BDA Mini Project

Google Search Results Analysis

Rajat Belgundi PC08
Kartik Bhutada PC09
Amey Bhide PC14
Divyang Bagla PC33

Introduction:

1. Google search results are very efficient and provide relevant web pages based on the keyword searches to its users in a matter of seconds.
2. Here, we try to find insights about how google ranks various websites based on different characteristics such as number of searches, content type, content length, etc on 3 very popular domains - Wikipedia, Fandom and Quora and use tableau as a visualization tool to help with the purpose.
3. We used MongoDB as our database to store the data and pymongo to show how queries can be performed, based on the requirement of the user.

Technology Stack:

1. MongoDB
2. Mongo Shell
3. Jupyter Notebook
4. Tableau
5. Python

Dataset:

The dataset link is available on:-
https://zenodo.org/record/3541092#.YVRV_ZpBxPZ

Dataset contains the following elements:

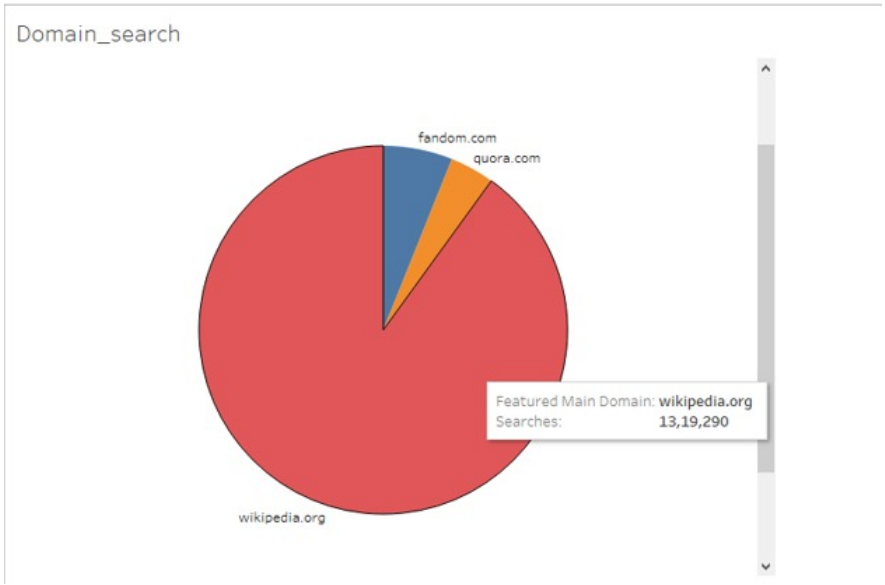
1. keyword,
2. number of monthly searches,
3. featured domain,

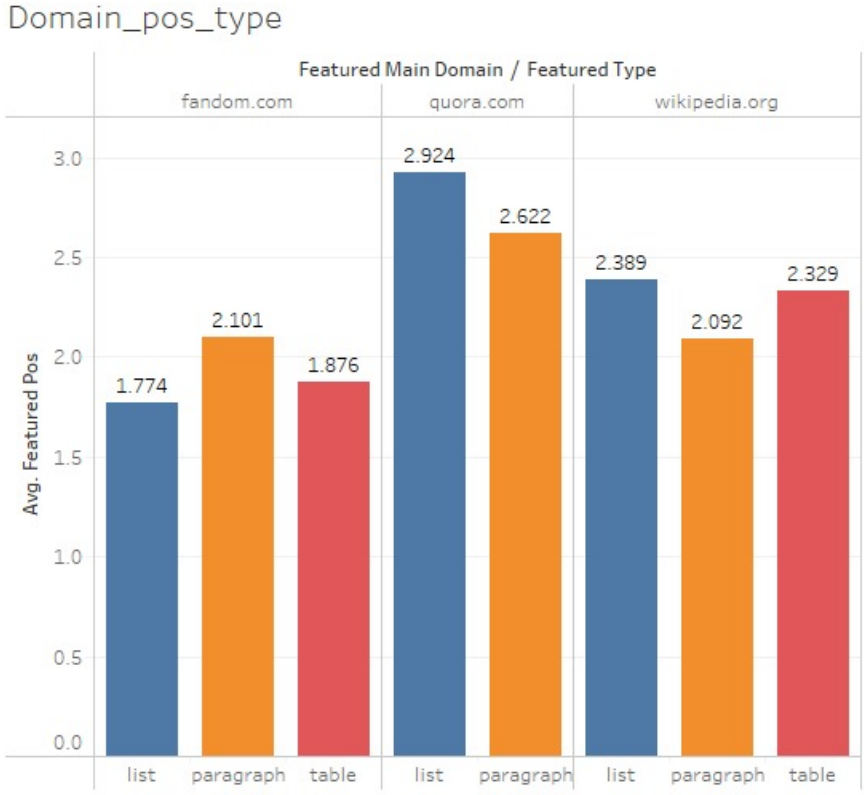
4. featured main domain,
5. featured position,
6. featured type,
7. featured url,
8. content,
9. content length.

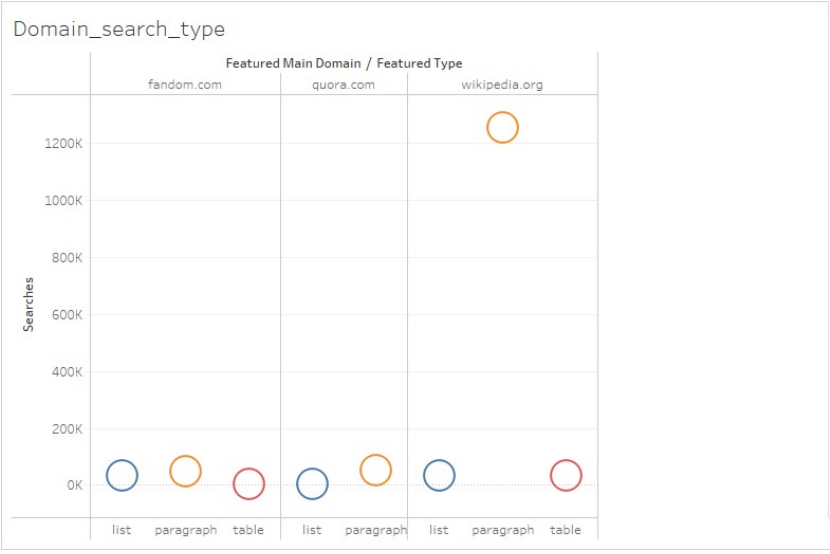
Queries executed:

1. Get Element By Keyword: find_one (create)
2. Get Elements with conditions
3. Sort Elements by Ascending and Descending order
4. CRUD operations
5. Group by featured_pos and then sort by count
6. Group by searches and get elements whose searches are greater than 50 and less than 200
7. Group by featured_type and get results whose count is greater than 1.

Tableau visualizations:




Domain_pos_type
Featured Main Domain / Featured Type
fandom.com
quora.com
wikipedia.org
Avg. Featured Pos
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1.774
2.101
1.876
2.924
2.622
2.389
2.092
2.329
list
paragraph
table
list
paragraph
list
paragraph
table


Domain_search_type
Featured Main Domain / Featured Type
fandom.com
quora.com
wikipedia.org
Searches
1200K
1000K
800K
600K
400K
200K
0K
list
paragraph
table
list
paragraph
list
paragraph
table

Conclusion:

MongoDB is a great tool to handle large datasets and perform complex queries on them. With pymongo you can run MongoDB queries in a python environment and retrieve data in a dictionary format for further processing in the environment itself.

Following things can be concluded after visualization:-

1. Wikipedia is the most searched domain followed by fandom and quora.
2. Most of the content is of type paragraphs in all 3 sites.
3. Sites having around 10 searches have most of their featured positions in 1st rank on google however interestingly, sites having more than 10000 searches do not have their featured position as 1 rather they are placed at 4/6th position.
4. Sites having content lengths of more than 200 are more likely to rank higher than other sites.