DIGITAL DOCUMENT SEARCH

TEAM: THE BULLS

MARIYA JOHAR

YASWANTH JAGILANKA

KOLLI SAI NITHIN REDDY

University at Buffalo The State University of New York

Motivation

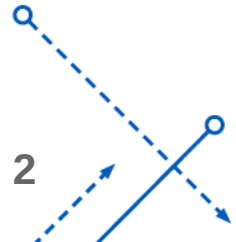
- Retrieve relevant research docs with high accuracy
- Access to huge pool of data in short span of time
- Focused approach to enhance user experience

Problem Statement

- Age of Big Data raised challenges in terms of data compression, computation, efficient algorithms & labor cost
- Losses in terms of time, money and resources are a hotspot
- Uphill task of quick & efficient retrieval of relevant documents
- Develop a Digital Document Search for a smooth experience

Research Questions

- Understanding hidden topics via topic modelling techniques such as BERT and LDA.
- User query evaluation to figure impact on model performance.
- Model credibility in events of highly diverse research disciplines.

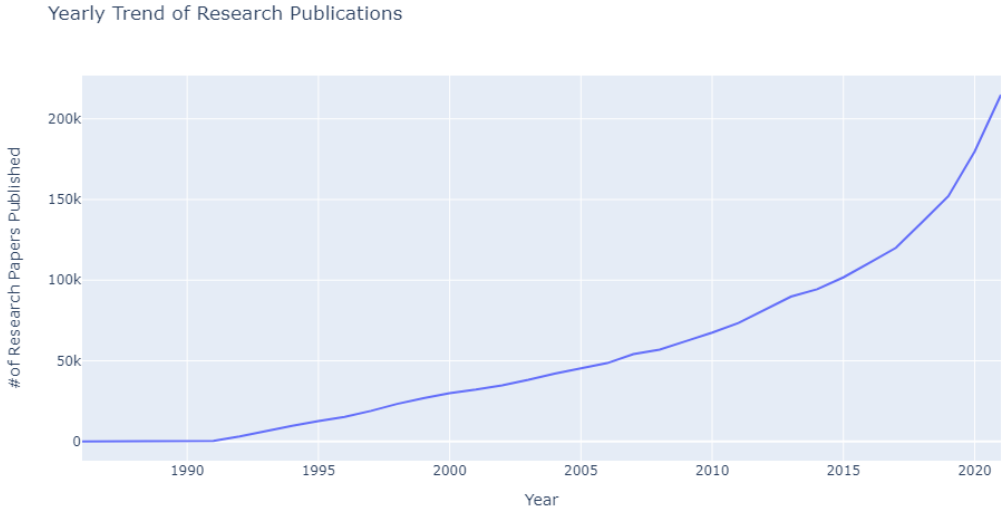


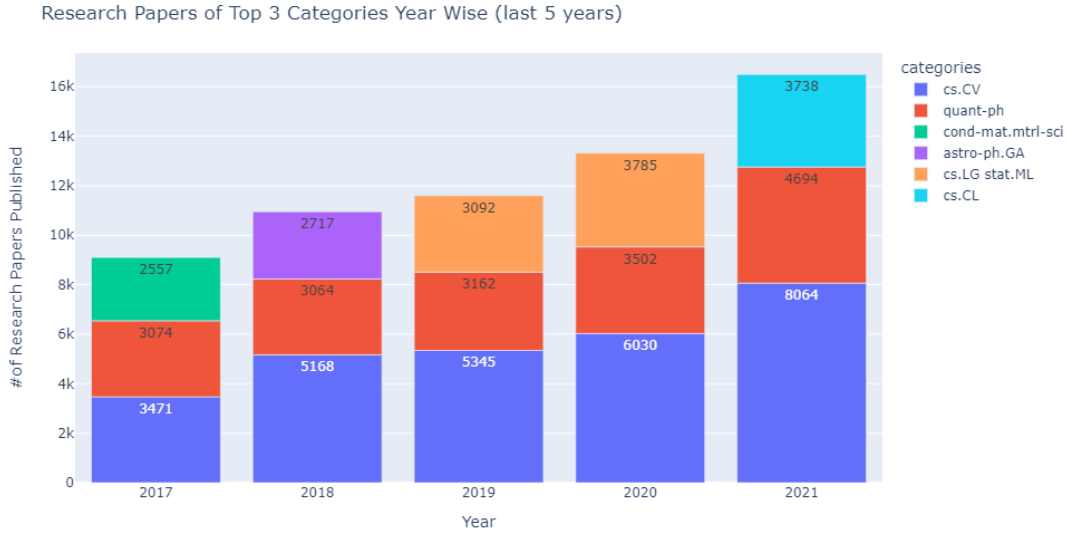
Understanding the Data

- The input file is a subset of ArXiv's repository.
- It captures information like id, title, abstract, authors, category etc., for each research article.
- These research articles are from multiple disciplines like computer science, Astro physics, quantum physics, etc.

Input file structure

```
[('abstract', 'string'),
 ('authors', 'string'),
 ('authors_parsed', 'array<array<string>>'),
 ('categories', 'string'),
 ('comments', 'string'),
 ('doi', 'string'),
 ('id', 'string'),
 ('journal-ref', 'string'),
 ('license', 'string'),
 ('report-no', 'string'),
 ('submitter', 'string'),
 ('title', 'string'),
 ('update_date', 'string'),
 ('versions', 'array<struct<created:string,version:string>>')]
```

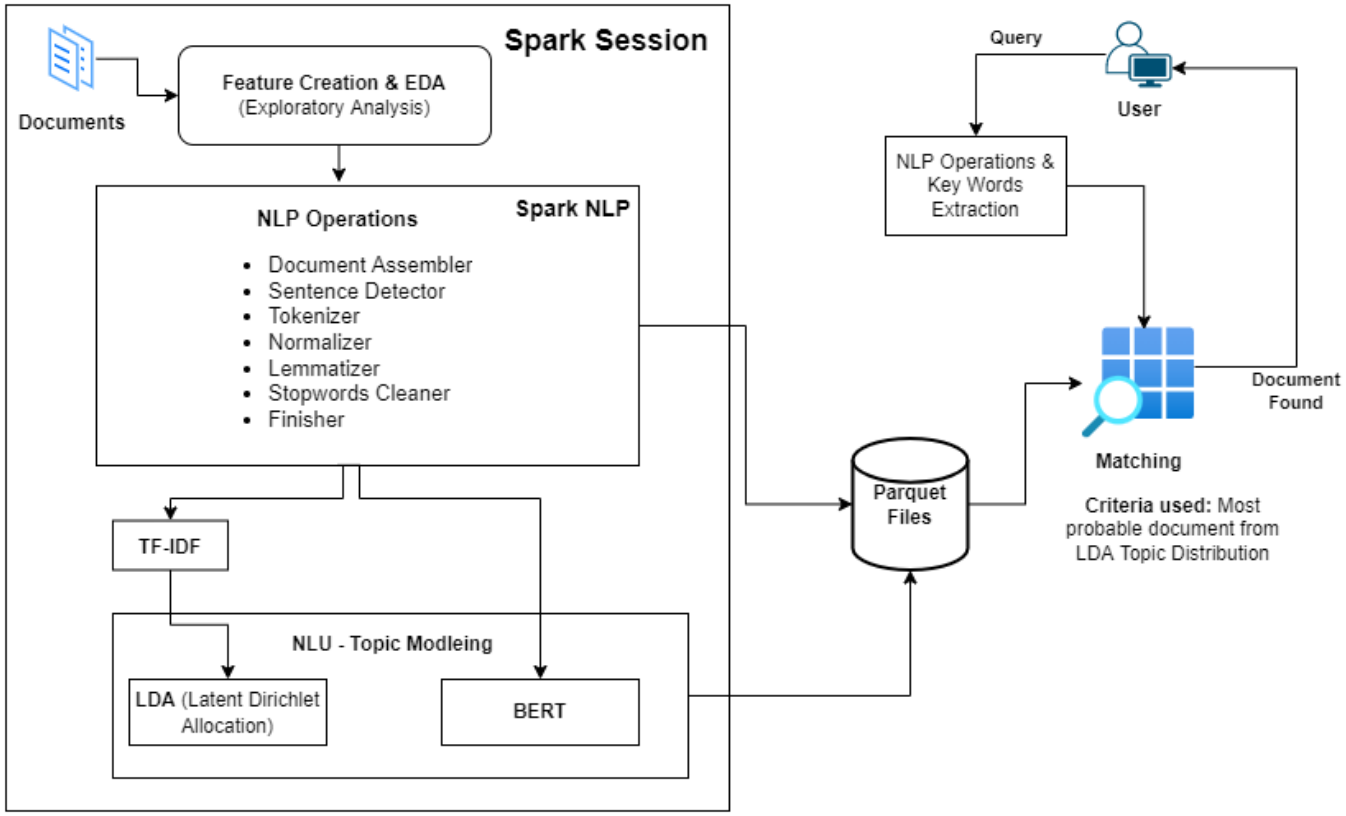




- Total number of **research papers** published is **2.03 million**.
- Number of **distinct authors** who have published research articles is **1.95 million**.

Approach

Technical Architecture



A Brief Overview of Techniques Used

Latent Dirichlet Allocation (LDA)

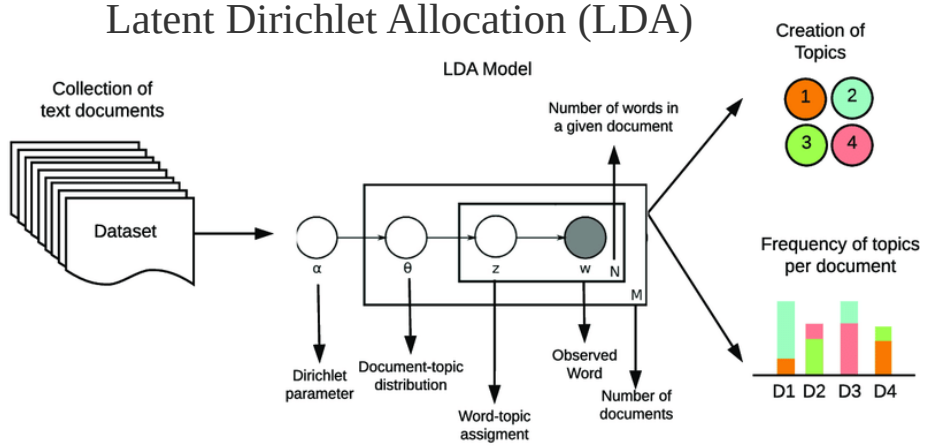
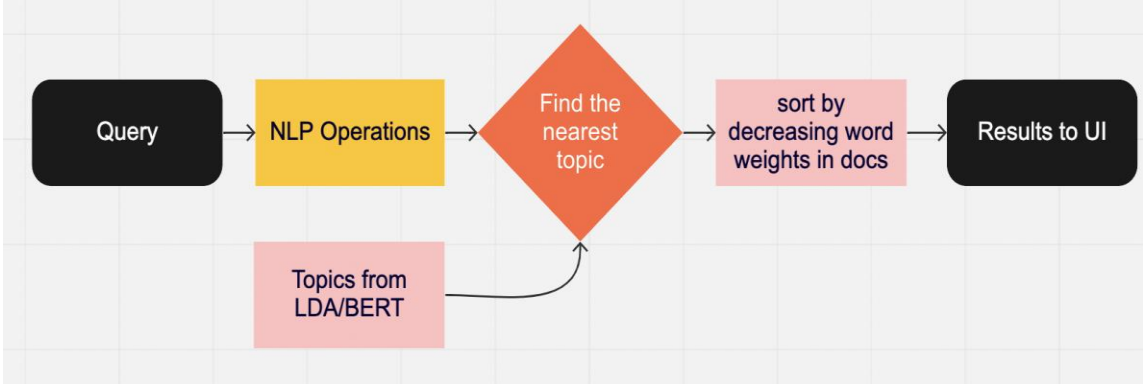
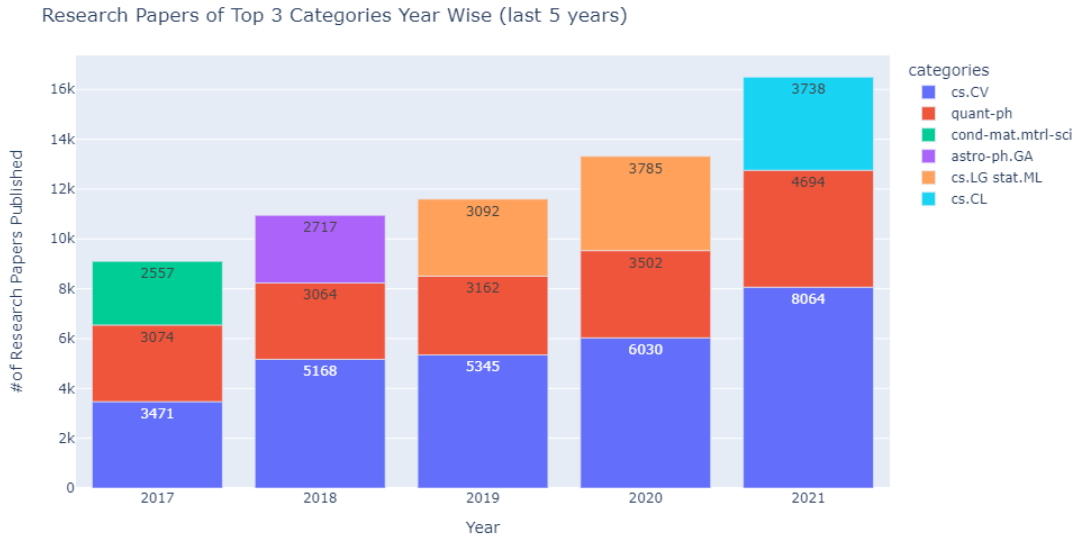


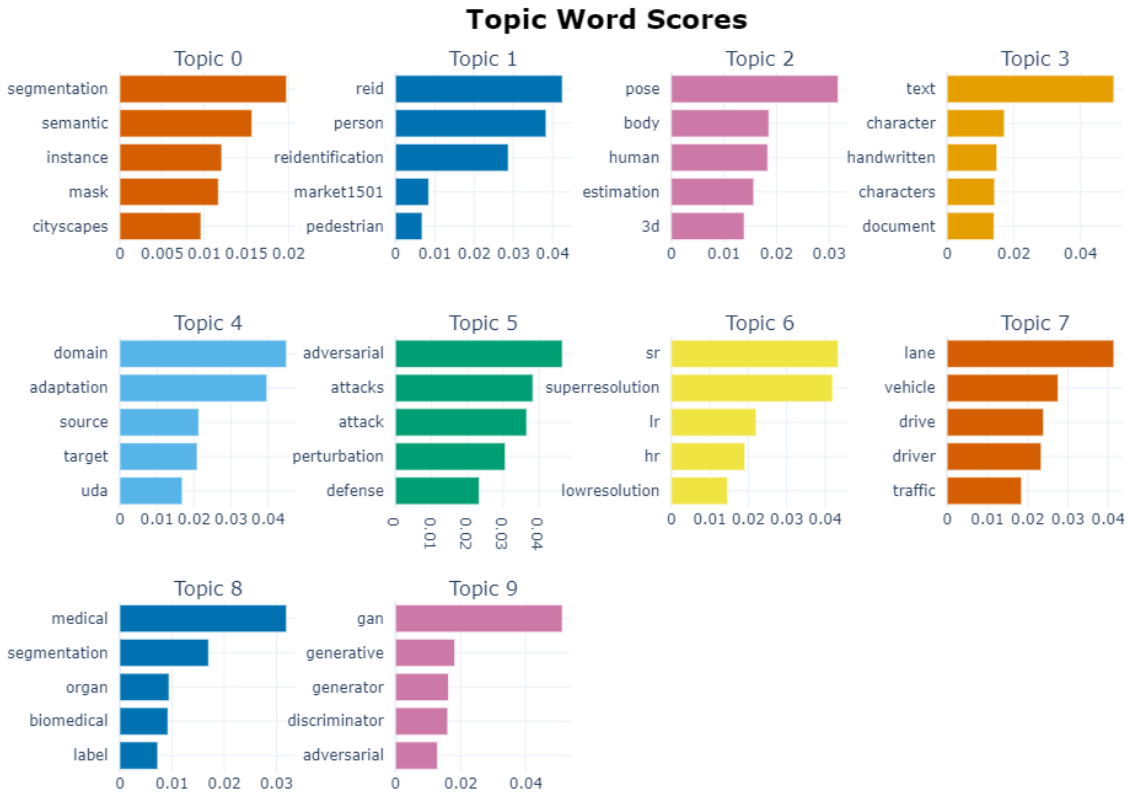
Image Source

Matching Criteria



Results and Application Demo






```
127.0.0.1 - - [02/May/2022 17:38:26] "GET / HTTP/1.1" 200 -
127.0.0.1 - - [02/May/2022 17:38:27] "GET /favicon.ico HTTP/1.1" 404 -
CVIP
127.0.0.1 - - [02/May/2022 17:38:29] "POST /query_input HTTP/1.1" 200 -
Seconds since epoch = 9.473736763000488
127.0.0.1 - - [02/May/2022 17:38:35] "POST /query_response HTTP/1.1" 200 -
```

Findings

Category : Computer Vision **Query1 :** scene detection using convolutional neural networks

Here are the top results:

1. Non-anchor-based vehicle detection for traffic surveillance using bounding ellipses
2. Road Surface Translation Under Snow-covered and Semantic Segmentation for Snow Hazard Index
3. Simultaneous Multi-View Camera Pose Estimation and Object Tracking with Square Planar Markers
4. Attention Based Semantic Segmentation on UAV Dataset for Natural Disaster Damage Assessment
5. Real-Time Trash Detection for Modern Societies using CCTV to Identifying \
   Trash by utilizing Deep Convolutional Neural Network

Category : Computer Vision **Query2 :** image transformers

Here are the top results:

1. LocalViT: Bringing Locality to Vision Transformers
2. Vision Transformer for Small-Size Datasets
3. Demystifying Local Vision Transformer: Sparse Connectivity, Weight Sharing, and Dynamic Weight
4. Reveal of Vision Transformers Robustness against Adversarial Attacks
5. Improved Robustness of Vision Transformer via PreLayerNorm in Patch Embedding

Outcomes and Impact

- Works best for specific topic-based query
- Faster and lighter
- Heavy on one-time processing for a specific category
- Implementations of other deep techniques can make it more robust and better for wide range queries
- Our application is able to retrieve the articles within computer vision in a fraction of seconds with high relevancy.
- It can be used across multiple business domains, not just for document retrieval in research articles. For example, this system can be used within an organization to help employees retrieve information about HR policies, maternity leaves, financial information, tax information, and so on.
- Not robust for broader query

Future Research Opportunities

- Our matching criteria were not that accurate with diverse datasets and it was taking longer time to process large number of diverse research articles. So, as a further research opportunity, we can make use of TF-TDF, topic distribution from BERT and LDA, and execute this on Hadoop clusters along with Spark.
- Transformation into fast search engines and recommendation systems.
- Security aspect of the project to fetch data by assigning privileges to the data
- Integration of cloud technologies such as Databricks and Snowflake.