ECS736P - Information Retrieval

Assignment 2 - Design of a Search Engine (Group 18)

Hannah Melkemaryam Claus, Berkay Dur, Stephanie Nicole Garibay Lim, Iqbal Singh

Introduction

In the context of news articles, IR systems are designed to help users find specific news articles or related news items quickly and easily. Saracevic (2010) highlights the importance of relevance in IR, stating that it is critical to ensure that users receive the most relevant and accurate information possible based on their search criteria.

The accuracy and timeliness of news articles are critical factors in shaping public opinion and influencing decision-making processes. Inaccurate or incomplete information can lead to misinformation and misinterpretation, resulting in negative consequences. IR systems can help address these concerns by ensuring that news articles are accurate, reliable, and timely. Thus, by designing and implementing such a system, the quality can be improved to ensure safe and accurate searches.

This report will only focus on proposing the design of an information retrieval system for news articles. It also includes organisational information regarding the future implementation of the system.

The target implementation and evaluation of the IR system is to create two levels of complexity in search results. The first layer concentrates on general news. Here the query will be applied to all available articles. The second layer will focus on one particular news topic, sports, thus, experimenting with information on a telescoped dataset.

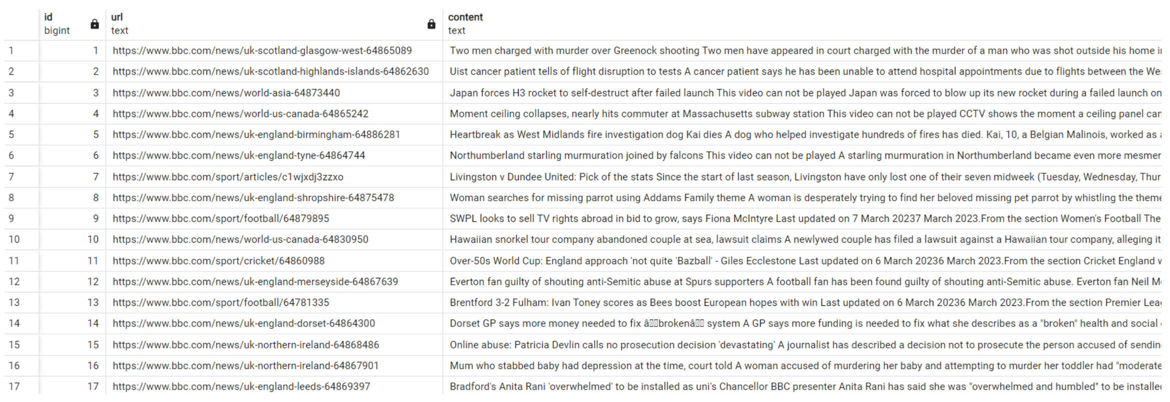
Dataset

The project will use a collection of BBC News articles that is scraped using Python. The news articles are under “/news” and “/sports”.

The project will face some limitations from using its own dataset, specifically, there are no queries and as such no relevancies which are required for testing. A few ways of overcoming this problem are:

- Self-labelling
- Labelling by ensembling the results of existing IR models
- Using the news category as a query (as given in the URL)
- News article title as a query

Due to the nature of the data (High amount of documents but no queries), the DESM (Dual Embedding Space Model) can only be trained on the Document Bodies. (Nalisnick et al., 2016)



	id	url	content
	bigint	text	text
1	1	https://www.bbc.com/news/uk-scotland-glasgow-west-64865089	Two men charged with murder over Greenock shooting Two men have appeared in court charged with the murder of a man who was shot outside his home i
2	2	https://www.bbc.com/news/uk-scotland-highlands-islands-64862630	Uist cancer patient tells of flight disruption to tests A cancer patient says he has been unable to attend hospital appointments due to flights between the We
3	3	https://www.bbc.com/news/world-asia-64873440	Japan forces H3 rocket to self-destruct after failed launch This video can not be played Japan was forced to blow up its new rocket during a failed launch on
4	4	https://www.bbc.com/news/world-us-canada-64865242	Moment ceiling collapses, nearly hits commuter at Massachusetts subway station This video can not be played CCTV shows the moment a ceiling panel car
5	5	https://www.bbc.com/news/uk-england-birmingham-64886281	Heartbreak as West Midlands fire investigation dog Kai dies A dog who helped investigate hundreds of fires has died. Kai, 10, a Belgian Malinois, worked as
6	6	https://www.bbc.com/news/uk-england-tyne-64864744	Northumberland starling murmuration joined by falcons This video can not be played A starling murmuration in Northumberland became even more mesmer
7	7	https://www.bbc.com/sport/articles/c1wjxdj3zzxo	Livingston v Dundee United: Pick of the stats Since the start of last season, Livingston have only lost one of their seven midweek (Tuesday, Wednesday, Thur
8	8	https://www.bbc.com/news/uk-england-shropshire-64875478	Woman searches for missing parrot using Addams Family theme A woman is desperately trying to find her beloved missing pet parrot by whistling the theme
9	9	https://www.bbc.com/sport/football/64879895	SWPL looks to sell TV rights abroad in bid to grow, says Fiona McIntyre Last updated on 7 March 20237 March 2023.From the section Women's Football The
10	10	https://www.bbc.com/news/world-us-canada-64830950	Hawaiian snorkel tour company abandoned couple at sea, lawsuit claims A newlywed couple has filed a lawsuit against a Hawaiian tour company, alleging it
11	11	https://www.bbc.com/sport/cricket/64860988	Over-50s World Cup: England approach 'not quite 'Bazball' - Giles Ecclestone Last updated on 6 March 20236 March 2023.From the section Cricket England v
12	12	https://www.bbc.com/news/uk-england-merseyside-64867639	Everton fan guilty of shouting anti-Semitic abuse at Spurs supporters A football fan has been found guilty of shouting anti-Semitic abuse. Everton fan Neil M
13	13	https://www.bbc.com/sport/football/64781335	Brentford 3-2 Fulham: Ivan Toney scores as Bees boost European hopes with win Last updated on 6 March 20236 March 2023.From the section Premier Lea
14	14	https://www.bbc.com/news/uk-england-dorset-64864300	Dorset GP says more money needed to fix â▯▯brokenâ▯▯ system A GP says more funding is needed to fix what she describes as a "broken" health and social
15	15	https://www.bbc.com/news/uk-northern-ireland-64868486	Online abuse: Patricia Devlin calls no prosecution decision 'devastating' A journalist has described a decision not to prosecute the person accused of sendin
16	16	https://www.bbc.com/news/uk-northern-ireland-64867901	Mum who stabbed baby had depression at the time, court told A woman accused of murdering her baby and attempting to murder her toddler had "moderate
17	17	https://www.bbc.com/news/uk-england-leeds-64869397	Bradford's Anita Rani 'overwhelmed' to be installed as uni's Chancellor BBC presenter Anita Rani has said she was "overwhelmed and humbled" to be installe

Figure A: Sample SQL data

The set of document bodies will consist of news articles that were scraped over the course of 10 days. This 10-day window was chosen because it allows for a range of queries that could yield different results, for example “who scored in the Arsenal game 6 days ago”, or “Trump mar a lago”.

Architecture

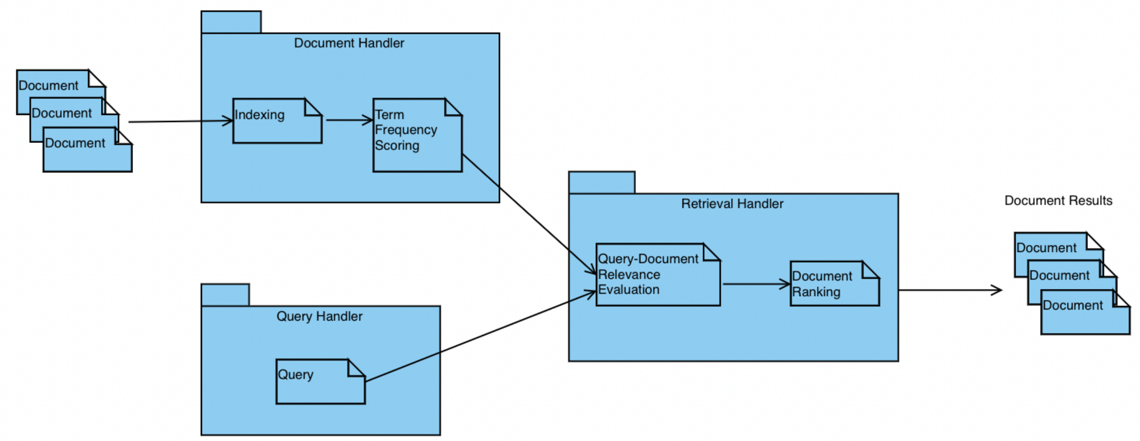


Figure B: Search Engine Framework

Description of the Search Engine Framework:

The Document Handler handles the document processing to help carry out the relevant document retrieval. The list of tasks it goes through are as follows:

1. Capture documents from the database
2. Carry out pre-processing of each document
3. Use the invert-index of the document
4. Implement the term frequency scoring in the document. This project will consider using the model BM25 as a baseline

The Query Handler handles the pre-processing of the query.

The Retrieval Handler handles the document retrieval based on the document-query relevance. The list of tasks it goes through are as follows:

1. Identification of Query-Document relevance. This involves the evaluation of the precision of how relevant the document is to the given query. Usage of recall, F1 scoring and precision will be used for evaluation. In addition, the use of the DESM model will be considered in identifying the query-document similarity. (Nalisnick et al., 2016)
2. The document will then be ranked and sorted in accordance with the evaluation results, the sorting will result in documents of higher relevance being placed on top of the list.

Finally, the sorted relevant documents will then be showcased to the user.

To evaluate which models perform better (i.e. BM25 or DESM), the project proposes to use the normalised discounted cumulative gain (NDCG) to calculate a quantitative score (Järvelin and Kekäläinen, 2017).

Software

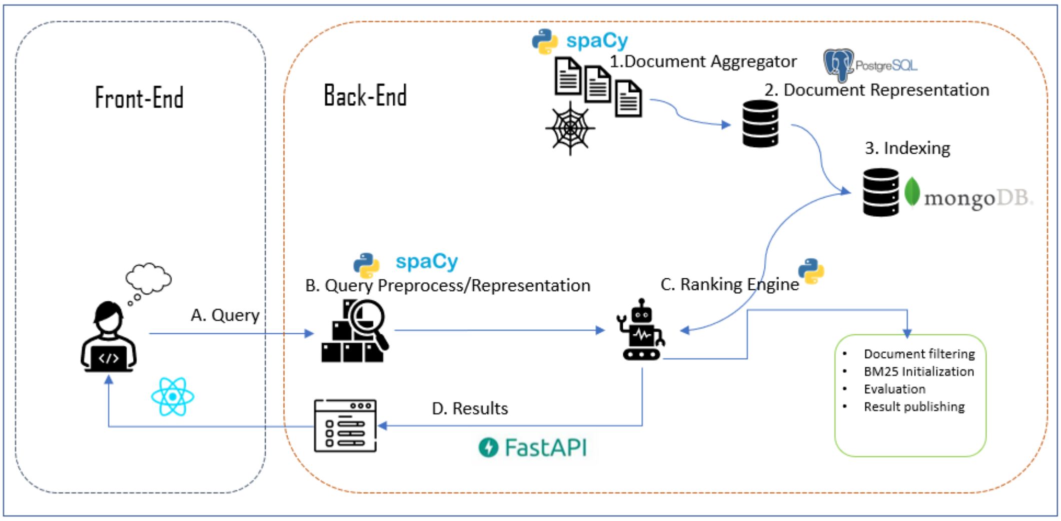


Figure C: Tech stack for search engine implementation

The technical implementation consists of a two-part process - part A, being the front end which consists of a user who types in the query and part B, the backend, which receives the query and returns the relevant news to the user through FastAPI.

The brief descriptions of the software used for each part and its sub-processes are below:

Area	Process	Software	Remarks
Front-End	• Capturing Get Response	FastAPI	Used to capture user query and to send it for back-end processing

Back-End	● Document aggregator ● Document Representation	Tech used: Python, PostgreSQL Libraries: Spacy, Beautiful Soup, psycopg2	**Beautiful soup** is used for document aggregation process, and text pre-processing steps, such as tokenisation, stemming, and stop word removal is done by leveraging on **Spacy** library. **psycopg2** is used to push clean data into the database. **Postgres** is utilised to store the articles along with its URL. Postgres is fast and has the ability to process long text data.
	● Indexing	Tech used: MongoDB, Python Libraries:psycopg2, pymongo	**MongoDB** is used to store inverted indexes, given its key-value-like data structure. Python is used for data creation, and MongoDB is used for storage. Feeding into MongoDB is done via the **pymongo** library and fetching articles from Postgres is done by **psycopg2**
	● Ranking Engine	Tech used: MongoDB, Python Libraries: psycopg2,pymongo, rank_bm25	Based on the user query, the recall set is derived using an inverted index, this is done by leveraging **MongoDB** where the index is stored. **pymongo** library connects python to **mongoDB** allowing data flow. **rank_bm25** library to create the term-frequency table and also to compute the relevancy score based on a pre-processed query. Model evaluation between BM25 and DESM will be using Python and human feedback. This evaluation result will be captured and stored in the database.
	● Query Preprocess	Tech used: Python Libraries: Spacy	**Spacy** is used in pre-processing of user query
	● Results	Tech used: Python Libraries: FastAPI	**FastApi** is used to publish results to the front-end and to capture human feedback for model evaluation.
Other Tools	Code Management	GitHub	
	Prototyping	Jupyter	Used for functional prototyping of the code.

	ML process	Pytorch, Tensorflow	Incorporation of Embedding-based search methods.

Roles and Responsibilities

Weekly Team Meetings: every Monday 1pm-4pm, every Wednesday during the labs

	Member	Role and Responsibility
1	Berkay Dur	Database Analysis and collection, Document Preparation, Implementation of retrieval framework
2	Hannah Melkemaryam Claus	Project Planning and Organisation, Literature Review, Document Preparation, Implementation of retrieval framework
3	Iqbal Singh	Analysis of tools and libraries, Document Preparation, Initial Search Engine Design Proposal, Implementation of retrieval framework
4	Stephanie Nicole Garibay Lim	Architecture Design, Document Preparation, Implementation of retrieval framework

Timeline

Weeks Left	Date	Tasks
7	27 Feb - 5 Mar	● Do research on potential datasets and models to implement ● Discuss which dataset to use ● Discuss which model to use ● Create a first draft of the design report
6	6 - 12 Mar	● Choose final dataset and model to use during the implementation ● Assign roles and responsibilities ● Test potential software to use during the implementation ● Receive permission to use the dataset ● Create a GitHub repository to document and store the design and future implementation: https://github.com/melkemaryam/search_engine ● Finish the design report and submit
5	13 - 19 Mar	● Setup environments for the implementation of the model ● Acquire required tools ● Familiarise with the model
4	20 - 26 Mar	● Develop the base model ● Evaluate the results ● Start developing new model

3	27 Mar - 2 Apr	● Continue developing new model ● Evaluate new results ● Experiment with parameters ● Compare different performances ● Create the first draft of the presentation slides
2	3 - 9 Apr	● Test and troubleshoot code ● Incorporate feedback from demonstrators and lecturer ● Create the second draft of the presentation slides
1	10 -12 April	● Finalise the presentation slides ● Finalise the code, add missing comments and explanations ● Record the demonstration video ● Finalise the report and submit

Investigation

The project uses the findings of Nalisnick et al. (2016) to build a News article Information Retrieval system. This paper suggests that using BM25 and DESM in a Mixture model results in a better performance than either model separately in a more general setting. To further test the findings of this paper, this project looks at using DESM in a more telescopic setting. So, the performance will be evaluated on two different levels of complexity:

1. Broad dataset of news articles
2. Telescoped dataset of news articles with the topic "sports"

The models we will look at are:

1. BM25
2. DESM
3. BM25 + DESM Mixture

References

- Saracevic, T. (2010). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. Journal of the American Society for Information Science and Technology, 61(2), 207-237.
- Nalisnick, E. et al. (2016) “Improving document ranking with dual word embeddings,” Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion, pp. 83–84. Available at: https://doi.org/10.1145/2872518.2889361.
- Järvelin, K. and Kekäläinen, J. (2017). IR evaluation methods for retrieving highly relevant documents. ACM SIGIR Forum, 51(2), pp.243–250. doi:https://doi.org/10.1145/3130348.3130374.