Mid-Term Project Review 1

Group 61

Updated baseline results (system/prototype)

In the previous report we had made strides in preprocessing the data and being able to generate word clouds as part of the retrieved system analytics.

In this deadline, we planned to create a robust search and retrieval system as well as a summarizer model based on LLM architecture.

Retrieval System

For optimisation purposes, after experimenting, we decided to augment our data and narrowed our search space to the titles and abstracts of the scientific papers we needed to search for.
For our document retrieval system, we searched and implemented a cutting-edge scientific paper: Mandikal, P., & Mooney, R. (2024). Sparse Meets Dense: A Hybrid Approach to Enhance Scientific Document Retrieval. arXiv preprint arXiv:2401.04055.
Link:[2401.04055] Sparse Meets Dense: A Hybrid Approach to Enhance Scientific Document Retrieval (arxiv.org)

The paper presents a hybrid approach for enhancing scientific document retrieval by combining traditional sparse bag-of-words vector representations with dense embeddings from a state-of-the-art transformer-based language model called SPECTER2.

The key techniques and findings discussed in the paper are:

1. Sparse Retrieval Model: The paper uses a classical vector space model with TF-IDF weighted bag-of-words vectors and cosine similarity for sparse document/query representations.

2. Dense Retrieval Model: The paper employs SPECTER2, a transformer-based model pre-trained on a large corpus of scientific documents using contrastive learning on citation networks to produce dense document embeddings.

3. Hybrid Retrieval Model: The paper proposes a hybrid approach that combines the sparse and dense retrieval models by taking a weighted sum of their document-query similarity scores.

4. Experiments on a cystic fibrosis document retrieval dataset show that the hybrid model significantly outperforms both the sparse and dense models individually on precision-recall and NDCG metrics.

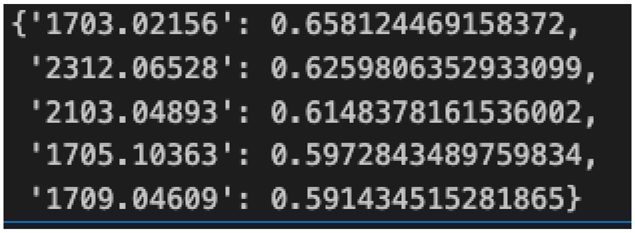
5. The optimal performance is achieved by weighting the dense SPECTER2 embeddings higher (λ=0.8) than the sparse vectors (λ=0.2) in the hybrid model.

6. Ablation studies show the SPECTER2 base model works better than explicitly trained adapter versions on this dataset.

The key novelty is demonstrating the benefits of integrating classical sparse vector techniques with modern dense embeddings from large language models in a straightforward hybrid approach for specialized scientific document retrieval.

Using this hybrid approach, we were able to get our top 5 results, and the model proved effective as a retrieval tool, the only limitation being the low computational capabilities available, making it not possible to run these retrieval models on a larger set of data.

The table below shows the id of paper and hybrid cosine similarity of the retrieval model when prompted with the query “Computer vision and AI.”



```
{'1703.02156': 0.658124469158372,
 '2312.06528': 0.6259806352933099,
 '2103.04893': 0.6148378161536002,
 '1705.10363': 0.5972843489759834,
 '1709.04609': 0.591434515281865}
```

Summarisation System

Using the top results of the system, we aimed to provide a summary of the top results at a glance(in the GUI) so that searching for relevance is easier as well as better. Using the ID of the paper, we retrieve the pdf to scan for its text and other parameters. After that, it is preprocessed and tokenised. Using a pre-trained LLM model we generate the summary of our given paper. Currently, we are experiencing difficulties getting coherent responses from LLMs, implying that we need to fine-tune them for our own usage so we can develop a better summarization feature

```
“Work on learning transferable representations is presented at the
ICLR 2017 workshop on the limits of representation learning with
label-based supervised learning (LSD). The work is presented by
Jiaming Song, Russell Stewart, Shengjia Zhao & Stefano Ermon with the
help of the Stanford Computer Science Department.
we show that supervised learning is upper-bounded in its capacity for
representation learning in ways that certain generative models, such
as Generative Adversarial Networks (GANs Goodfellow et al. (2014))
are not. We hope that our analysis will provide a rigorous motivation
for further exploration of generative representation learning.
some domain (e.g. com-
 vision), and there exists a set of good features Fg that we would
like to learn from a dataset that contains an exponentially large set
of bad features that we want to learn. The goal is to learn all the
good features from the dataset in the process of using a deep neural
network to perform certain tasks…….”
```

This is a snippet from our current summarization model, which is based on a paper with arxiv id 1703.02156

The proposed method (features/ data analysis)

Going forward, our group plans to do the following tasks to improve and augment our system.

Topic modelling and clustering

Why would we like to do topic modelling?
Discovering Trends: By analyzing the distribution and evolution of topics over time, researchers can identify emerging trends or shifts in focus within their field of study. This can inform future research directions or highlight areas of increasing importance.

Organizing and Summarizing Information: Topic modelling can aid in organizing and summarizing large amounts of textual data. By clustering related documents together based on their thematic similarities, researchers can effectively navigate through the corpus and extract key insights more efficiently.

Enhancing Information Retrieval: Once topics are identified, they can be used to improve information retrieval systems. By associating documents with relevant topics, researchers can better search, filter, and prioritize information based on their specific interests or needs.

To perform topic modelling effectively, we can follow a structured approach:

Data Preprocessing: This involves cleaning and transforming raw text data into a format suitable for analysis. Steps include removing noise (stopwords, punctuation), normalizing text (lowercasing, lemmatization), and converting text into numerical representations (such as TF-IDF vectors or word embeddings).

Dimensionality Reduction: Optionally, high-dimensional representations like word embeddings can be reduced to lower dimensions using techniques like PCA, t-SNE, or UMAP. This step helps in improving computational efficiency and sometimes enhances the quality of clustering.

Clustering: Grouping related documents together using clustering algorithms like K-Means, DBSCAN, or HDBSCAN. The number of clusters can be determined using techniques like the elbow method or silhouette analysis.

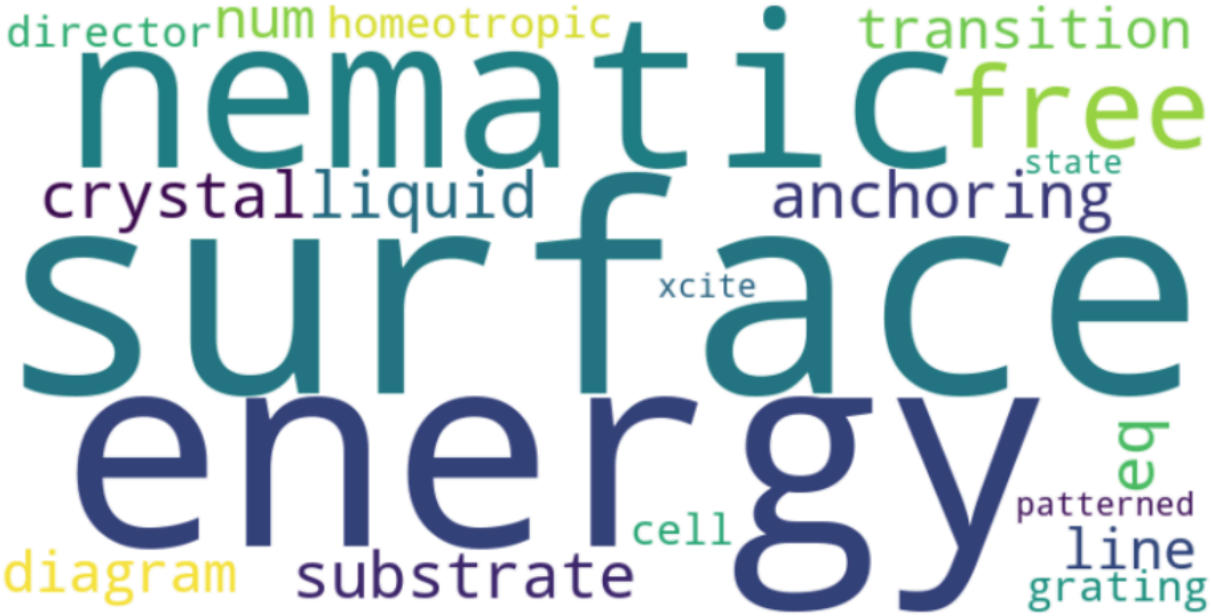
Topic Extraction: Within each cluster, applying topic modeling algorithms like Latent Dirichlet Allocation (LDA) or Non-Negative Matrix Factorization (NMF) are used to extract the prominent themes or topics. These algorithms automatically identify groups of words that frequently co-occur within documents.

Evaluation and Interpretation: Assessing the quality and coherence of extracted topics using metrics like topic coherence or manual inspection. Interpreting and labelling the topics based on domain knowledge is essential for meaningful analysis.

Analysis and Visualization: Analyzing the extracted topics to understand trends, relationships, and patterns within the corpus. Visualizing topics using techniques like pyLDAvis or word clouds can aid in the interpretation and communication of results.

GUI and visual aids

By using a GUI, we want to provide a better experience for the users of the project with sophisticated modalities and interactivity.
Summarisation and word cloud generation are some of the techniques we have tried to see so that we can better the context of the results that the user searches for.



Seeing graphs and word clouds similar to this can, at a glance, provide the user with contextual information about the paper, and a decision on whether it is worth reading can be made faster. Our team believes that giving more such analytical tools as well as a summary of the paper at an accessible glance will provide the user with more capabilities and make their experience as well as decision-making better.