DIGITAL DOCUMENT SEARCH

Team: *The Bulls*

Members:

Raghav Kumar
Sri Hari Gopinath Thota

Group : ***The Bulls***

Problem Statement and Research Questions

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 :

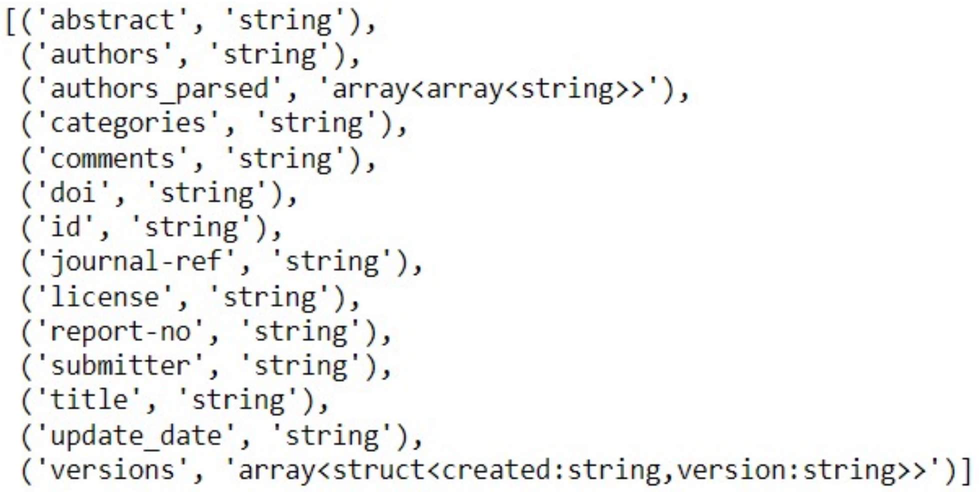
- Understanding hidden topics via topic modelling techniques such as BERT and LDA.
- User query evaluation to figure impact on model performance.
- Model credibility in events of highly diverse research disciplines.

*Group : **The Bulls***

Understanding the Data

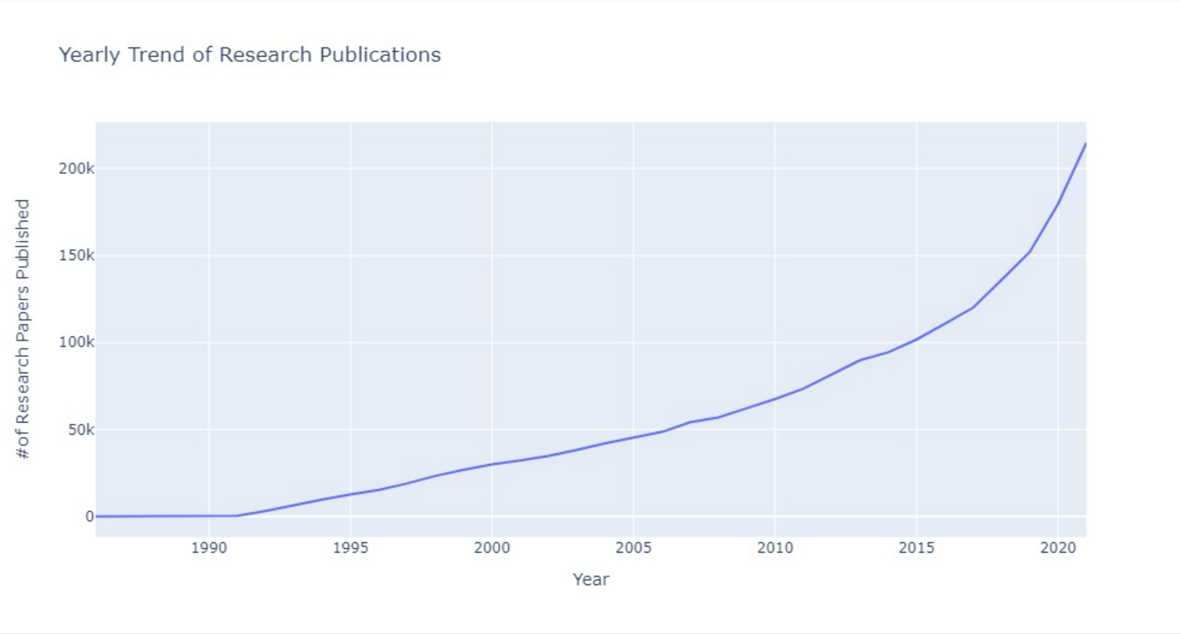
- The input file is a subset of ArXiv's repository.
- It captures information like id, title, abstract, authors, category etc., for each research article.
- These research articles are from multiple disciplines like computer science, quantum physics etc.

Input file structure

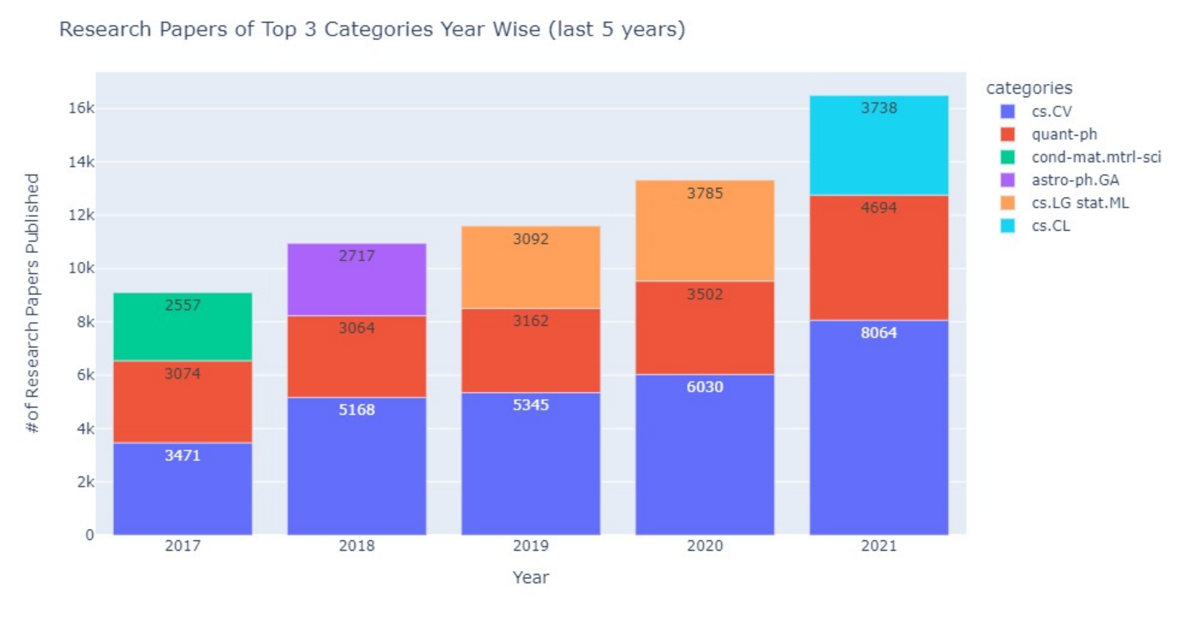


```
[('abstract', 'string'),
 ('authors', 'string'),
 ('authors_parsed', 'array<array<string>>'),
 ('categories', 'string'),
 ('comments', 'string'),
 ('doi', 'string'),
 ('id', 'string'),
 ('journal-ref', 'string'),
 ('license', 'string'),
 ('report-no', 'string'),
 ('submitter', 'string'),
 ('title', '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**.



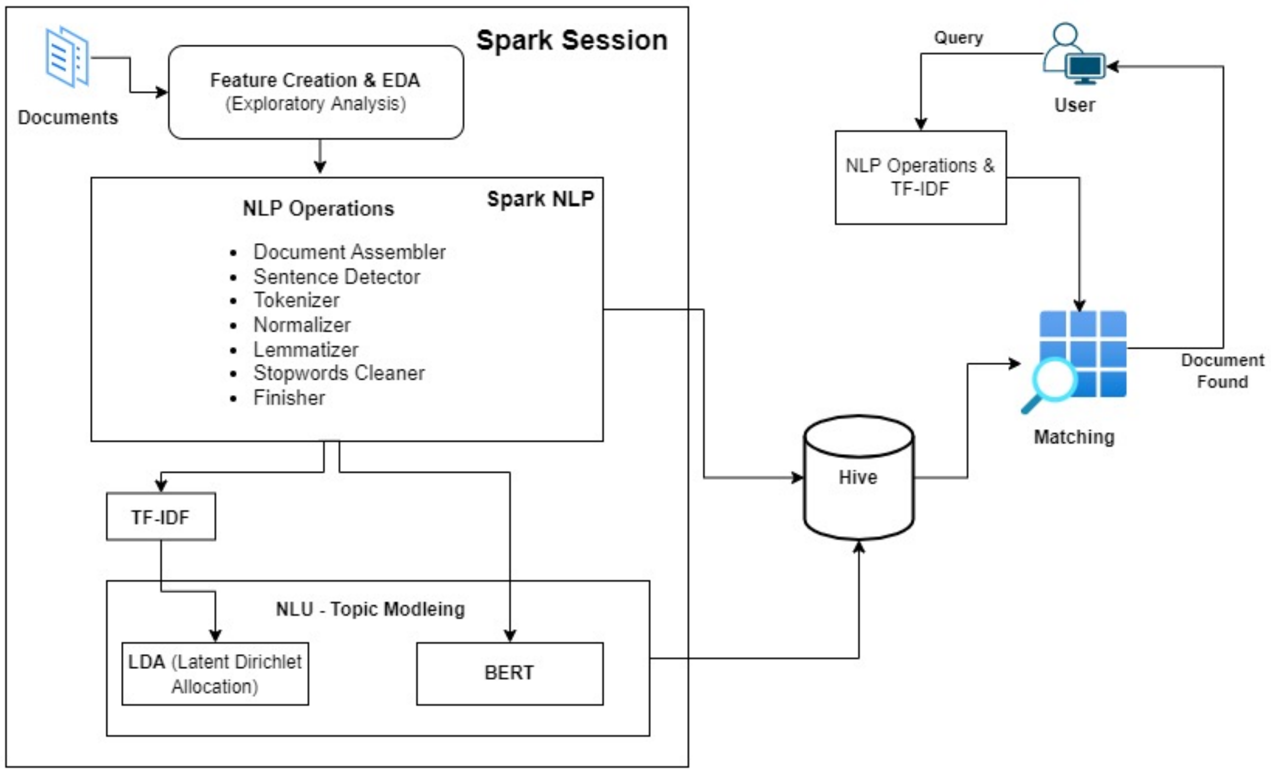
Yearly Trend of Research Publications



Research Papers of Top 3 Categories Year Wise (last 5 years)

*Group : **The Bulls***

Proposed Approach



Creation of Spark session, data cleansing and EDA

- As a first step, we have created spark session and loaded the input file. Performed EDA (Exploratory Data Analysis) to understand the data better.

NLP (Natural Language Processing) Operations

- Created a pipeline for NLP operations to perform on the abstract column of input file.

NLU (Natural Language Understanding)

- Used LDA (Latent Dirichlet Allocation) and BERT topic modelling methods to extract the hidden topics from abstracts of research articles.

Creation of Hive Database

- Database to be created on Hive to store the TF-IDF outputs and results of topic modelling obtained from LDA and BERT methods.

Query matching

- User’s query will be matched against results from topic modelling after performing NLP operations. Matching criteria is cosine similarity followed by retrieval of relevant documents.

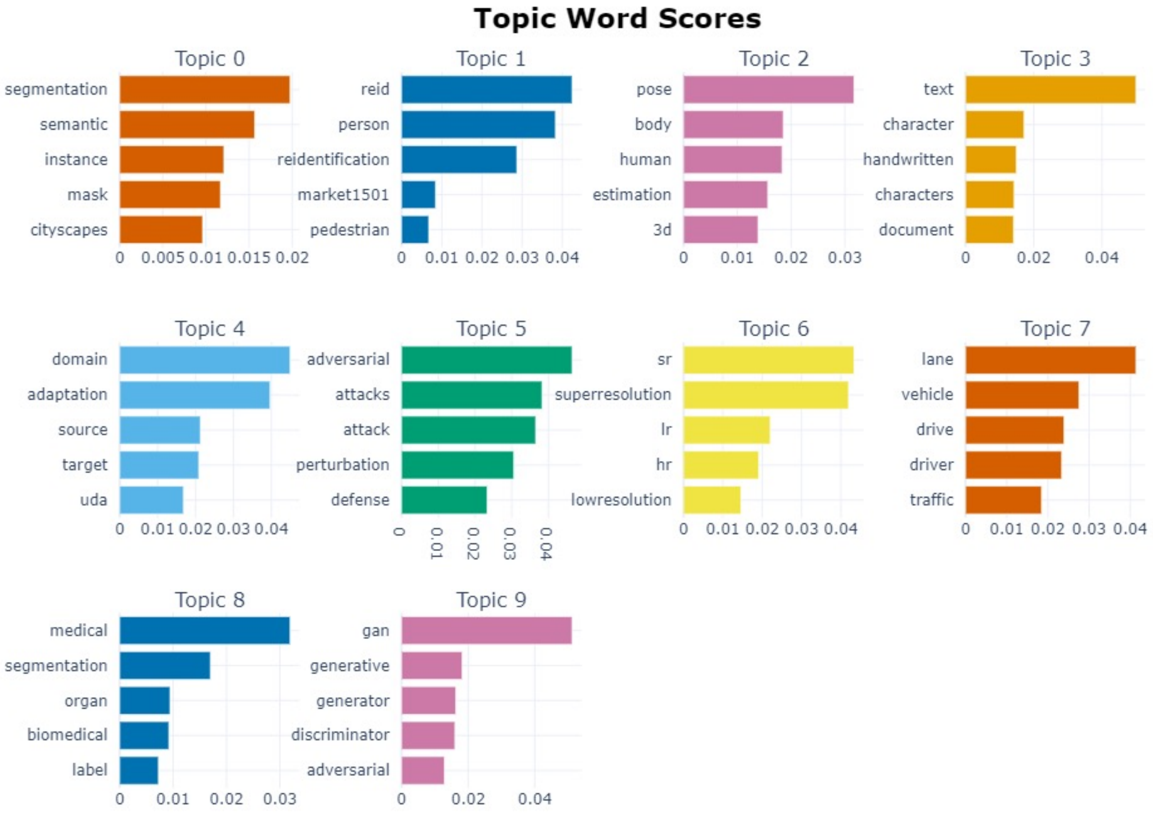
*Group : **The Bulls***

Intermediate Results

BERT Topic Modeling

- In this model, first the documents are embedded using “paraphrase-MiniLM-L12-v2” model.
- Performs dimensionality reduction using UMAP.
- Finally, the clustering of UMAP is done using HDBSCAN.

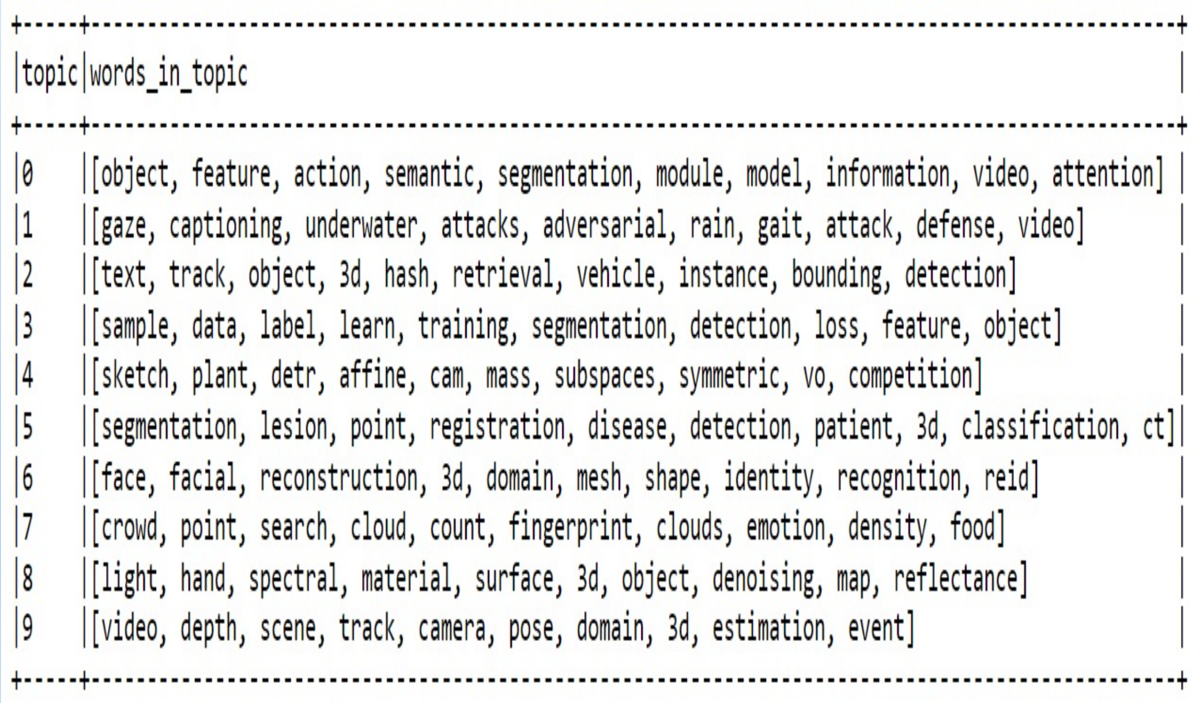
Below chart shows top 5 key words within each topics extracted from **abstracts of computer vision** research articles using **BERT**



LDA Topic Modeling

- Creates a Dirichlet distribution-based topics per document and words per topic model.
- This model takes TF-IDF weight vectors of each document as input to find the hidden topics from the abstracts of research articles.

Below chart shows top 10 key words within each topics extracted from **abstracts of computer vision** research articles using **LDA**



```
+-----+--------------------------------------------------------------------------------------------------+
|topic|words_in_topic                                                                                    |
+-----+--------------------------------------------------------------------------------------------------+
|0    |[object, feature, action, semantic, segmentation, module, model, information, video, attention]  |
|1    |[gaze, captioning, underwater, attacks, adversarial, rain, gait, attack, defense, video]          |
|2    |[text, track, object, 3d, hash, retrieval, vehicle, instance, bounding, detection]                |
|3    |[sample, data, label, learn, training, segmentation, detection, loss, feature, object]            |
|4    |[sketch, plant, detr, affine, cam, mass, subspaces, symmetric, vo, competition]                   |
|5    |[segmentation, lesion, point, registration, disease, detection, patient, 3d, classification, ct]|
|6    |[face, facial, reconstruction, 3d, domain, mesh, shape, identity, recognition, reid]              |
|7    |[crowd, point, search, cloud, count, fingerprint, clouds, emotion, density, food]                 |
|8    |[light, hand, spectral, material, surface, 3d, object, denoising, map, reflectance]               |
|9    |[video, depth, scene, track, camera, pose, domain, 3d, estimation, event]                         |
+-----+--------------------------------------------------------------------------------------------------+
```