The Bulls

Digital Document Search

Yash Rathi
yrathi@buffalo.edu

Raghav Kumar
raghavku@buffalo.edu

Anoop Mathew Peringalloor
anoopmat@buffalo.edu

Yaswanth Jagilanka
yjagilan@buffalo.edu

Mariya Johar
mariyajo@buffalo.edu

Kolli Sai Nithin Reddy
sainithi@buffalo.edu

Sri Hari Gopinath Thota
sthota2@buffalo.edu

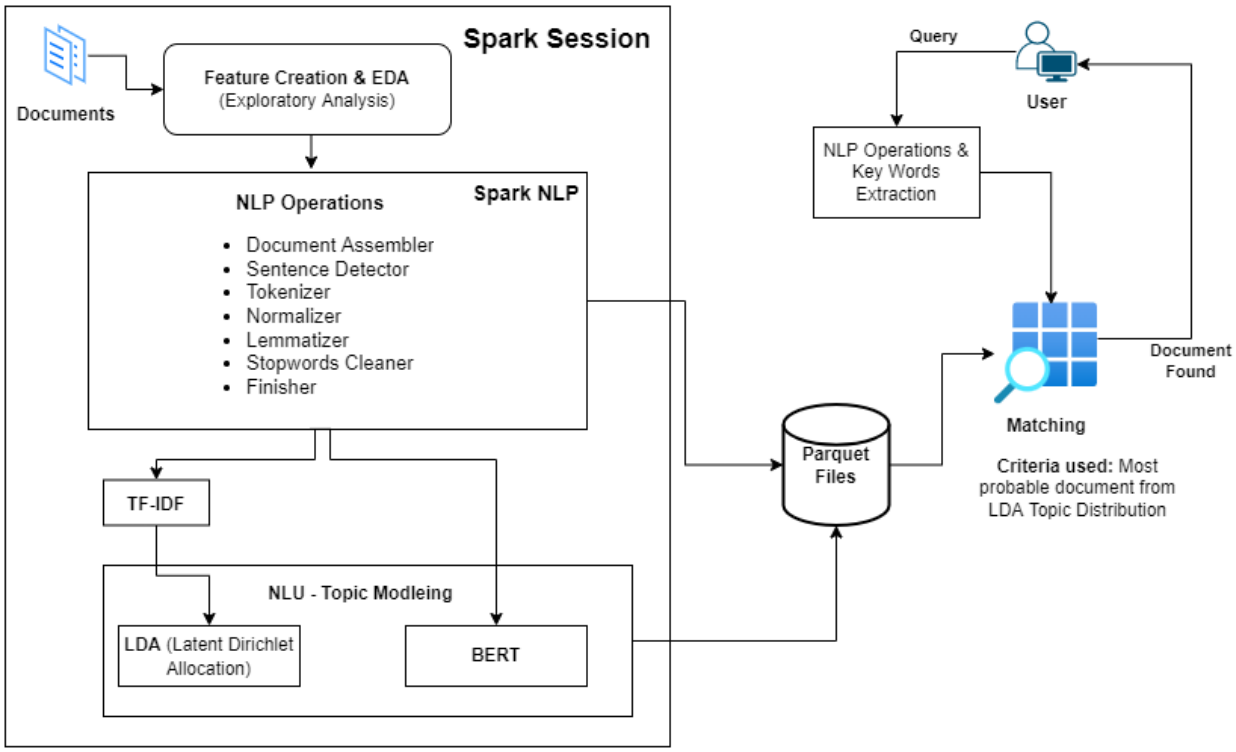
Qa

For our problem statement we defined the need to retrieve relevant research documents with good relevancy considering the huge set of data with respect to their respective domain and user-friendly interface to access significant data. The challenges we aimed to overcome where to handle increasing data set, avoid performance lag while dealing with these huge data pool and increase the accuracy for search results. Research questions focused on include using TF-IDF to convert textual content into meaningful numerical representations and using topic modeling for TF-IDF results.

Towards the end of our project, we could restate our problem statement as to create a simplified user-friendly interface to retrieve research papers with respect to a specific domain with relevant information for a huge pool of data. The challenges we addressed with the project are dealing with large data using Hive Database, improving performance by using NLP operations including TF-IDF to vectorize the most frequently used words in a document and finally to increase accuracy by using cosine similarity to display the most relevant data. These similarities are matched with the results from topic modeling operation Latent Dirichlet Allocation (LDA) and BERT to find the semantics of frequently occurring words from an abstract.

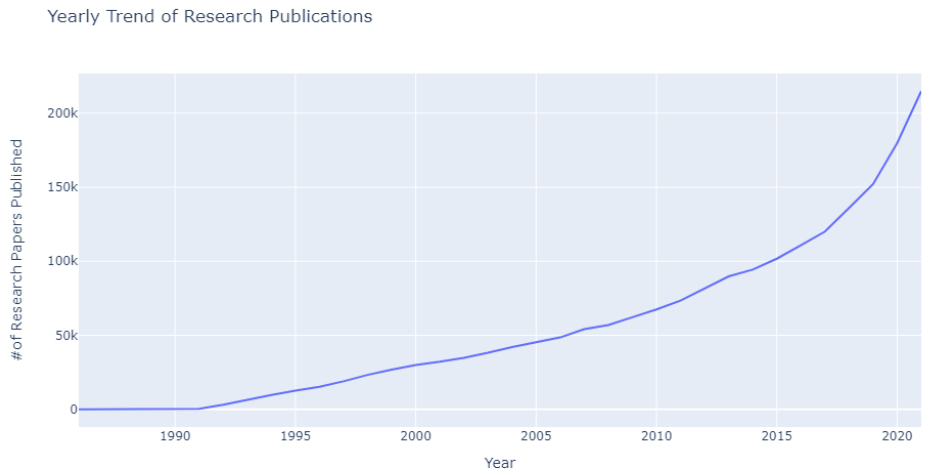
Qb

We are working with Cornell University's ArXiv dataset which includes over 1.7 million articles in JavaScript Object Notation. For each article, the following fields are considered: id, submitter, authors, title, categories, abstract, versions, and other information about the publication, such as the Digital Object Identifier (DOI). We tried to implement our model on all the articles but due to the vast amount of data and lack of computation power, we were not able to do so. Instead, we filtered the data to cover just the Computer Vision based research articles for the last 5 years, that is from 2017 to 2021. Our choice to select just the Computer Vision articles was due to the fact that articles from this field have dominated all other areas of research for the last 5 years. On filtering, we get a total of 28000+ articles. We read the data from the JSON file and filtered the features as per the requirements, then passed this data forward for NLP operations. As mentioned in the previous assignment, we applied NLP operations like document assembler, sentence tokenization (splitting the text into sentences), removing punctuations, stop words, converting the text to lower case, word tokenization (splitting the sentence into words), Stemming and Lemmatization (converting the words into base stem words while considering the context of the word). This is done to prepare our data for Topic Modeling like Latent Dirichlet Allocation (or LDA for short) - to find the topics hidden in the latent layer, and TF-IDF weighted indexing - to get the most important keywords from the data. We used Count Vectorization to find the term frequency for a document. All these processes are done in Spark session using Spark NLP. We tried to implement these steps without Spark, it took around 40-45 minutes just to preprocess the data and perform TF-IDF on over 28000 documents. Using the Spark NLP this time was reduced to a mere 10-12 minutes for the same dataset.

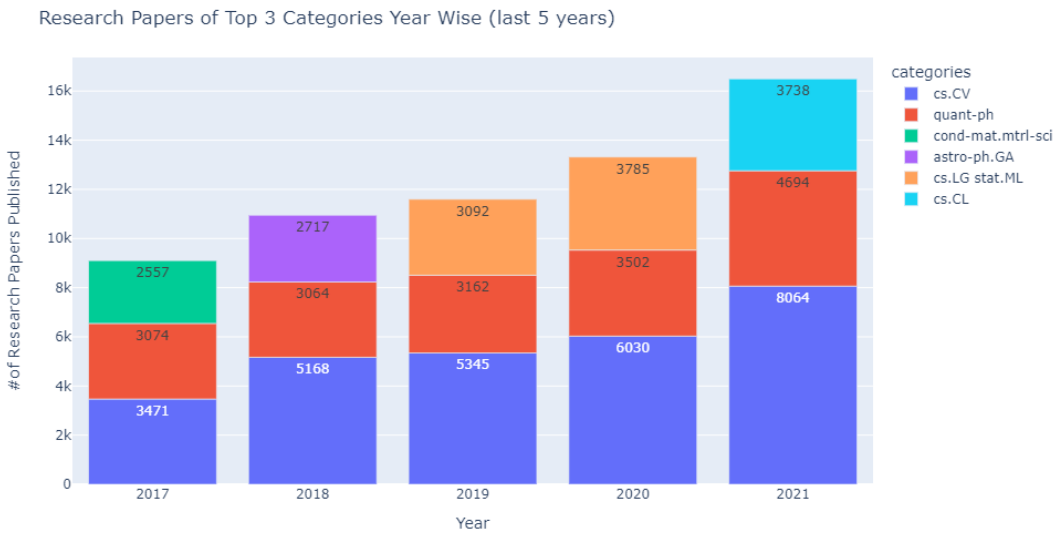


Application Architecture

Performed EDA (**Exploratory Data Analysis**) to understand the data better.



- We can observe from above chart that number of research papers published from 1990 to present almost follows exponential trend.



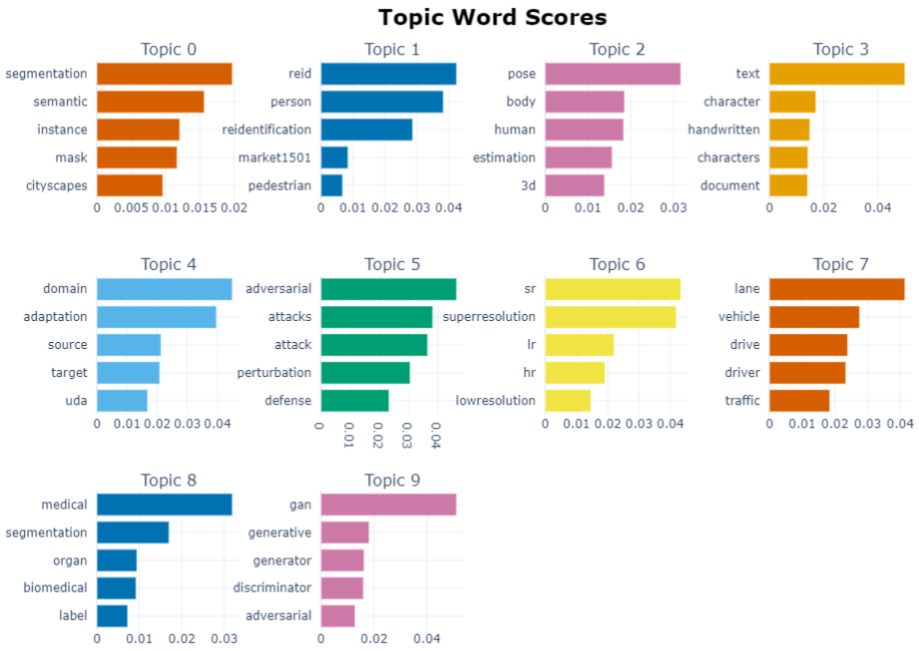
- The above graph shows the top 5 categories by volume of research papers published over last 5 years. We can observe that **research papers in the domains of computer vision (cs.CV) and quantum physics dominate the most**.
- Other key metrics like **total number of research papers published, and total number of distinct authors are 2032539 and 1946914 respectively**.

Top 10 Topics from LDA modelling on abstract column in the dataset

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]

- We also performed BERT topic modelling on the abstract column after performing all the NLP operations as mentioned in above section. In this model, first the documents are embedding is done using “paraphrase-MiniLM-L12-v2” model and performs dimensionality reduction using UMAP. Finally, the clustering of UMAP is done using HDBSCAN.

Top 10 topics from BERT modeling



- The results from both the topic modelling are stored in parquet files which will be used for matching user's query by cosine similarity.

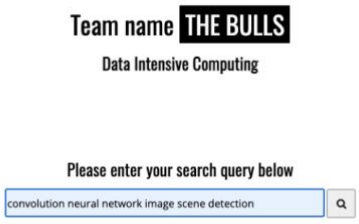
User interface from where user can enter his/her query will be developed in the next phase. We will also perform cosine similarity on user's query after extracting key words with the results from TF-IDF weighting and topic modelling results and compare the accuracy.

User query to Topic matching: The query entered by the user is processed using the nltk library for stopwords, lemmatization and other basic NLP operations, this is then converted into a list of words. Corresponding to these words and the weights associated with these words in the topic distribution, we find the topic with which this given word is closely related to. Based on the topic selected, we sort the generic database of documents according to weight of each document for the respective topic and output the top 15 articles in the selection.

UI and Result discussion:

In the application we have 3 stages as shown below.

Stage 1: Select the category of articles



Stage 2: Enter you query

And the stage3 shows the respective results as shown below.

Team name **THE BULLS**

Data Intensive Computing

Relevant results for the given query from the Arxiv Database

title

	id	title	abstract	authors	Year
3052	2104.05707	LocalViT: Bringing Locality to Vision Transformers	We study how to introduce locality mechanisms into vision transformers. The transformer network originates from machine translation and is particularly good at modelling long-range dependencies within a long sequence. Although the global interaction between the token embeddings could be well modelled by the self-attention mechanism of transformers, what is lacking a locality mechanism for information exchange within a local region. Yet, locality is essential for images since it pertains to structures like lines, edges, shapes, and even objects. We add locality to vision transformers by introducing depth-wise convolution into the feed-forward network. This seemingly simple solution is inspired by the comparison between feed-forward networks and inverted residual blocks. The importance of locality mechanisms is validated in two ways: 1) A wide range of design choices (activation function, layer placement, expansion ratio) are available for incorporating locality mechanisms and all proper choices can lead to a performance gain over the baseline, and 2) The same locality mechanism is successfully applied to 4 vision transformers, which shows the generalization of the locality concept. In particular, for ImageNet2012 classification, the locality-enhanced transformers outperform the baselines DeiT-T and PVT-T by 2.6\% and 3.1\% with a negligible increase in the number of parameters and computational effort. Code is available at \url{https://github.com/ofsoundof/LocalViT}.	Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, Luc Van Gool	2021
8008	2112.13492	Vision Transformer for Small-Size Datasets	Recently, the Vision Transformer (ViT), which applied the transformer structure to the image classification task, has outperformed convolutional neural networks. However, the high performance of the ViT results from pre-training using a large-size dataset such as JFT-300M, and its dependence on a large dataset is interpreted as due to low locality inductive bias. This paper proposes Shifted Patch Tokenization (SPT) and Locality Self-Attention (LSA), which effectively solve the lack of locality inductive bias and enable it to learn from scratch even on small-size datasets. Moreover, SPT and LSA are generic and effective add-on modules that are easily applicable to various ViTs. Experimental results show that when both SPT and LSA were applied to the ViTs, the performance improved by an average of 2.96% in Tiny-ImageNet, which is a representative small-size dataset. Especially, Swin Transformer achieved an overwhelming performance improvement of 4.08% thanks to the proposed SPT and LSA.	Seung Hoon Lee, Seunghyun Lee, Byung Cheol Song	2021
4216	2106.04263	Demystifying Local Vision Transformer: Sparse Connectivity, Weight Sharing, and Dynamic Weight	Vision Transformer (ViT) attains state-of-the-art performance in visual recognition, and the variant, Local Vision Transformer, makes further improvements. The major component in Local Vision Transformer, local attention, performs the attention separately over small local windows. We rephrase local attention as a channel-wise locally-connected layer and analyze it from two network regularization manners, sparse connectivity and weight sharing, as well as weight computation. Sparse connectivity: there is no connection across channels, and each position is connected to the positions within a small local window. Weight sharing: the connection weights for one position are shared across channels or within each group of channels. Dynamic weight: the connection weights are dynamically predicted according to each image instance. We point out that local attention resembles depth-wise convolution and its dynamic version in sparse connectivity. The main difference lies in weight sharing - depth-wise convolution shares connection weights (kernel weights) across spatial positions. We empirically observe that the models based on depth-wise convolution and the dynamic variant with lower computation complexity perform on-par with or sometimes slightly better than Swin Transformer, an instance of Local Vision Transformer, for ImageNet classification, COCO object detection and ADE semantic segmentation. These observations suggest that Local Vision Transformer takes advantage of two regularization forms and dynamic weight to increase the network capacity.	Qi Han, Zejia Fan, Qi Dai, Lei Sun, Ming-Ming Cheng, Jiaying Liu,\n Jingdong Wang	2021
6292	2109.12801	Effect Of Personalized Calibration On Gaze Estimation Using Deep-Learning	With the increase in computation power and the development of new state-of-the-art deep learning algorithms, appearance-based gaze estimation is becoming more and more popular. It is believed to work well with curated laboratory data sets, however it faces several challenges when deployed in real world scenario. One such challenge is to estimate the gaze of a person about which the Deep Learning model trained for gaze estimation has no knowledge about. To analyse the performance in such scenarios we have tried to simulate a calibration mechanism. In this work we use the MPIIGaze data set. We trained a multi modal convolutional neural network and analysed its performance with and without calibration and this evaluation provides clear insights on how calibration improved the performance of the Deep Learning model in estimating gaze in the wild.	Nairit Bandyopadhyay, S\'ebastien Riou, Didier Schwab	2021
4192	2106.03734	Reveal of Vision Transformers Robustness against Adversarial Attacks	The major part of the vanilla vision transformer (ViT) is the attention block that brings the power of mimicking the global context of the input image. For better performance, ViT needs large-scale training data. To overcome this data hunger limitation, many ViT-based networks, or hybrid-ViT, have been proposed to include local context during the training. The robustness of ViTs and its variants against adversarial attacks has not been widely investigated in the literature like CNNs. This work studies the robustness of ViT variants 1) against different Lp-based adversarial attacks in comparison with CNNs, 2) under adversarial examples (AEs) after applying preprocessing defense methods and 3) under the adaptive attacks using expectation over transformation (EOT) framework. To that end, we run a set of experiments on 1000 images from ImageNet-1k and then provide an analysis that reveals that vanilla ViT or hybrid-ViT are more robust than CNNs. For instance, we found that 1) Vanilla ViTs or hybrid-ViTs are more robust than CNNs under Lp-based attacks and under adaptive attacks. 2) Unlike hybrid-ViTs, Vanilla ViTs are not responding to preprocessing defenses that mainly reduce the high frequency components. Furthermore, feature maps, attention maps, and Grad-CAM visualization jointly with image quality measures, and perturbations' energy spectrum are provided for an insight understanding of attention-based models.	Ahmed Aldahdooh, Wassim Hamidouche, Olivier Deforges	2021
			Vision transformers (ViTs) have recently demonstrated state-of-the-art performance in a variety of vision tasks, replacing convolutional neural networks (CNNs). Meanwhile, since ViT has a	Bum Jun Kim,	

Coming to the results of the given input:

Stage 1: Computer Vision

Stage 2: convolutional neural network image scene detection

Here are the top results:

1. LocalViT: Bringing Locality to Vision Transformers
2. Demystifying Local Vision Transformer: Sparse Connectivity, Weight Sharing, and Dynamic Weight
3. Vision Transformer for Small-Size Datasets
4. Gaze Estimation using Transformer
5. PAConv: Position Adaptive Convolution with Dynamic Kernel Assembling on Point Clouds

The highlighted terms in the results show the co-relation between the query and suggested articles as expected. Vision transformers for locality detection and image scene detections lie under the same problem statement proving the results good enough. But the robustness of this approach has to be evaluated over series of queries and datasets which is another future scope for this project.

Results and findings addressed the following research problem statements

- Understood the hidden topics from the corpus along with how these topics are distributed across the documents with help of topic modelling techniques like LDA and BERT.
- In the above section, we have seen an example of when user searches query our application fetches relevant documents in less span of time.
- Our application most accurate in Computer vision, but there is more work to be done to improve the accuracy of the application on diverse datasets by combining HDFS, PySpark, TF-IDF, LDA, and BERT.

Qc

Getting the relevant information from huge piles of data is a hectic task and this gave us the motivation to come up with an approach that helps in the information retrieval of the relevant research paper based on the user query in the most efficient way in terms of relevancy and response time. There have been classical models like Boolean, vector and probabilistic models that help in retrieving the relevant information. We implement the TF-IDF vector model to retrieve the relevant research paper. The biggest difference that our project makes in this domain is the utilization of big data technology, the Spark Engine, to handle our massive metadata of 1.7 million articles (ArXiv dataset) which includes the fields like abstract, title, authors, category, DOI etc. The usage of Spark engine helps us retrieve the relevant research paper in the most time efficient way. We observe a 60%-time reduction in processing of 28k research articles from the point of user query to its respective research paper retrieval. Implementation of such a framework for the IR purposes would result in a system with high efficiency and accuracy, which are the biggest challenges in this domain and our project aims to rectify them.

Qd

The massive increase in the amount of online text available and the need for access to various sorts of information has revived interest in a wide range of IR-related topics other than just simple and normal document retrieval. However, with data increasing day by day, retrieving the relevant information has been quite challenging. One of the challenges we had in our project was processing 160K articles for only CS 2021 research publications, which took approximately two hours even with PySpark. Some of the other research challenges include, due to synonymy and different related meanings of a word, there is a problem with vocabulary mismatch. The same word might have multiple meanings and if proper contexts are not provided, a search engine may be unable to discern the correct meaning. Handling complex programming data and understanding Natural Language could be two more research challenges. Another challenge faced in our project is, when we searched the full field of Computer Science, our matching criteria were not very accurate. Web documents have a larger possibility of containing incorrect information than traditional IR collections because there are often no content restrictions on anything released on the Web. As a result, retrieving correct content will become very difficult in this case. A better way to tackle how to handle incorrect information is one of the biggest research challenges.

Qe

In the age of Big Data, as more and more computing power is utilized, challenges in domain of researchers is emerging. More and more research organization are putting capital in research and development to create effective and efficient document retrieval systems. Also, the losses in money and time are becoming a pain area as they are slowing down the pace of the progress. Future research opportunities associated with our project of 'document digital search' is integration of cloud technologies such as Databricks, snowflake to raise the bar of speed and efficiency. Researchers can transform this project into fast search engines and recommendation systems. These modifications can assist the institutions and academics grow and learn by getting recommended articles and explore new areas. In addition, researchers can further develop this project to create a quicker database to retrieve data from. Apart from it, researchers can further work on the security aspect of the project to fetch data by assigned privileges to the data. This way we can ensure that sensitive and classified documents are not revealed to unknown

users and confidentiality is maintained in the system. Therefore, there are many dimensions to this project which can be explored and capitalized to forge a creative solution.

References:

https://static.googleusercontent.com/media/research.google.com/en//people/jeff/WSDM09-keynote.pdf
http://kth.diva-portal.org/smash/get/diva2:589139/FULLTEXT01.pdf