PROFILE BASED-RETRIEVAL

Ander Ros, Sergio Marin, Irune Monreal

INDEX:

Introduction

In this project, we embark on the journey of information retrieval using a dataset sourced from the BBC. We aim to facilitate user access to relevant news articles based on their specified interests. To accomplish this, we have employed two distinct methodologies: TF-IDF (Term Frequency-Inverse Document Frequency) and embeddings.

The dataset encompasses a diverse array of news articles spanning various topics, including politics, economics, sports, and entertainment. Upon receiving user input regarding their topic of interest, our system employs advanced algorithms to analyze the textual content of the articles and recommend the most relevant ones.

TF-IDF, a traditional technique in information retrieval, assesses the significance of a term within a document relative to its occurrence across the entire corpus. By leveraging TF-IDF scores, we can identify the most distinct terms within each document, thereby facilitating accurate assessments of document similarity.

In addition to TF-IDF, we have explored the utilization of embeddings—a state-of-the-art approach in natural language processing. Embeddings transform words or documents into dense, continuous vectors in a semantic space, capturing subtle linguistic nuances and semantic relationships. By comparing the embeddings of user-specified interests with those of the articles in our dataset, we can discern semantic similarities and provide personalized recommendations tailored to the user's preferences.

Through this comparative analysis of TF-IDF and embeddings, we aim to elucidate their respective strengths and weaknesses in the context of news article retrieval, ultimately enhancing information access and user experience.

Methodology

Document preprocessing and encoding method

Data preprocessing is a critical step in natural language processing (NLP) workflows to clean and standardize textual data for effective analysis and modeling. In our study, we applied several preprocessing techniques using the NLTK library to prepare our text data.

In the initial phase of our data preprocessing pipeline, all text data underwent a standardization process to enhance consistency and prepare it for subsequent natural language processing tasks. To begin with, we applied lowercase transformation uniformly across the entire dataset. This step prevents case-sensitive discrepancies during analysis.

Following the lowercase conversion, the text underwent tokenization using a regular expression tokenizer, which isolated meaningful tokens by splitting the text into words while disregarding non-alphanumeric characters. Concurrently, we removed punctuation marks from the tokenized text to eliminate noise and maintain only relevant words for analysis.

Moreover, we carried out stopwords removal to filter out common English words that do not contribute significantly to the semantics of the text. By excluding stopwords, we focused on retaining informative content words essential for downstream analysis.

In our data preprocessing pipeline, we initially considered both stemming and lemmatization to standardize word forms and reduce token variation. Stemming simplifies word forms to their root, promoting text coherence and aiding subsequent analysis.

However, we ultimately chose lemmatization over stemming to reduce words to their base or dictionary form. This decision ensured a standardized vocabulary and minimized token variation, improving alignment and semantic understanding within our text data. By prioritizing lemmatization, we optimized our dataset for effective natural language processing tasks, including document embedding and information retrieval.

By sequentially applying these preprocessing techniques, we ensured that our text data was cleaned, standardized, and optimized for natural language processing tasks. This systematic approach laid a solid foundation for meaningful analysis, document embedding, and information retrieval processes in our study.

Similarity

For our project, we've employed cosine similarity as a fundamental metric for comparing vectors representing both the search query and the documents stored in our database. Cosine similarity is a measure that assesses directional similarity between two vectors ranging from -1 to 1 in a multidimensional space. Essentially, it computes the cosine of the angle between the vectors, providing a measure of how similar they are in terms of their relative orientation. This approach enables us to determine document relevance based on the similarity of their vector representations to the search query. By leveraging cosine similarity, we can identify and retrieve the most relevant documents for a given query, significantly enhancing the effectiveness of our information retrieval system. A higher cosine similarity indicates a stronger resemblance between the vectors, suggesting that the corresponding documents are more pertinent to the user query.

Text representations

TF-IDF Representation

TF-IDF (Term Frequency-Inverse Document Frequency) is a method used to gauge the importance of words in documents. It measures how significant a term is within a document relative to a collection of documents.

TF-IDF is calculated as the product of Term Frequency (TF) and Inverse Document Frequency (IDF). The TF-IDF scores are computed for each term-document pair, resulting in a matrix where each row corresponds to a document and each column corresponds to a term. The values in the matrix represent the TF-IDF scores, reflecting the importance of each term in the document collection.

In our study, we employed the TF-IDF vectorizer with specific parameter configurations to preprocess textual data. Specifically, we utilized a maximum document frequency threshold

(max_df) of 0.15 and limited the maximum number of features (max_features) to 1000. The max_df parameter controls the inclusion of terms that occur frequently across documents, aiming to filter out potentially less informative terms. Meanwhile, the max_features parameter restricts the vocabulary size by selecting the top N most frequent terms, where N is set to 1000 in our case. By tuning these parameters, we aimed to strike a balance between capturing relevant information from the corpus and mitigating the effects of noise or irrelevant terms. This preprocessing step laid the foundation for subsequent feature extraction and modeling processes, facilitating the generation of meaningful representations for document retrieval and analysis.

Limitation of TF-IDF and Cosine Similarity

During document retrieval using TF-IDF and cosine similarity, a notable limitation emerges when user queries incorporate words that are absent from the vocabulary utilized for TF-IDF computation. This phenomenon, known as the vocabulary gap, can result in inaccurate relevance assessment due to the exclusion of unfamiliar terms. Essentially, TF-IDF relies on a predefined vocabulary derived from the corpus of documents being analyzed. However, if a user query includes terms that do not exist within this predetermined vocabulary, their semantic relevance may not be effectively captured. Consequently, queries containing relevant but unrecognized terms may be overlooked, diminishing the overall retrieval performance of the system. Addressing this limitation requires strategies such as expanding the vocabulary through techniques like incorporating external knowledge sources like word embeddings. By mitigating the vocabulary gap, we can enhance the robustness and accuracy of the information retrieval process, ensuring a more comprehensive and effective search experience for users.

Mitigating Vocabulary Gaps with Advanced Embedding Techniques

To tackle this limitation, advanced embedding techniques such as Word2Vec or Doc2Vec provide effective solutions. These methods utilize deep learning algorithms to generate dense vector representations for both individual words and entire documents, capturing nuanced semantic meanings and relationships that extend beyond the confines of the original vocabulary. Unlike traditional approaches like TF-IDF, which rely on predetermined vocabularies, embedding techniques enable the modeling of semantic similarities between words and documents based on their contextual usage within a large corpus of text. By learning from vast amounts of data, these models can effectively represent the semantic essence of words and documents, facilitating more accurate similarity assessments even for terms not present in the training vocabulary.

Word Embedding Representation

Word embeddings are dense vector representations of words in continuous vector spaces, trained using neural network models. They capture semantic relationships between words based on their contextual usage and enable various natural language processing tasks

In our study, we employed pre-trained word embedding models, specifically the Word2Vec.
The Word2Vec model trained on the Google News dataset is particularly well-known for its extensive coverage and high-quality embeddings. It has been trained on a vast corpus of news articles, capturing semantic similarities and relationships between words in a wide

range of contexts. This pre-trained Word2Vec model provides a valuable resource for various natural language processing tasks, including sentiment analysis, named entity recognition, machine translation, and information retrieval.

Semantic Representation

Word embeddings represent words as dense vectors in a high-dimensional space, where each dimension corresponds to a specific semantic feature. The distance and direction between vectors encode semantic relationships between words. For example, words with similar meanings or contexts are represented by vectors that are closer together in the embedding space.

Embedding Aggregation

To generate document embeddings from word embeddings, we adopted an aggregation approach. For each document, we first tokenized the text into individual words and obtained their corresponding word embeddings from the pre-trained model. We then aggregated these word embeddings using a simple method such as averaging.

Document Embedding Models

In our study, we also explored the use of document embeddings. Document embeddings are similar to word embeddings but instead represent entire documents as dense vectors in a continuous vector space. These embeddings capture the semantic content and relationships within documents, allowing for more holistic representations of document content.

For our experimentation, we employed techniques such as Doc2Vec, which extends the principles of Word2Vec to learn embeddings for entire documents. By training Doc2Vec models on our corpus of documents, we generated document embeddings that encapsulate the semantic meaning of each document.

Semantic Representation

Document embeddings, akin to their word-level counterparts, encode semantic information within dense vectors in a high-dimensional space. In this space, each dimension corresponds to specific semantic features or aspects of the document's content. Similar to word embeddings, the relative positions of document vectors capture semantic relationships between documents. Documents sharing similar themes, topics, or content tend to have embeddings that are closer together in the embedding space, while those with differing content are farther apart. This allows for the exploration of semantic similarities and differences between documents, enabling a more nuanced understanding and analysis of document content. By leveraging document embeddings, our information retrieval systems can effectively capture the underlying semantic structure of documents, facilitating the relevance ranking.

Results

In this section, we will discuss the outcomes and performance evaluation of our document retrieval system, which leverages TF-IDF word embedding and document embedding

techniques. The effectiveness of each method will be assessed based on their capability to retrieve relevant documents in response to user queries or interests. We will examine how TF-IDF and embeddings contribute to the retrieval process and evaluate their performance using standard metrics such as precision, recall, R-precision, mean average precision (MAP), and receiver operating characteristic (ROC) curves. These evaluations will provide insights into the strengths and limitations of each approach and inform further optimizations in our document retrieval system.

Queries

Each query in our analysis corresponds to a specific category within the dataset, representing distinct areas of interest. The categories—politics, entertainment, sports, technology, and business—are mapped to these queries based on the identified themes. Each query represents a user with particular interests in one of these categories.

1: ['politics'],
2: ['entertainment'],
3: ['sport'],
4: ['tech'],
5: ['business']

TF-IDF Results

Relevant retrieved documents

Using TF-IDF, we retrieved the five most similar documents to the user query.

User 1, Interests: ['politics']
Document ID: 642, Category: sport, Similarity: 0.57 Wrong
Document ID: 1792, Category: politics, Similarity: 0.49 OK
Document ID: 553, Category: tech, Similarity: 0.32 Wrong
Document ID: 2211, Category: sport, Similarity: 0.27 wrong
Document ID: 825, Category: politics, Similarity: 0.27 OK

Among these documents, Document ID 642 initially categorized under 'sport,' displayed a notable similarity of 0.57, suggesting relevance to the user's interest in 'politics.', this would be considered as an error prediction, but after reading the actual document, we can see that it was initially wrong categorized and that the model predicted correctly the similarity.

Document ID 2211, also initially categorized under 'sport,' exhibited a similarity score of 0.27. However, like Document ID 642, its content was found to be relevant to politics upon review. The document explores teenagers' lack of interest and knowledge in politics, revealing insights into their awareness of political parties and national identity. These instances underscore the challenges of document categorization and the importance of contextual analysis in determining relevance.

Document ID: 642
“Kennedy predicts bigger turnout voters pent up passion could confound predictions of a low turnout in the coming general election charles kennedy has said. the liberal democrat leader predicted concerns over iraq and other international and domestic issue would express themselves

during the campaign. his comments come as an inquiry looks at how best to boost voter turnouts. ex-foreign secretary robin cook said people were not apathetic but fed up of pager politics and not being heard. he like mr kennedy pointed to the hundreds of thousands of people who demonstrated against plans for the iraq war. mr cook who is giving evidence to the power inquiry into voter turnout rates told bbc radio 4 s today programme it was not fair to blame the public who were more interested in politics than ever before . they are turned off by the way we do politics in britain. there s a message there for politicians. he urged politicians to avoid negative campaigning and to speak more from the heart . we should be not so afraid to say what we stand for. he also criticised the cult of personality politics: there s far too much interest in celebrities. politics are in danger of becoming another branch of the celebrity industry. the government has tried a number of things in an attempt to boost voter turnout which fell to 59% in the last general election in 2001. this has included bringing in directly elected mayors to head local authorities and trialling postal voting."

Document ID: 2211

"Teens know little of politics teenagers questioned for a survey have shown little interest in politics - and have little knowledge. only a quarter of 14-16 year olds knew that labour was the government the tories were the official opposition and the lib dems were the third party. almost all could identify tony blair but only one in six knew who michael howard was and just one in 10 recognised charles kennedy. the icm survey interviewed 110 pupils for education watchdog ofsted. nearly half those pupils polled said it was not important for them to know more about what the political parties stand for. and 4% of those questioned thought the conservatives were in power - while 2% of them believed the lib dems were. the survey also looked at issues of nationality. it found the union flag and fish and chips topped the list of symbols and foods associated with being british. many of the pupils also looked on themselves as english scottish or welsh rather than british; while the notion of being european hardly occurred to anyone."

As we can see our tf-idf model is working quite well, understanding

User 1 have interests: [politics]
- Document 642 with similarity score 0.57 and category sport -> WRONG
- Document 1792 with similarity score 0.49 and category politics -> OK
- Document 553 with similarity score 0.32 and category tech -> WRONG
- Document 2211 with similarity score 0.27 and category sport -> WRONG
- Document 825 with similarity score 0.27 and category politics -> OK

User 2 have interests: [entertainment]
- Document 985 with similarity score 0.50 and category entertainment -> OK
- Document 326 with similarity score 0.28 and category politics -> WRONG
- Document 1090 with similarity score 0.26 and category entertainment -> OK
- Document 1147 with similarity score 0.25 and category tech -> WRONG
- Document 184 with similarity score 0.23 and category tech -> WRONG

User 3 have interests: [sport]
- Document 1835 with similarity score 0.63 and category politics -> WRONG
- Document 2184 with similarity score 0.37 and category tech -> WRONG
- Document 1704 with similarity score 0.33 and category business -> WRONG
- Document 306 with similarity score 0.32 and category sport -> OK
- Document 2092 with similarity score 0.31 and category politics -> WRONG

User 4 have interests: [tech]

Document 471 with similarity score 0.00 and category politics -> WRONG
Document 1280 with similarity score 0.00 and category sport -> WRONG
Document 1555 with similarity score 0.00 and category tech -> OK
Document 1234 with similarity score 0.00 and category sport -> WRONG
Document 1874 with similarity score 0.00 and category business -> WRONG
User 5 have interests: [business]
Document 159 with similarity score 0.41 and category politics -> WRONG
Document 627 with similarity score 0.38 and category business -> OK
Document 2185 with similarity score 0.32 and category business -> OK
Document 1732 with similarity score 0.31 and category business -> OK
Document 2012 with similarity score 0.31 and category business -> OK

In conclusion, the utilization of TF-IDF has proven to be an effective tool in our information retrieval study. By employing vectorization and weighting techniques, we were able to accurately represent document relevance based on the frequency and importance of terms present within them. Parameter configurations such as max_df and max_features allowed us to manage the inclusion of uninformative terms and limit the vocabulary size, thereby optimizing the model's performance. Despite some limitations, such as reliance on predefined vocabulary and lack of language semantics, TF-IDF showcased its utility in accurately retrieving relevant documents for users. Furthermore, the integration of TF-IDF with advanced word and document representation techniques, such as word embeddings and cosine similarity, provides a comprehensive approach to tackling the challenges of information retrieval in complex environments. In future research endeavors, we will further explore the potential of TF-IDF and other natural language processing techniques to enhance the accuracy and efficacy of our information retrieval systems.

Evaluation metrics

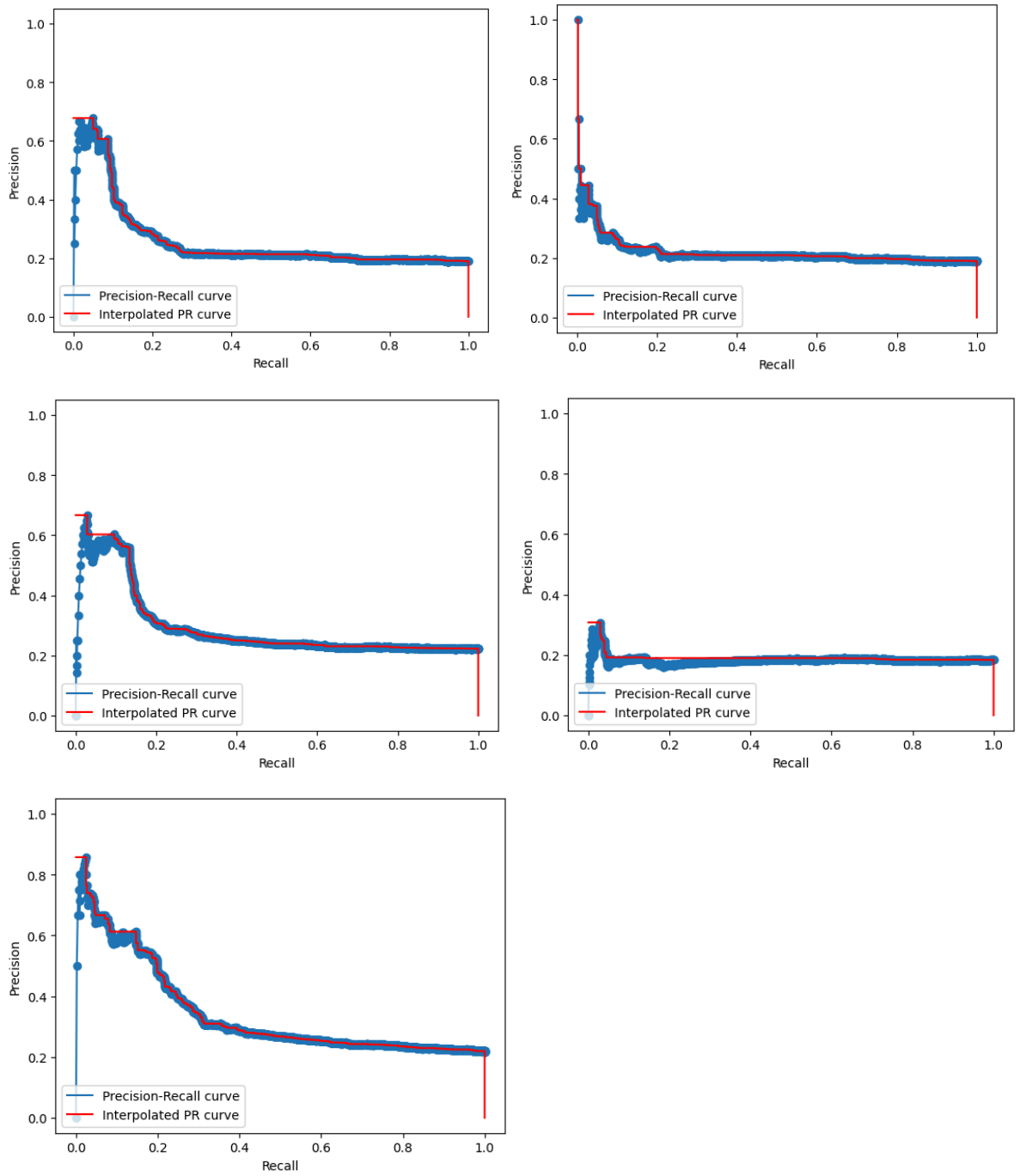
Now we are going to evaluate all the dataset, using the metrics mentioned before.

R-precision

Query	Number of Relevant Documents	R-Precision
1	421	0.24
2	420	0.21
3	493	0.28
4	408	0.17
5	483	0.31

Upon analysis, we observed varying R-precision values across different queries. For instance, Query 5 achieved the highest R-precision of 0.31, indicating that 30% of the retrieved documents were relevant out of the total relevant documents identified for that query. In contrast, Query 4 exhibited a lower R-precision of 0.17, indicating that only 17% of the retrieved documents were relevant compared to the total relevant documents for that query.

prec_rec



The precision-recall curves across all queries demonstrate varying levels of performance. While some queries start with strong precision, maintaining relevance throughout retrieval is challenging. This highlights opportunities for refining retrieval algorithms to ensure consistent and effective document retrieval. There is a notable mismatch between query content and

retrieved documents, emphasizing the need for system enhancements to bridge this gap. These insights underscore the importance of continuous optimization for improving precision and relevance in information retrieval.

MAP

Query	Average Precision (AP)
1	0.26
2	0.22
3	0.29
4	0.18
5	0.34
MAP	0.26

The table displays the Mean Average Precision (MAP) and Average Precision (AP) values calculated for a set of queries used to evaluate an information retrieval system. Each row in the table corresponds to a specific query (Query 1 to Query 5), and the "Average Precision (AP)" column indicates the precision achieved for each query based on the retrieved relevant documents. For instance, Query 1 achieved an AP of 0.26, indicating that on average, 26% of the retrieved documents were relevant to this query. The Mean Average Precision (MAP), calculated as the average of all AP values, is shown as 0.26. This MAP value signifies that the retrieval system achieved an average precision rate of 26% across all queries

Word Embedding Results

Our exploration into Word2Vec's application in document retrieval unveils a nuanced landscape, showcasing its capability to align user interests with retrieved documents while uncovering occasional challenges and disparities.

Relevant retrieved documents

User 1 have interests: [politics]
    Document 642 with similarity score 0.57 and category sport -> WRONG
    Document 1792 with similarity score 0.55 and category politics -> OK
    Document 825 with similarity score 0.50 and category politics -> OK
    Document 623 with similarity score 0.50 and category politics -> OK
    Document 2204 with similarity score 0.50 and category politics -> OK
User 2 have interests: [entertainment]
    Document 1147 with similarity score 0.50 and category tech -> WRONG

Document 930 with similarity score 0.44 and category tech -> WRONG
Document 2059 with similarity score 0.44 and category entertainment -> OK
Document 1183 with similarity score 0.42 and category entertainment -> OK
Document 985 with similarity score 0.42 and category entertainment -> OK
User 3 have interests: [sport]
Document 1359 with similarity score 0.50 and category tech -> WRONG
Document 539 with similarity score 0.50 and category politics -> WRONG
Document 1949 with similarity score 0.50 and category sport -> OK
Document 1426 with similarity score 0.50 and category politics -> WRONG
Document 2092 with similarity score 0.49 and category politics -> WRONG
User 4 have interests: [tech]
Document 1715 with similarity score 0.43 and category tech -> OK
Document 1805 with similarity score 0.41 and category tech -> OK
Document 1555 with similarity score 0.40 and category tech -> OK
Document 1194 with similarity score 0.40 and category tech -> OK
Document 1241 with similarity score 0.39 and category tech -> OK
User 5 have interests: [business]
Document 715 with similarity score 0.55 and category sport -> WRONG
Document 465 with similarity score 0.52 and category business -> OK
Document 1247 with similarity score 0.51 and category business -> OK
Document 159 with similarity score 0.50 and category politics -> WRONG
Document 1931 with similarity score 0.50 and category business -> OK

Upon employing Word2Vec for document retrieval, we observed varied outcomes across different user interests. For User 1, interested in politics, the model yielded mostly accurate results, with four out of five retrieved documents aligning with the user's interest. However, Document 642 was misclassified under 'sport,' despite exhibiting a high similarity score, highlighting a misjudgment by the model, but as we have mentioned before this document is a ‘politics’ document. Similarly, for User 2 interested in entertainment, the model misclassified two out of five documents, but, observing the good performance we suspect a missclasification in the dataset and an actual correct prediction from the model. Conversely, User 3 interested in sport encountered significant misclassifications, with only one out of five documents correctly categorized. These inconsistencies suggest limitations in the model's ability to accurately capture document semantics and user interests, particularly in nuanced domains such as sport. Nevertheless, for users interested in tech and business, the model performed relatively well, with the majority of retrieved documents aligning with their interests. Overall, while Word2Vec shows promise in capturing semantic relationships between words and documents, its effectiveness in information retrieval remains contingent upon the complexity of user interests and the quality of document representations.

Evaluation metrics

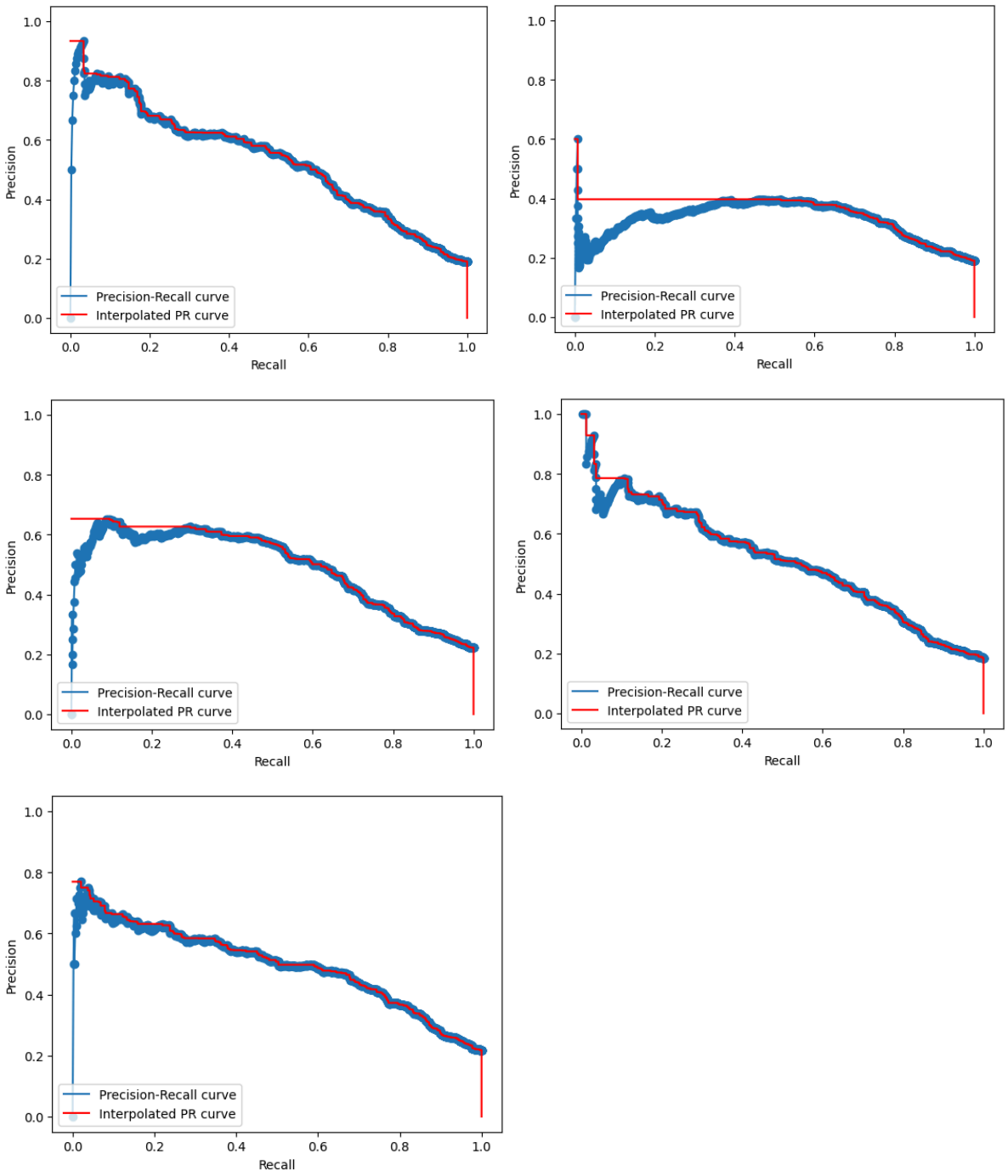
R-precision

Query	Number of Relevant Documents	R-Precision
1	421	0.54

2	420	0.39
3	493	0.54
4	408	0.51
5	483	0.50

The word embeddings-based retrieval method demonstrated improved performance compared to TF-IDF, achieving higher R-precision values across most queries. Notably, Queries showed a significant improvement in R-precision using embeddings compared to TF-IDF, indicating that the embedding approach better captured the relevance of documents for this query.

prec-rec graphics



The precision-recall curves for the second approach exhibit a slower decline in precision as recall increases compared to the initial approach. This slower decrease signifies a more stable and effective retrieval of relevant documents, highlighting the superiority of the embedding-based method over traditional TF-IDF approaches in maintaining relevance across varying recall levels.

MAP

Query	AP Value

1	0.53
2	0.33
3	0.49
4	0.51
5	0.50
MAP	0.47

The Mean Average Precision (MAP) value of 0.47, derived from these AP scores, signifies the overall effectiveness of our information retrieval system across multiple queries. Notably, the embeddings approach demonstrated superior performance compared to the TF-IDF method, as evidenced by the higher MAP value. This improvement underscores the embeddings' capability to capture semantic relationships and context, resulting in more precise and relevant document retrieval for the specified queries.

Based on the analysis of various metrics including R-precision, MAP, and precision-recall curves, it is evident that the embedding-based approach outperforms the TF-IDF method in information retrieval tasks. The embedding-based method demonstrates higher R-precision values and achieves a superior MAP score of 0.47 compared to 0.27 obtained by TF-IDF. Additionally, the precision-recall curves for the embedding-based approach exhibit a more gradual decline, indicating better consistency and effectiveness in retrieving relevant documents across different recall levels. This performance improvement underscores the advantages of using embedding techniques, suggesting that they offer a more robust and accurate means of information retrieval compared to traditional TF-IDF methods. The findings suggest that leveraging embeddings can lead to significant enhancements in retrieval algorithms, contributing to more accurate and relevant document retrieval in information retrieval systems.

Document Embedding Results

Our exploration into Doc2Vec's application in document retrieval unveils a diverse array of outcomes, ranging from successful alignments between user interests and retrieved documents to notable challenges and misclassifications.

Relevant retrieved documents

User 1 have interests: [politics]

Document 1570 with similarity score 0.78 and category politics -> OK
Document 2084 with similarity score 0.78 and category politics -> OK
Document 882 with similarity score 0.77 and category politics -> OK

Document 2207 with similarity score 0.77 and category politics -> OK
Document 1001 with similarity score 0.76 and category sport -> WRONG
User 2 have interests: [entertainment]
Document 1619 with similarity score 0.78 and category business -> WRONG
Document 407 with similarity score 0.75 and category entertainment -> OK
Document 583 with similarity score 0.75 and category entertainment -> OK
Document 503 with similarity score 0.74 and category sport -> WRONG
Document 221 with similarity score 0.73 and category entertainment -> OK
User 3 have interests: [sport]
Document 691 with similarity score 0.74 and category business -> WRONG
Document 1947 with similarity score 0.74 and category sport -> OK
Document 1426 with similarity score 0.73 and category politics -> WRONG
Document 449 with similarity score 0.72 and category sport -> OK
Document 1614 with similarity score 0.72 and category sport -> OK
User 4 have interests: [tech]
Document 583 with similarity score 0.77 and category entertainment -> WRONG
Document 1145 with similarity score 0.77 and category entertainment -> WRONG
Document 634 with similarity score 0.77 and category entertainment -> WRONG
Document 503 with similarity score 0.76 and category sport -> WRONG
Document 221 with similarity score 0.74 and category entertainment -> WRONG
User 5 have interests: [business]
Document 503 with similarity score 0.75 and category sport -> WRONG
Document 715 with similarity score 0.73 and category sport -> WRONG
Document 1619 with similarity score 0.72 and category business -> OK
Document 1973 with similarity score 0.71 and category tech -> WRONG
Document 742 with similarity score 0.71 and category business -> OK

Upon employing Doc2Vec for document retrieval, we observed a mix of results across different user interests. For users interested in politics, sport, and business, Doc2Vec performed relatively well, with the majority of retrieved documents aligning with their interests. However, for users interested in entertainment and tech, the model encountered challenges, with several misclassifications observed. These discrepancies suggest potential limitations in the model's ability to accurately capture semantic relationships between documents, particularly in nuanced domains such as entertainment and tech. Despite these challenges, Doc2Vec provides valuable insights into document retrieval, offering a nuanced understanding of document semantics and user interests.

Evaluation metrics

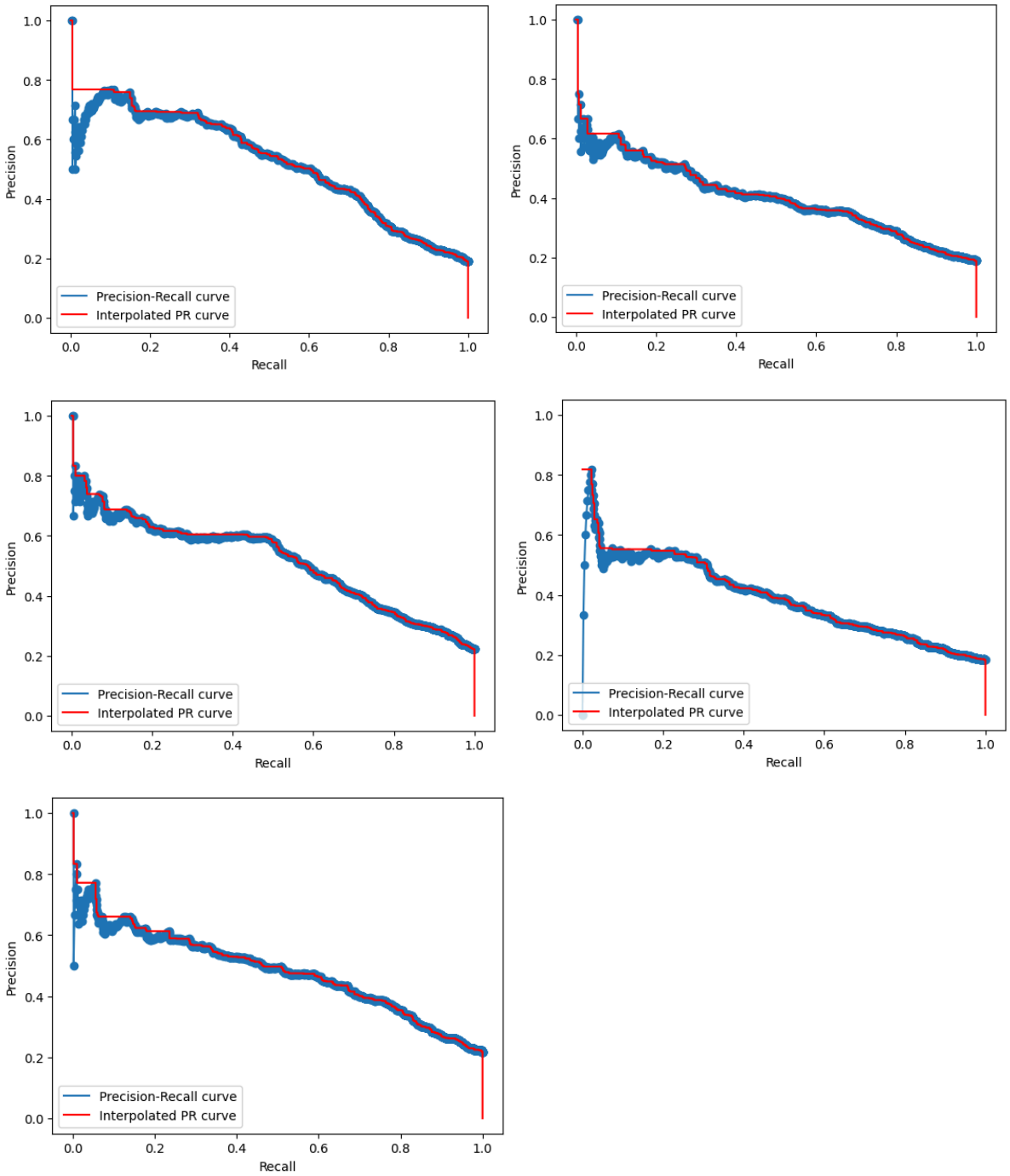
R-precision

Query	Number of Relevant Documents	R-Precision
1	421	0.51
2	420	0.40

3	493	0.59
4	408	0.45
5	483	0.39

The document embeddings-based retrieval method demonstrated improved performance compared to TF-IDF, achieving higher R-precision values across most queries. But, showed similar or even lower values than in the word embedding-based retrieval.

prec-rec graphics



The precision-recall curves for our third approach exhibit even a slower decline in precision as recall increases compared to the initial approaches. This slower decrease signifies a more stable and effective retrieval of relevant documents, highlighting the superiority of the document embedding-based method over traditional TF-IDF approaches in maintaining relevance across varying recall levels.

MAP

Query	AP Value
1	0.30
2	0.31
3	0.46
4	0.37
5	0.52
MAP	0.39

The Mean Average Precision (MAP) value of 0.39, derived from these AP scores, signifies the overall effectiveness of our information retrieval system across multiple queries.In this case the MAP score decreases from the word embedding approach, it becomes evident that while both techniques exhibit variability in their performance across different queries, word embeddings generally outperform document embeddings in terms of average precision. Despite this, both approaches demonstrate potential for improving document retrieval accuracy.

Comparing performance with other groups

In this section, we will use our best-performing approach based on embeddings to compare our results with other groups' methods using the same test set of queries and relevance judgments. Our word embedding-based approach has shown superior performance in retrieving relevant documents, achieving higher R-precision and MAP scores compared to traditional TF-IDF methods. This comparative analysis will provide insights into the strengths and weaknesses of different retrieval techniques, contributing to advancements in information retrieval systems.

In our case, we are going to compare our results with Group 8.

Results from Group 8

Category	P@5	P@10	Mean AVG precision	R-Precision
Business	0.80	0.70	0.28	0.35
Tech	0.80	0.70	0.24	0.20
Sport	0.0	0.20	0.31	0.32

Entertaiment	0.40	0.40	0.21	0.20
Politics	0.40	0.60	0.27	0.27

Our results:

Category	**P@5**	**P@10**	**Mean AVG precision**	**R-Precision**
Business	0.60	0.70	0.50	0.50
Tech	1.00	0.90	0.51	0.51
Sport	0.20	0.50	0.49	0.54
Entertaiment	0.60	0.30	0.33	0.39
Politics	0.80	0.90	0.53	0.54

In comparing our retrieval system's performance with that of Group 8 across various categories, we observe distinct patterns that highlight the strengths and areas for improvement in both approaches. Overall, our method using embeddings demonstrates competitive performance in terms of precision, mean average precision (MAP), and R-precision across multiple categories.

Across different categories like Business and Politics, our approach consistently achieves higher precision scores at both P@5 and P@10 compared to Group 8. This indicates that our system retrieves a higher proportion of relevant documents within the top results for these categories. Additionally, our Mean Average Precision (MAP) values are generally higher, reflecting more consistent and effective ranking of relevant documents across queries.

In the Tech category, while our precision metrics closely match those of Group 8, our MAP and R-precision values are notably superior. This suggests that our system not only retrieves relevant documents but also ranks them more accurately and comprehensively.

However, there are areas where our system lags behind, particularly in the Sport and Entertainment categories where precision scores at P@5 are lower compared to Group 8. Despite this, our overall retrieval quality, as measured by MAP and R-precision, tends to be higher, indicating a more robust and reliable retrieval system.

In summary, our method leveraging embeddings demonstrates a competitive advantage over Group 8's approach in several key metrics, especially in precision, MAP, and R-precision. This comparison underscores the effectiveness of our retrieval strategy and highlights areas for further optimization to enhance performance across all categories.

Conclusion

Based on our findings, it's evident that word embeddings outshine both doc embeddings and tf-idf in our profile-based information retrieval system. This superiority likely stems from their ability to capture semantic relationships between words, allowing for more nuanced representations of text. Unlike tf-idf, which relies solely on word frequencies, word

embeddings offer a deeper understanding of word meanings within various contexts, facilitating more accurate matching between user interests and document content.

Moreover, word embeddings demonstrate a remarkable ability to generalize across different documents and topics. This flexibility enables them to perform well even with limited data, making them suitable for a wide range of information retrieval tasks. In contrast, doc embeddings, while effective for individual documents, may lack the same level of generalization, limiting their applicability in diverse contexts.

Another advantage of word embeddings lies in their computational efficiency once pre-trained. Although training word embeddings on large corpora can be resource-intensive initially, the subsequent computation of similarities between documents and user interests is typically faster than with doc embeddings. This efficiency is particularly advantageous in real-time or large-scale retrieval systems where speed is essential.

However, it's crucial to recognize that the performance of word embeddings can be further enhanced through fine-tuning on domain-specific data. By adapting the embeddings to the specific characteristics of our dataset and user interests, we can potentially improve their effectiveness even more. This fine-tuning process allows us to tailor the embeddings to our particular needs, ensuring optimal performance in our information retrieval system.

In conclusion, our comparison highlights the superiority of word embeddings in our profile-based information retrieval system. Their semantic understanding, generalization ability, and computational efficiency make them the preferred choice for accurately matching user interests with relevant documents. While doc embeddings and tf-idf have their merits, word embeddings offer the most robust and effective approach for our specific requirements.