraping html files section. (from the assumption that most users focus on the results that appear on the first page) then, we rank each website content in the result using a number of different factors like: clarity of the definition, quantity of content related to the subject, visualization related to the subject and more parameters that will help us determent the quality of the search.

we will add more parameters later in the project when we figure out more meta data factors that indicate about quality content.

for more details about the parameters we gathered so far - check the README file.

3. Preliminary results

From primary analysis on the data we collected from Google search of Mammals in four languages. We found huge differences between the amount of search results in different languages. by order: english, russian, arabic and hebrew.

This finding shows that the amount of information available in some languages is significantly greater than in other languages.

In addition, the question arose as to whether a large amount of information pushes quality information into the first search results as a result of Google's algorithm.

After manually ranking from the first pages in each language, we summarized the average rankings for each language and category and obtained that indeed the English search received the highest ranking.

The surprising result is that Hebrew mean ranking we received rated higher than languages with more search results. This may be the result of our ranking based on Google's translation and additional samples are required to determine this with more certainty.

See visualization below for graphs.

4. Data analysis plan

we decided of two stages for this project:

Stage 1 - come up with a valid ranking system for page quality

We will scrape more Google results from all four languages. Then we are going to look for more parameters that indicate about quality content using the following method:

a) manually rank result quality. (preference for collaboration with people who speak different languages)

b) Build regression model that will try and decide which parameters can predict well the manual ranking that indicates the quality of the content on the results sites.

c) chose a quality content score function that calculate all the relevant parameters and return a valid quality ranking for each result.

examples:

check the connection between number of appearances of the search world in the page text to our ranking of the content.

split the result into categorical such as: academic articles, marketing sites, digital library / online encyclopedia or un related sites. Then check if there is a connection between the category and the clearance of the scientific definition.

Stage 2 - explore the differences between languages

Once come up with a ranking system of measuring content quality, we will examine the claim that there is a difference between the quality of search results in different languages.

Then we will examine if the term category of the search is influencing the gap between the languages. for example - dose searches of terms in anatomy and space have the same language gap?

Teamwork :

At stage one, we will split the manual ranking between us.

Omer - write scraping scripts and preforming stage one parameters analysis.

Ido - stage two analysis and visualization.

Appendix

Data README

```
# SISE2601 Project data description
================
Team 0 - Omer Gazit and Ido Bouhnik
```

```
This Markdown file describes the data folder structure and organization ...

data_change.csv - contain the search result sites and manual ranking of the google search results.
fields:
  topic - the Google search term (English)
    language - the search language
    url - the result website url
    title - the search result url name
# ranking system
    clearance definition - how easy was to find the term definition on the site :
        1 - not found at all
        2 - found but not immediately
        3 - found very easily
    content - how mach information was found about the topic on the site :
        1 - not relevant information
        2 - only the definition and not farther
        3 - a lot of information and reference to more sources
    visualization - is there a visualization related to the term :
        1 - no visualization at all
        2 - 1 to 4 visualizations
        3 - more then 4 visualization
    professional looking site - How much is invested in the site :
        1 - amateur site
        2 - medium
        3 - professional website
    adds - is the site advertising something? (Yes or No filed)

html_classifire.csv - csv file that hold paths to local html files of google search.

Result_num_changed.csv - hold the number of result for every Google search
```

Source code

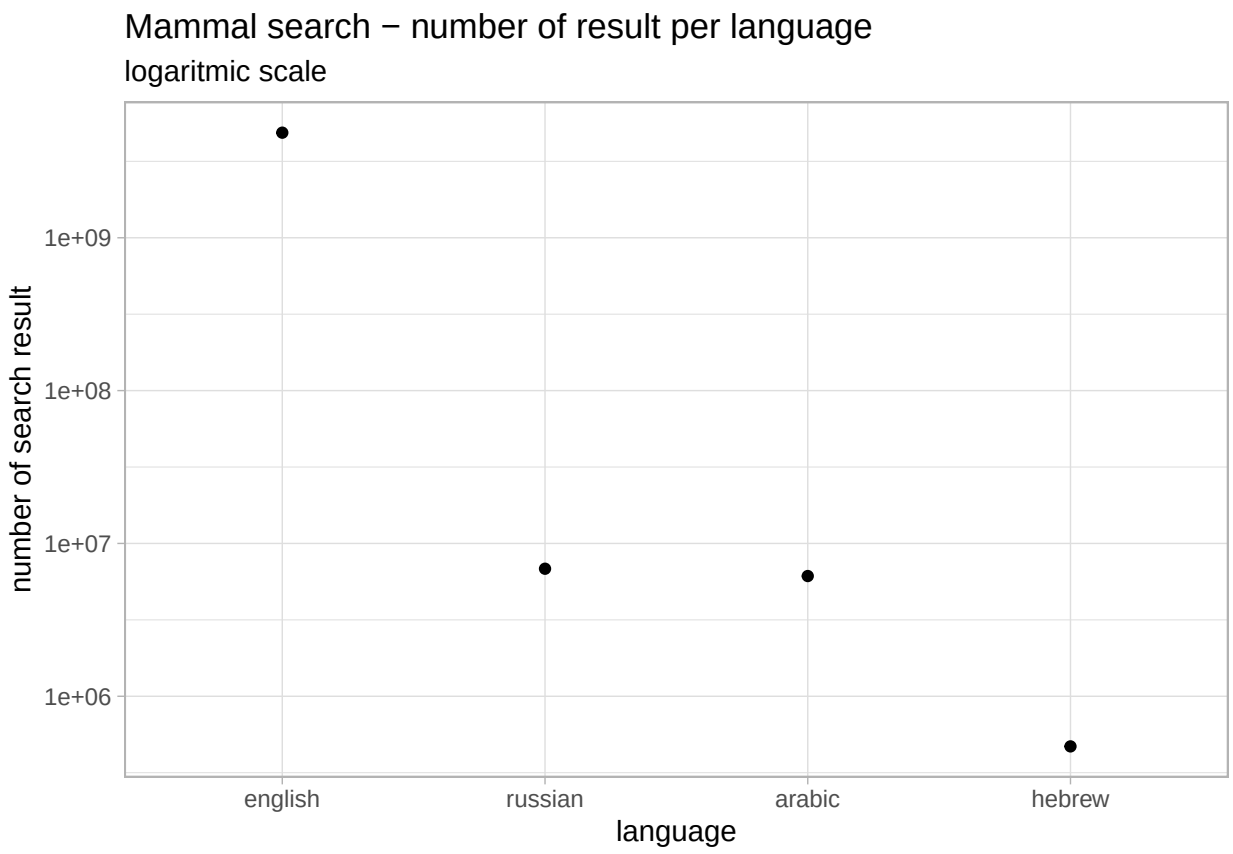
analysis

```
result_num_set <- read_csv("../data/result_num_chaneged.csv")
```

```
## New names:
## Rows: 4 Columns: 4
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (2): topic, language dbl (2): ...1, result_num
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
ggplot(data = result_num_set, mapping = aes(x =  factor(language, level = c('english', 'russian', 'arab
  geom_point() +
  scale_y_continuous(trans = 'log10') +
  theme_light()+
  labs(x = "language",
```

```
       y = "number of search result",
       title = "Mammal search - number of result per language",
       subtitle = "logaritmic scale")
```




```
result_ranking <- read_csv("../data/data_change.csv")
```

```
## New names:
## Rows: 35 Columns: 11
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (6): topic, language, url, title, summery, adds dbl (5): ...1, Clearance
## definition, Content, Visualization, Profesional loo...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
Clearance_mean <- result_ranking %>%
  group_by(language) %>%
  summarise_at(vars("Clearance definition"), list("Clearance definition" = mean))

Content_mean <- result_ranking %>%
  group_by(language) %>%
  summarise_at(vars("Content"), list("Content" = mean))

Visualization_mean <- result_ranking %>%
```

```
  group_by(language) %>%
  summarise_at(vars("Visualization"), list("Visualization" = mean))

Profesional_mean <- result_ranking %>%
  group_by(language) %>%
  summarise_at(vars("Profesional looking site"), list("Profesional looking site" = mean))

advertize <- result_ranking %>%
  group_by(language) %>%
  count(adds)

advertize <- advertize %>%
  group_by(language) %>%
  summarise(sites_with_adds = sum(n[adds =="Yes"])/sum(n))

combined <- Clearance_mean %>%
  inner_join(Content_mean) %>%
  inner_join(Visualization_mean) %>%
  inner_join(Profesional_mean) %>%
  inner_join(advertize)
```

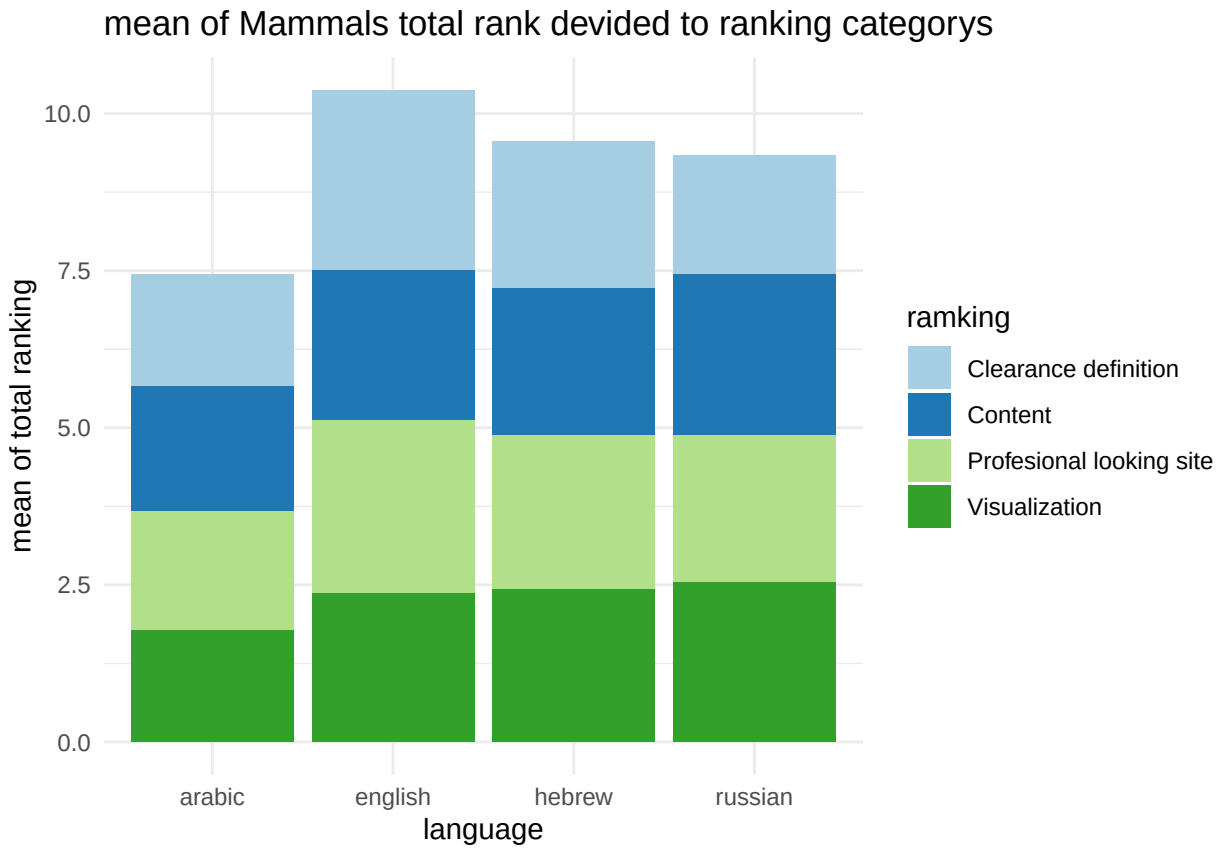
```
## Joining, by = "language"
## Joining, by = "language"
## Joining, by = "language"
## Joining, by = "language"
```

```
combined
```

```
## # A tibble: 4 x 6
##   language `Clearance definition` Content Visualization `Profesional looking s~`
##   <chr>                     <dbl>   <dbl>         <dbl>                    <dbl>
## 1 arabic                     1.78    2             1.78                     1.89
## 2 english                    2.88    2.38          2.38                     2.75
## 3 hebrew                     2.33    2.33          2.44                     2.44
## 4 russian                    1.89    2.56          2.56                     2.33
## # ... with 1 more variable: sites_with_adds <dbl>
```

```
longer_combined <- pivot_longer(combined,
             cols = c("Clearance definition", Content, Visualization,"Profesional looking site"),
             names_to = "ramking")

ggplot(data=longer_combined, aes(x=language, y=value, fill=ramking)) +
  geom_bar(stat="identity")+
  scale_fill_brewer(palette="Paired")+
  theme_minimal()+
  labs(title = "mean of Mammals total rank devided to ranking categorys",
       y = "mean of total ranking")
```



scraping html files

```
search_page_scrape <- function(html_path, language, topic){
  # scraping google search html file and return data frame with the values (topic, language, result url
  rawHTML <- read_html(html_path)
  if(language == "english" || language == "russian"){
    urls <- rawHTML %>%
      html_nodes(".tjvcx") %>%
      html_text()
    urls <- urls[seq(1, length(urls), 2)]
    urls <- gsub("\\ .*", "", urls)
  }
  else{
    urls <- rawHTML %>%
      html_nodes(".tjvcx > span:nth-child(1)") %>%
      html_text()
    urls <- urls[seq(1, length(urls), 2)]
  }

  titles <- rawHTML %>%
    html_nodes(".DKVOMd") %>%
    html_text()

  summery <- rawHTML %>%
    html_nodes(".lyLwlc") %>%
```

```
    html_text()

  fsummery <- rawHTML %>%
    html_nodes(".hgKElc") %>%
    html_text()
  fsummery

  summery <- c(fsummery, summery)
  typeof(summery)

  google_search <- tibble(
    topic = topic,
    language = language,
    url = urls,
    title = titles,
    summery = summery
  )
  return(google_search)
}

all_scraper <- function(){
  # reads a csv file contain (term, language, path to html) and write 2 csv files: first page results a
  exceldata = read.csv("../data/html_classefire.csv")
  df <- data.frame(exceldata)
  topics <- df$topic
  languages <- df$language
  paths <- df$path
  new_df <- search_page_scrape(paths[1], languages[1], topics[1])
  result_df <- scraping_number_of_result(paths[1], languages[1], topics[1])
  for(i in 2:length(topics)){
    new_df <- rbind(new_df, search_page_scrape(paths[i], languages[i], topics[i]))
    result_df <- rbind(result_df, scraping_number_of_result(paths[i], languages[i], topics[i]))
  }
  write.csv(new_df, "../data/data.csv", row.names=TRUE)
  write.csv(result_df, "../data/result_num.csv", row.names=TRUE)
}

scraping_number_of_result <- function(html_path, language, topic){
  # return number of results for Google search html file
  rawHTML <- read_html(html_path)
  result_num <- rawHTML %>%
    html_node("#result-stats") %>%
    html_text() %>%
    str_extract("[0-9]{1,3}( [0-9]{3})+|[0-9]{1,3}(,[0-9]{3})+") %>%
    str_remove_all(",") %>%
    str_remove_all(" ") %>%
    as.numeric()

  result_set <- tibble(
    topic = topic,
    language = language,
    result_num = result_num)
}
```

```
all_scraper()
```

```
## Warning in read.table(file = file, header = header, sep = sep, quote = quote, :
## incomplete final line found by readTableHeader on '../data/html_classefire.csv'
```