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

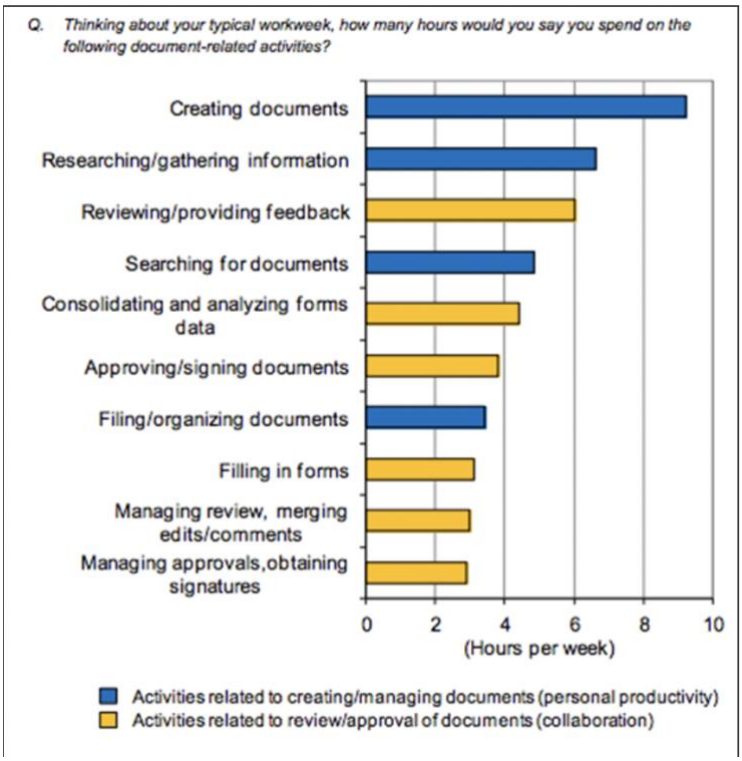
Kolli Sai Nithin Reddy
sainithi@buffalo.edu

Sri Hari Gopinath Thota
sthota2@buffalo.edu

Q1

1.a

As the age of Big Data has set in, every device around us from, a cell phone to a M1 chip engineered MacBook to a smart watch, is generating data all the time. With the onset of this massive data, new challenges pertaining to storage, computation, power cost and energy efficiency have emerged which needs to be addressed. One common problem faced by the academics in this domain is of document retrieval. Within the silos of research papers that are released, it becomes difficult to fetch the most relevant documents accurately and efficiently. Given the circumstance, our project intends to design a document retrieval system which could provide results within fraction of seconds. It is a vital issue because scientific researchers that aim to build on foundations of any domain of science needs to refer, collaborate, and learn from ongoing findings in the world to provide ground-breaking results.
As per popular firm McKinsey, it was discovered that an individual would spent around 1.8 hours per day looking for the right documents which they need. Another major company Gartner found that it takes average 18 minutes to find the exact documents. Surveys conducted by IDC came to conclusion that an employee spends almost 8.8 hours/week looking for documents. Therefore, we not only want to reduce upon this effort and time but also enhance efficiency of the system.



Source:https://resources.m-files.com/blog/how-long-does-it-actually-take-to-find-a- document-dissecting-the-many-stats-out-there

Data driven solutions can be crucial to address this problem as data driven solutions such as Hadoop, Apache Spark etc. offer the option to ingest data not only from a variety of data sources but also deliver with high performance. Data driven solutions are equipped with features such as scalability, high throughput, high compatibility which makes retrieval of data easier and faster.

1.b

The advent of Big data era has raised new challenges and issues pertaining to data size. The world is moving away from paperless to digital platform which is raising new challenges in terms of data management and storage. The traditional tools and technologies are not equipped to handle, create, and process the data of vast size. With this big data, academic domain is also affected since the time and pressure for producing research results has shrunk. Therefore, new research areas such as data compression is gaining popularity as storage is another issue that comes with Big data. Similarly, many new research fields have opened up with the rise of Big Data. Our project not only provides a basic search mechanism but also facilitates the scope of future by adding multiple layers on top of it. New avenues have definitely opened up such as text classification, recommendation systems or semantic visualizations.

Potential challenges would include finding the right context of a word while searching. The principles, concepts or theorems have different or synonymous names across domains. Understanding the context behind application of the word in a sentence would be another area of research that has come to light. On the other hand, probable challenges include storage and power optimization. Although we have come a long way working our way from storing data in floppy disks to migrating to cloud, we still have a long way to go. Another obstacle is the performance of the computation. Though we have increased the processors speed on hardware chips, the performance seems to have reached saturation where we are unable to bring any drastic change in current eco-system. In addition, not everyone can afford the cost that comes while assembling the equipment required for Big data. Questions such as reliability, sustainability and usability still exist.

1.c

With companies investing more in research and development to deliver innovation, organizations and individuals are in constant need of effective and efficient document retrieval at massive scale. And this requirement is only going to go up with each passing day. In addition, often, the existing technology is facing the challenge of integration with the Big Data implementation.
The loss in terms of time, money and unused resources is becoming a major hotspot for firms to rectify and remediate immediately. One common problem faced during fetching documents is the room for improvement in searching algorithms. Many times, we are not able to retrieve exactly what we require due to the algorithmic loopholes which don't match up to the requirements of the user. Therefore, this project can integrate algorithms to suit different user needs. Moreover, when a student is setting out to learn a topic, he is exploring documents which closely aligns with the topic of his/her interest. By adding layers to our project, there is a vast potential of creating a recommendation system which can offer articles by top of the field. Apart from it, there is a very close attention paid to the critical research and findings taking place in an institute. Authorities ensure to provide limited access to these classified and sensitive documents which can be resolved by our system. Our project can be designed to safekeep the sensitive documents with access only to authorized personnel's and not display these documents under results for layman. This will take out the needs for establishments for shelling out on security tools to safeguard their classified data.
Overall, we can see that our project capitalizing on Hadoop and the storage features of HBase distributed database, will address the issues faced above by the end-users in the EdTech and research domains. The project can properly manage the database and retrieve results with blazing speed to increase the efficiency of the system.

Q2

We have reviewed research papers published in the domain of information retrieval especially document retrieval as the work in these field is directly related to our project which is document search from pool of research papers. We understood intrinsic challenges and uncertainties and ways to mitigate these using major advancements in information retrieval using TF-IDF, BERT, Word2Vec and LDA techniques combined with different ML algorithms for clustering. Let us understand about these challenges and techniques in depth below.

The explosion of unstructured, semi-structured, structured, or heterogeneous data being created and stored, has resulted in a growth in applications that incorporate new types of information, such as multimedia and scientific data. Uncertainty is that it's not easy to define a user's information need and users frequently have a hazy understanding of the information they require. A query is nothing more than a hazy and partial representation of the information required [1]. Some of the obstacles are large-scale, as a result of the vast number of Webpages available on the Internet, as well as inheriting any text-based information retrieval system. Incorporating techniques from Information Retrieval will play a significant role in constructing DSSP because certain operations on Dataspace naturally include some degree of data uncertainty. Future work will focus on developing a new acceptable information retrieval model that can be employed in Dataspace to improve query outcomes.

In natural language processing applications, vectorization is crucial for converting textual content into meaningful numerical representations, for performing multi-document summarization. In the paper [2], researchers proposed different vectorization models to communicate with the machines for performing Natural Language Processing tasks on news articles related to top trending topics. Models include term frequency-inverse document frequency (TF-IDF), GloVe, Deep learning, Word2Vec, SentenceToVec. Modern models are ELMo and BERT. The authors summarize multiple documents into a single narrative using a Hybrid Multi-Document Summarization, using Deep Learning architecture, with a cascade of Abstractive and Extractive summarization approaches. The proposed work introduces three vectorization methods for capturing the semantic similarity of news articles to identify unifiable groups. The experimental results indicate that tf-idf vectorization produces more appropriate clusters than Word2Vec and Doc2Vec. Challenges are clusters should be refreshed upon the arrival of new chunks of articles and limitation in handling corpora with continuously changing vocabularies. The authors investigated existing text vectorization methods and proposed to apply multi-level word embeddings using ELMo and BERT for clustering news articles.

The processing of structured or semi-structured data is becoming increasingly complex in all businesses as the volume of data has expanded dramatically. TF-IDF is one such approach [3]. It's a numerical metric that illustrates how relevant keywords are to certain texts. The TF-IDF method combines two approaches. Term Frequency is a metric for counting how many times a term appears in a document. The Inverse Document Frequency gives less weight to frequently occurring words and more weight to infrequently occurring ones. The term frequency (TF) and the inverse document frequency (IDF) are simply multiplied (IDF). The main limitation of TF-IDF is that it cannot identify words even if the tense is changed slightly. To address the challenges, you can use the stemming process and stop word removal to ensure that you get useful terms as output. To achieve even better results, TF-IDF can be integrated with other approaches such as Naive Bayes.

In [4], Kim and Gil proposed research paper classification system based on Term Frequency – Inverse Document Frequency (TF-IDF) and Latent Dirichlet Allocation (LDA) schemes. This work uses Hadoop for TF-IDF, Spark MLlib in Python for LDA calculations and K-means clustering using Scikit-learn library for clustering vast number of research papers based on extracted frequently occurring keyword using TF-IDF and topics extracted from abstract by LDA topic modelling. The method used in this paper clusters the research papers more accurately as it is using correlation between keywords and topics extracted from TF-IDF and LDA techniques. If we only use keywords input by user, it might

not be guaranteed that these keywords are always correct to group papers with similar subjects. On the other hand, LDA extracts topics automatically from abstract. But this work did not use diverse types of datasets across multiple disciplines like math, physics, electronics, biology etc. for their study. So, there is scope for further research which can include research papers from multiple disciplines.

In this work [5], the author discusses about dealing with typos in passage retrieval. Most of current effective approaches rely on deep language model-based retrievers and rankers that have been pre-trained. Here we state the facts that the traditional models such TF-IDF and BM25 use exact keywords and these limits the capability to retrieve passages that are semantically relevant but use different keywords. The techniques DR and BERT re-ranker have been introduced and they use the latent embedding space to estimate the relevance of a query to a passage. We mainly aimed at 3 research questions for this work 1. Impact of typos in queries, 2. impact of typos on effectiveness on BERT, 3. Does proposed typo awareness training impact effectiveness. Different typo errors have been introduced using techniques like randinsert, randdelete, swap adjacent etc. These models are evaluated on MS MARCO dataset using large-bert-uncased and the empirical results are discussed. The results showed decrease in losses in MRR@10 halve for DR (from 52:3% to 24:3%) and reduce by one third for BERT re-ranker (from 34% to 22:7%).

References

1. https://www.researchgate.net/publication/306124580_Information_Retrieval_System_and_challenges_with_Dataspace
2. https://thesai.org/Downloads/Volume10No7/Paper_42-Vectorization_of_Text_Documents.pdf
3. (PDF) Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents (researchgate.net)
4. Kim and Gil Hum. Cent. Comput. Inf. Sci. (2019) https://doi.org/10.1186/s13673-019-0192-7
5. https://arxiv.org/abs/2108.12139