IR EVALUATION ON SOCIAL EVENT DETECTION

Miguel Javier Garrido Aguarón

CHUNG-ANG UNIVERSITY

Index

1. Introduction

In this project we have to collect data, analyze it and evaluate our models of Information Retrieval through various methods in order to see if they are efficient and calculate their precision and their recall.

2. Data collection

2.1 Methods

The data collected for the project is from various online newspapers websites.

For the data collection, various Python scripts were created, using web crawling and web scraping techniques to extract URLs from a specific section of the site. The primary goal of this script is to automate the process of gathering news articles from "The Korean Herald" (https://www.koreaherald.com/) and from “Korea Times” (https://www.koreatimes.co.kr/www2/index.asp).

Since each online newspapers webpage html code is different, two different scripts were created, one for each online newspaper, “The Korea Herald” and ___. In both scripts you just must pass the URL from a section of the online newspaper (for example: Sports) and execute it. It will automatically crawl through each page of that section and scrap all the news related to that section (following the previous example, it would only collect news related to sports). From each news, metadata is collected and saved into a txt file, this data is the title of the news, its text and the publication date.

2.2 Script functions

The script has two main functions:

1. **`information_collector(url, contents_filename)`:** This function extracts the title, content, and publication time of a given news article using web scraping methods.

2. **`article_collector(section_url, save_directory)`:** This function identifies the URLs of articles within a specified section of the website, then calls the `information_collector` function to retrieve and save the content of each article. By automating the collection of news articles, this script streamlines the process of accessing and organizing valuable information from "The Korean Herald" and “Korea Times” for our Information Retrieval systems.

2.3 How does the script work

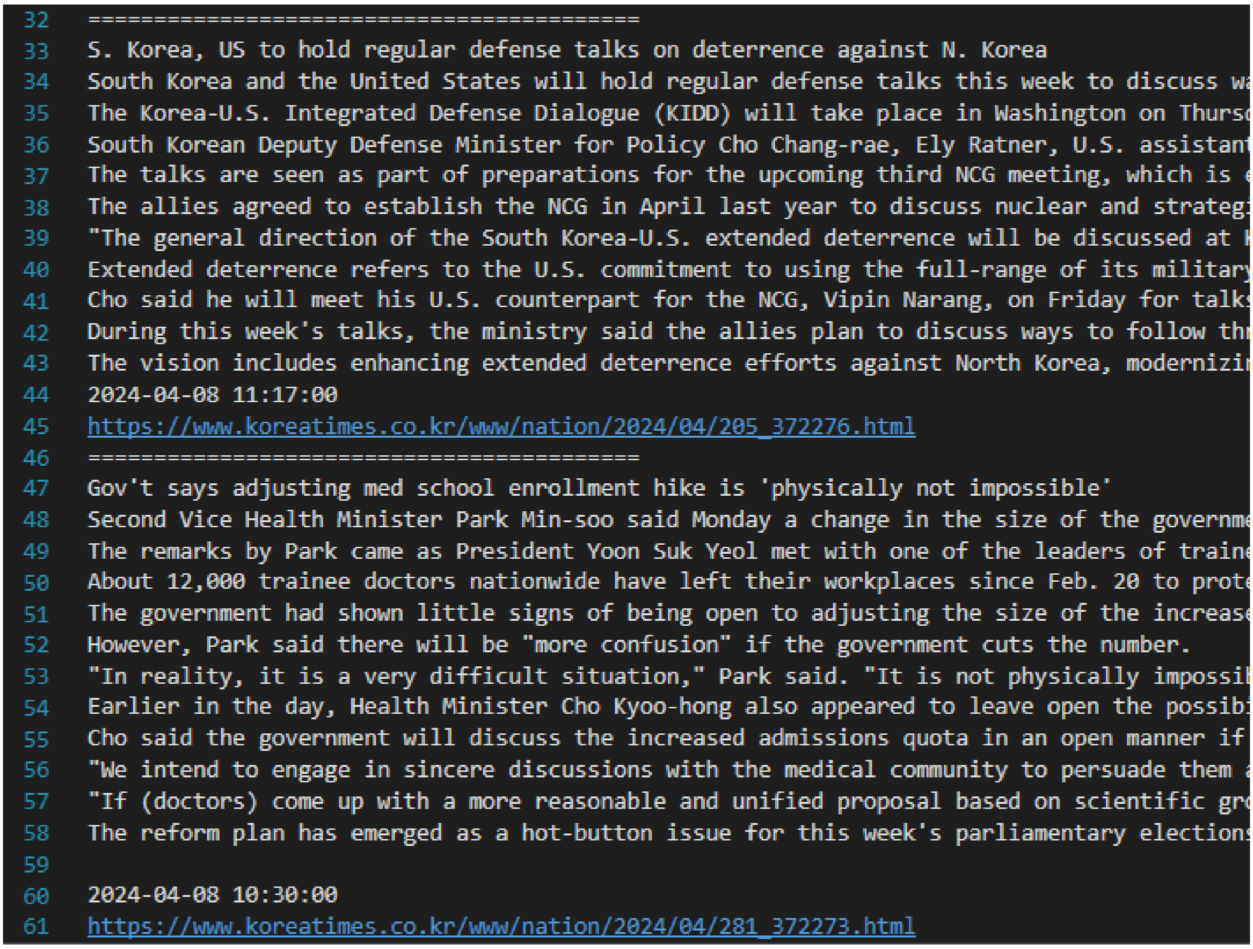
As mentioned previously, it is a scraper and a crawler. What the script does is iterating over every news on a section of the online newspaper (i.e. https://www.koreaherald.com/list.php?ct=020500000000&np=1), for each article in the page, it enters into it’s URL and extracts, thanks to the library BeautifulSoup, whatever elements we want from the html code. In this case, we retrieve the title, text and the date of the article. We also save this information with the article’s url.

After retrieving all the news from a page, it crawls into the second page of news, because not all news are on the same page, I am able to do this thanks to the list-like structure of the webpage, working by index. To give an example, the first page, that has the most actual news could have this link “https://www.koreaherald.com/list.php?ct=020500000000” or this link “https://www.koreaherald.com/list.php?ct=020500000000&np=1”, this means that the next page will have the following link https://www.koreaherald.com/list.php?ct=020500000000&np=2. So for accessing different pages we change the URL’s by https://www.koreaherald.com/list.php?ct=020500000000&np=**X** in the case of “The Korea Herald” webpage.

Thanks to the libraries “NewsPlease” and “Datetime” we can extract the date of every webpage and set the code to extract news until a desired date.

2.4 Outputs

As previously mentioned, two .txt files are created, one contains the scraping output and the other one the crawling URLs. In other words, one of the .txt files will have all the metadata from all the news collected and the other one will have all the URLs from the news collected.



```
==========================================
S. Korea, US to hold regular defense talks on deterrence against N. Korea
South Korea and the United States will hold regular defense talks this week to discuss wa
The Korea-U.S. Integrated Defense Dialogue (KIDD) will take place in Washington on Thursd
South Korean Deputy Defense Minister for Policy Cho Chang-rae, Ely Ratner, U.S. assistant
The talks are seen as part of preparations for the upcoming third NCG meeting, which is
The allies agreed to establish the NCG in April last year to discuss nuclear and strategi
"The general direction of the South Korea-U.S. extended deterrence will be discussed at
Extended deterrence refers to the U.S. commitment to using the full-range of its military
Cho said he will meet his U.S. counterpart for the NCG, Vipin Narang, on Friday for talks
During this week's talks, the ministry said the allies plan to discuss ways to follow thr
The vision includes enhancing extended deterrence efforts against North Korea, modernizin
2024-04-08 11:17:00
https://www.koreatimes.co.kr/www/nation/2024/04/205_372276.html
==========================================
Gov't says adjusting med school enrollment hike is 'physically not impossible'
Second Vice Health Minister Park Min-soo said Monday a change in the size of the governme
The remarks by Park came as President Yoon Suk Yeol met with one of the leaders of traine
About 12,000 trainee doctors nationwide have left their workplaces since Feb. 20 to prote
The government had shown little signs of being open to adjusting the size of the increase
However, Park said there will be "more confusion" if the government cuts the number.
"In reality, it is a very difficult situation," Park said. "It is not physically impossib
Earlier in the day, Health Minister Cho Kyoo-hong also appeared to leave open the possibi
Cho said the government will discuss the increased admissions quota in an open manner if
"We intend to engage in sincere discussions with the medical community to persuade them
"If (doctors) come up with a more reasonable and unified proposal based on scientific gr
The reform plan has emerged as a hot-button issue for this week's parliamentary election

2024-04-08 10:30:00
https://www.koreatimes.co.kr/www/nation/2024/04/281_372273.html
```

Each news article, which we will refer to them as documents, have the same structure: **Title**: The first line

Text: Several lines, depending on the length of the news, it varies amongst them.

Date: One line

URL: Last line of the document

For separating the documents I use a separator, which is basically a string of “==================” for facilitating the implementation of the retrieval systems.

3. Data

As mentioned before, I retrieved all the news from the online newspapers “The Korea Herald” and “Korea Times”, all this news are from the last 10 days as told to do.

In total, I managed to retrieve **950 documents** from both online newspapers combined. For evaluating the precision and recall of the systems, only 950 documents where used to I could know for a specific topic how many documents of that topic there are in total. Later on for time performance analysis, datasets of 2816 documents and 4689 documents will be used.

3.1 The Korea Herald

The Korea Herald, established in 1952, is one of the biggest English-language newspapers in South Korea. It provides information and analysis aimed at an international audience interested in developments within South Korea and also its relations with the world.

All the news from the last 10 days were extracted from all its different sections: national, business, life&culture, sports, world, opinion and k-pop.

3.1.1 The Korea Herald data analysis

A total of **377 documents** were retrieved, each section having a different number of documents, using R-Studio we can see how many documents each section has as seen in the image below.

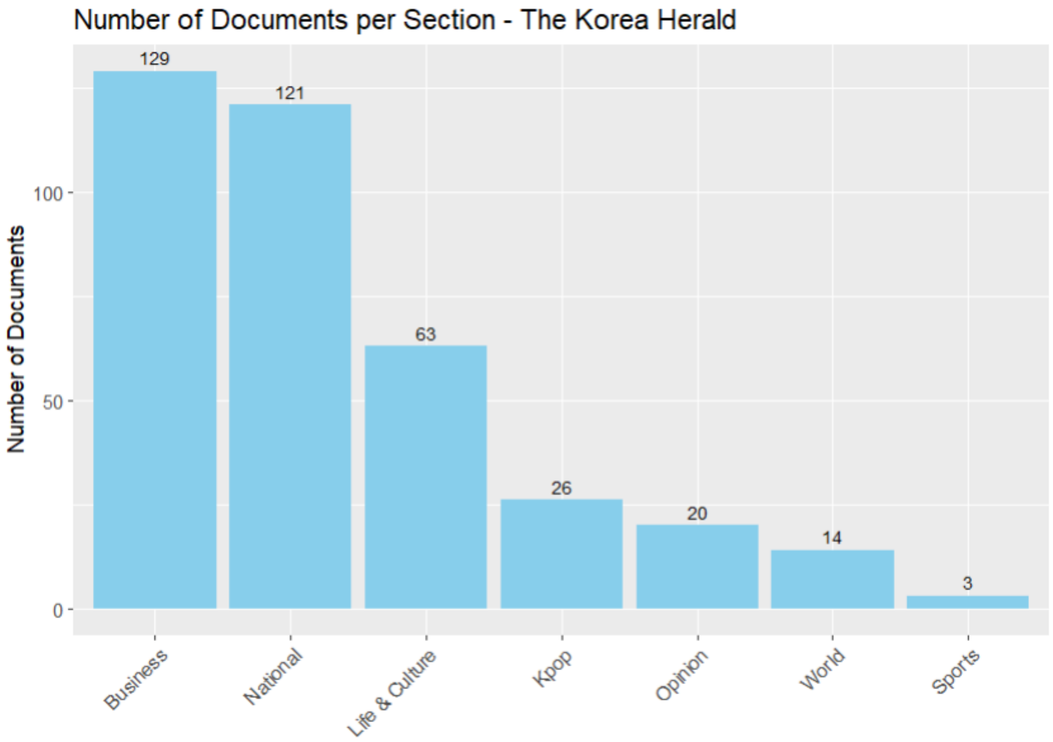


Figure 1 Number of documents per section - "The Korea Herald"

As we can see, The Korea Herald prioritizes news related to business and national news. We can also easily see how it doesn't publish much about sports nor world news. This meant that if we made queries with an Information Retrieval System about topics like sport only having data from The Korea Herald, we would have a very low precision and success rate.

3.2 The Korea Times

Established in 1950 with a long-standing history of reporting Korean issues, The Korea Times is one of the leading English-language newspapers in South Korea. With a large covering of topics such as: North Korea, national topics, business, finance, lifestyle, entertainment&art, sports, world topics and opinion. This newspaper provides insights into Korean affairs and serves as a good source of information for both local and international readers interested in developments of South Korea.

3.2.1 The Korea Times data analysis

With **573 documents** retrieved from news of the last 10 days as told to do, I extracted like in The Korea Herald, all news from all sections that were published in the last 10 days.

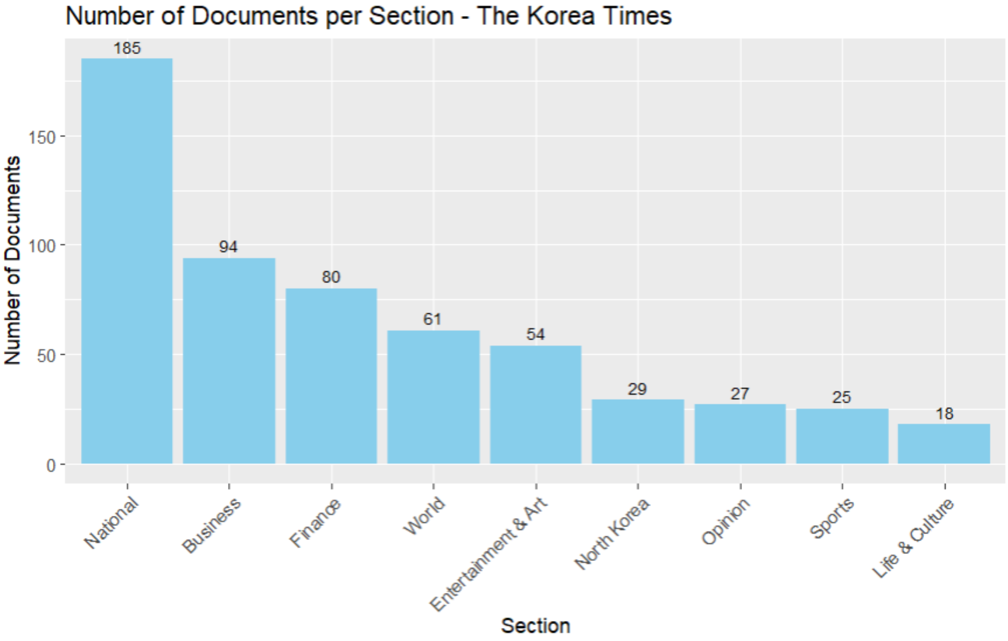


Figure 2 Number of documents per section - "The Korea Times"

With the usage of R-Studio we can plot all news from all sections and analyze them. We can see the sections with the most documents is the *national* section, maybe due to the elections and the medical protest. We also can see that there are not many news about topics like *life&culture* compared to others, which would influence our IR System.

3.3 Data analysis

Combining the documents extracted from both online newspapers we have a total of 950 documents with 10 different sections.

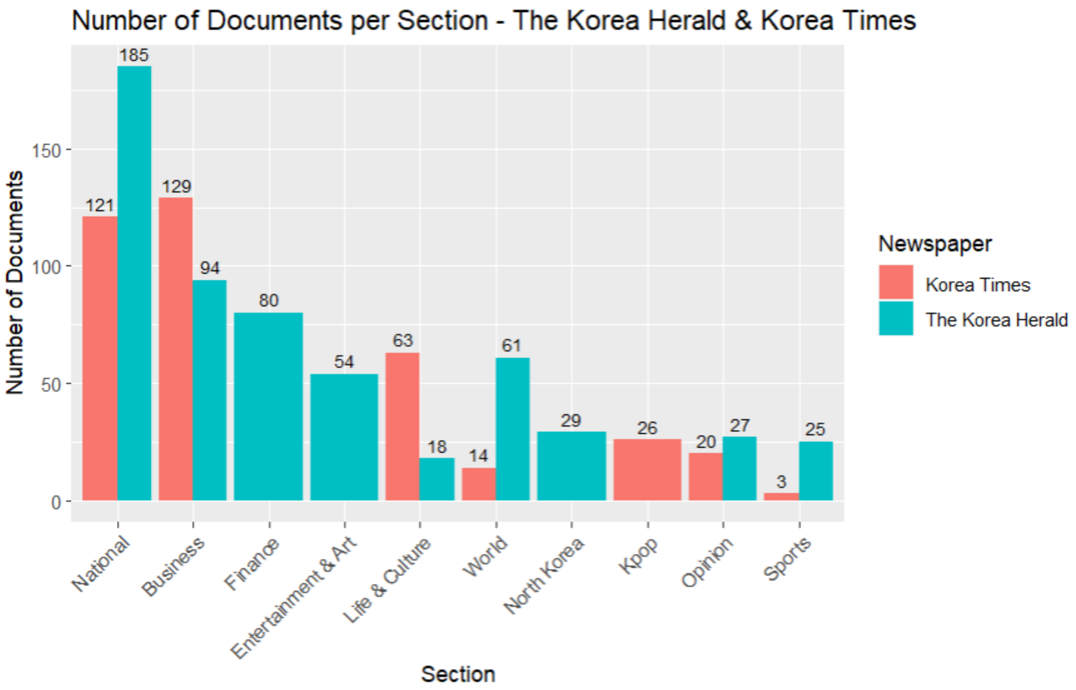


Figure 3 Number of documents per section combined

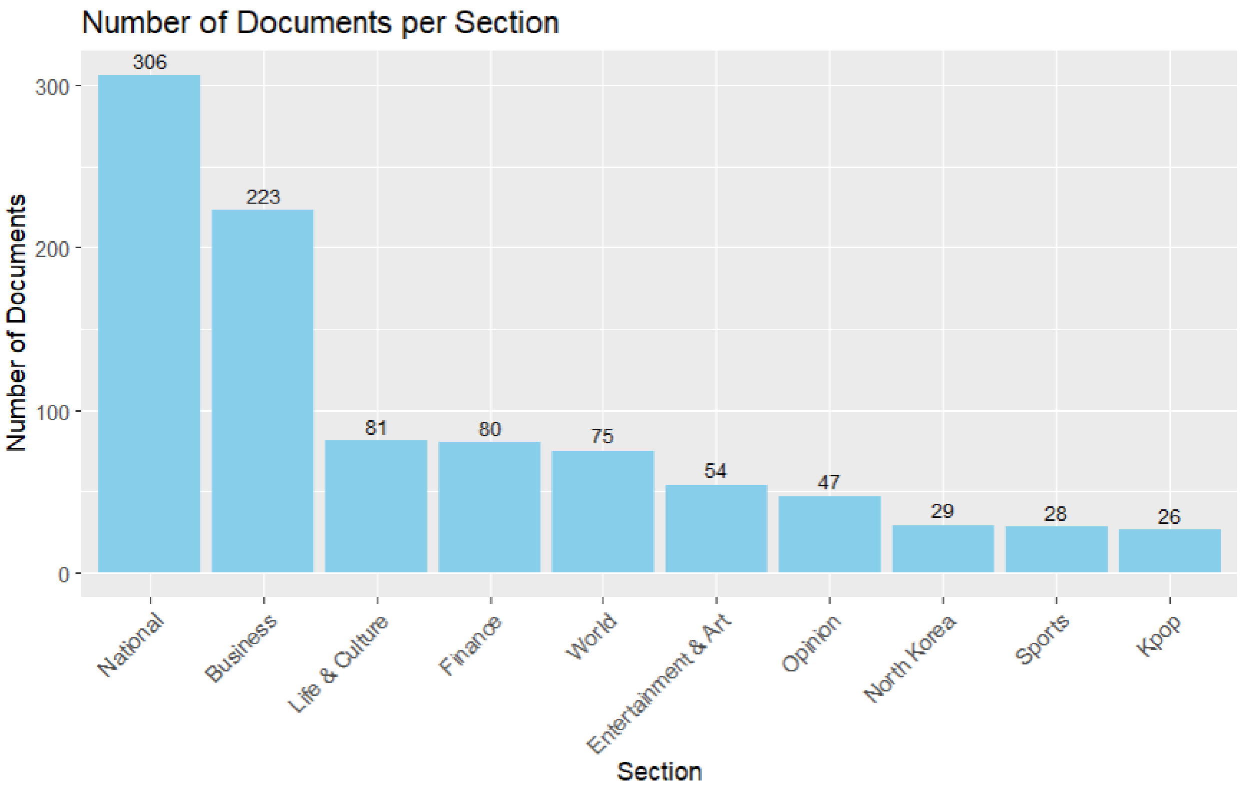


Figure 4 Number of documents per section (total)

Knowing how many documents we have per topic or section will help us understand the results. This is due to Information Retrieval systems using techniques such as cosine similarity can be affected by the distribution of documents across different topics. When there are lots of documents on one topic compared to few documents on another topic, it can impact the performance of the IR System. This is due to several reasons like imbalance in relevance, if our dataset is not balanced the IR system can become biased towards some topics. This will lead to more relevant documents being retrieved from the topic with highest representation and overlooking the actual relevant documents we are looking for.

Also, this can have an impact on term frequency in techniques like cosine similarity. When there are more documents on a certain topic, the terms associated with that topic are likely to appear more frequently on the dataset, this can skew similarity scores towards documents from that specific topic, in conclusion making it harder for documents from less represented topics to compete.

4. Boolean System

4.1 Summary

A Boolean System was implemented to efficiently retrieve documents based on Boolean queries. Boolean queries. These are queries that use Boolean operators such as “AND”, “OR” & “NOT” to combine or exclude terms and retrieve only relevant documents from the database that contain the terms mentioned in the query (or not contained in the case of “NOT”). For doing this, a combination of tokenization, inverted indexing and Boolean logic process had to be made.

4.2 Code components explanation

1. Initialization

Upon instantiation, the class is initialized with two attributes: ‘index’. This will be the fundamental data structure that will be worked on in the system. This is a dictionary that serves as an index and maps tokens to list document IDs where the tokens appear. Each key in the index is a unique token extracted from all the documents. Each value in the other side, corresponds to a list of document IDs that contain where the token appears. For example, if we had *{rain: 3, 5}* would mean that the documents 3 and 5 contain the token "rain".

2. Tokenization

The tokenization process, in the code the 'tokenize' method, is responsible for transforming the text into tokens, tokens in Natural Language Processing are single and meaningful units of text, individual words. Also in the tokenize process all tokens are transformed into lowercase and punctuation signs are removed. This process will return a list of tokens from the input text.

3. Adding documents

The 'add_document' method adds a document to the inverted index. It calls the tokenize function for tokenizing the title and text of the documents, and then iterate over each token to update the index with the corresponding ID.

4. Searching

The 'search' method is used to retrieve documents based on a given query input by the user. It processes the query and evaluates Boolean expressions. Then returns a list of document IDs that match the query.

4.3 Boolean system process

1. Document parsing

As the documents are read from our text database, that consists of a text file, and they are parsed to extract titles, text and URLs. Each of these documents are also added to a list for later processing.

2. Indexing

As every document is being parsed, it is added to the 'InvertedIndex' object using the 'add_document' method. This will build the inverted index, mapping tokens to document IDs.

3. User query

The user is prompted to input a Boolean query. Then it is processed using the 'search' method.

4. Search execution

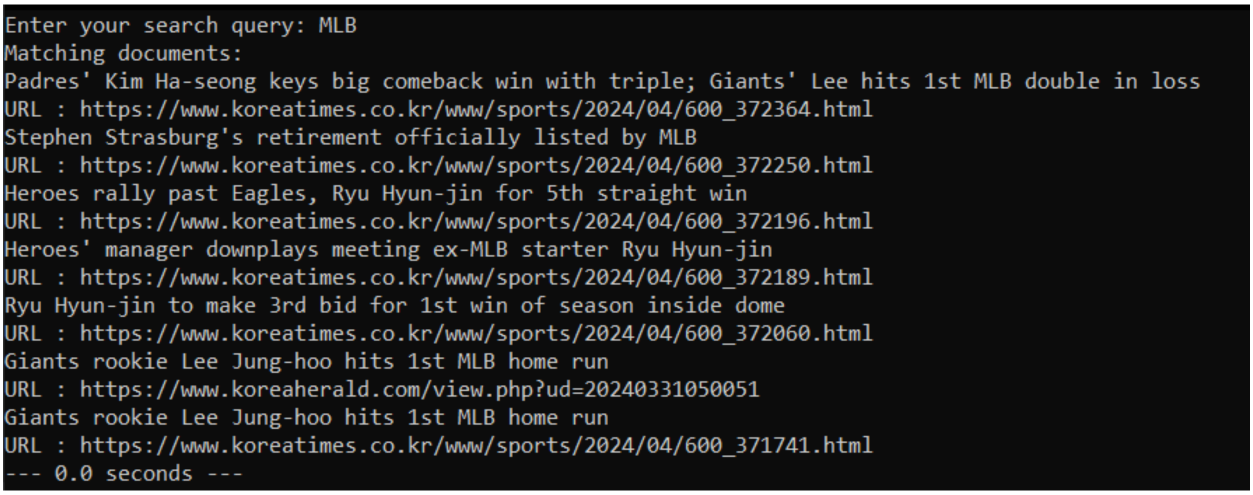
The search query is processed and Boolean operators are identified. Then based on the Boolean logic of the query, relevant documents are retrieved from the inverted index.

5. Result Display

If any matching documents are found, their titles and their corresponding URLs will be displayed to the user. Otherwise, a message will appear indicating no matching documents were found.

4.4 Execution and evaluation

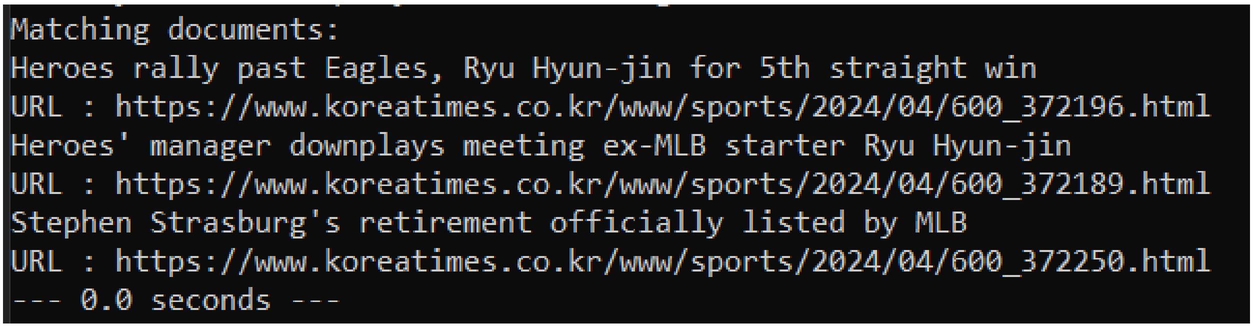
For the Boolean system the execution is really simple, you type the Boolean query and it retrieves the documents that have that Boolean query.



```
Enter your search query: MLB
Matching documents:
Padres' Kim Ha-seong keys big comeback win with triple; Giants' Lee hits 1st MLB double in loss
URL : https://www.koreatimes.co.kr/www/sports/2024/04/600_372364.html
Stephen Strasburg's retirement officially listed by MLB
URL : https://www.koreatimes.co.kr/www/sports/2024/04/600_372250.html
Heroes rally past Eagles, Ryu Hyun-jin for 5th straight win
URL : https://www.koreatimes.co.kr/www/sports/2024/04/600_372196.html
Heroes' manager downplays meeting ex-MLB starter Ryu Hyun-jin
URL : https://www.koreatimes.co.kr/www/sports/2024/04/600_372189.html
Ryu Hyun-jin to make 3rd bid for 1st win of season inside dome
URL : https://www.koreatimes.co.kr/www/sports/2024/04/600_372060.html
Giants rookie Lee Jung-hoo hits 1st MLB home run
URL : https://www.koreaherald.com/view.php?ud=20240331050051
Giants rookie Lee Jung-hoo hits 1st MLB home run
URL : https://www.koreatimes.co.kr/www/sports/2024/04/600_371741.html
--- 0.0 seconds ---
```

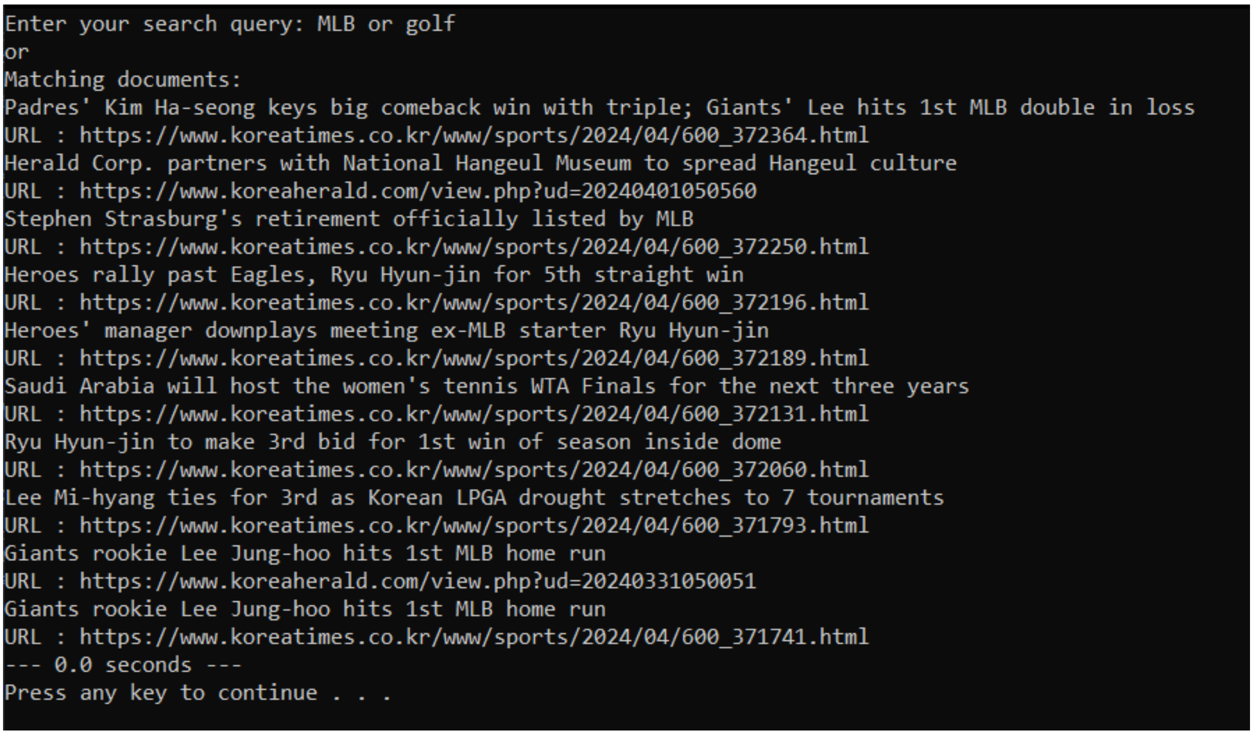
Searching “MLB”:

Searching “MLB AND NOT giants”



```
Matching documents:
Heroes rally past Eagles, Ryu Hyun-jin for 5th straight win
URL : https://www.koreatimes.co.kr/www/sports/2024/04/600_372196.html
Heroes' manager downplays meeting ex-MLB starter Ryu Hyun-jin
URL : https://www.koreatimes.co.kr/www/sports/2024/04/600_372189.html
Stephen Strasburg's retirement officially listed by MLB
URL : https://www.koreatimes.co.kr/www/sports/2024/04/600_372250.html
--- 0.0 seconds ---
```

Searching for “MLB OR golf”



```
Enter your search query: MLB or golf
or
Matching documents:
Padres' Kim Ha-seong keys big comeback win with triple; Giants' Lee hits 1st MLB double in loss
URL : https://www.koreatimes.co.kr/www/sports/2024/04/600_372364.html
Herald Corp. partners with National Hangeul Museum to spread Hangeul culture
URL : https://www.koreaherald.com/view.php?ud=20240401050560
Stephen Strasburg's retirement officially listed by MLB
URL : https://www.koreatimes.co.kr/www/sports/2024/04/600_372250.html
Heroes rally past Eagles, Ryu Hyun-jin for 5th straight win
URL : https://www.koreatimes.co.kr/www/sports/2024/04/600_372196.html
Heroes' manager downplays meeting ex-MLB starter Ryu Hyun-jin
URL : https://www.koreatimes.co.kr/www/sports/2024/04/600_372189.html
Saudi Arabia will host the women's tennis WTA Finals for the next three years
URL : https://www.koreatimes.co.kr/www/sports/2024/04/600_372131.html
Ryu Hyun-jin to make 3rd bid for 1st win of season inside dome
URL : https://www.koreatimes.co.kr/www/sports/2024/04/600_372060.html
Lee Mi-hyang ties for 3rd as Korean LPGA drought stretches to 7 tournaments
URL : https://www.koreatimes.co.kr/www/sports/2024/04/600_371793.html
Giants rookie Lee Jung-hoo hits 1st MLB home run
URL : https://www.koreaherald.com/view.php?ud=20240331050051
Giants rookie Lee Jung-hoo hits 1st MLB home run
URL : https://www.koreatimes.co.kr/www/sports/2024/04/600_371741.html
--- 0.0 seconds ---
Press any key to continue . . .
```

As seen in this examples works fine and you can retrieve all the documents that contain (or not) the term or terms you want, all with a Boolean query.

5. Vectorial Information Retrieval System

5.1 Summary

This information retrieval system aims to rank documents based on their relevance to a given query input by the user. In this case, the query of the user doesn't need to be a Boolean query, it is a free query where the user can write the query however he wants. For the ranking system, the TF-IDF (Term Frequency-Inverse Document Frequency) weighting is used.

5.2 Code components explanation

1. Initialization

The RankedRetrieval class initializes several important attributes. The doc_frequency attribute is a dictionary that keeps track of document frequencies. Weights is a dictionary and idf a list that are used to store term weights and inverse document frequencies, respectively. Finally, the max_doc_id attribute holds the maximum document ID.

2. Tokenization

Same process than the one used in the Boolean system

3. Frequency calculation

Term frequency is calculated and document frequency is updated. For all tokens in a document we iterate over them and counting how many times they appear. Also frequencies are normalized by dividing them between the largest frequency of all of them. When calculated, "doc_frequency" is updated, where the key is the term and the value is a list, that has the same number of elements as documents there are and inside this list there is a tuple of (doc_id, frequency).

4. Inverse Document Frequency Calculation

The get_idf method calculates the inverse document frequency (IDF) for each term in a document using the document frequencies stored in doc_frequency. We do this because terms that appear in many different documents are less indicative of overall topic. For calculating the *idf* for a term we calculate the logarithm base 2 of the division of the total number of documents divided by the document frequency of that term. With this we can calculate the weight of that term by multiplying its idf with its frequency and store it in the dictionary "weights". In this dictionary each key is a term and the values are tuples of (doc_id, weight).

5. Query processing

We do the same we did for the documents but for the user's query. We tokenize it and create a TF-IDF weighted vector based on the tokens of the query.

6. Computing cosine similarity

For computing the cosine similarity we have to iterate over every token in the query vector and retrieve the TF-IDF weights for the term in each document that it appears.For each document containing the current term, the method calcaulates the dot product of the TD-IDF weighted vector for the query and the TF-IDF weighted vector for the document, the magnitude (Euclidian norm) of the TF-IDF weighted vector is also calculated for each document. Using the dot product and magnitudes, the cosine similarity between the query vector and each document vector is calculated.

The similarity scores are normalized by dividing the dot product by the product of the magnitudes of the query vector and the document vector.

If the magnitude of the document vector is zero (indicating that the document does not contain any of the query terms), the similarity score for that document is set to zero.

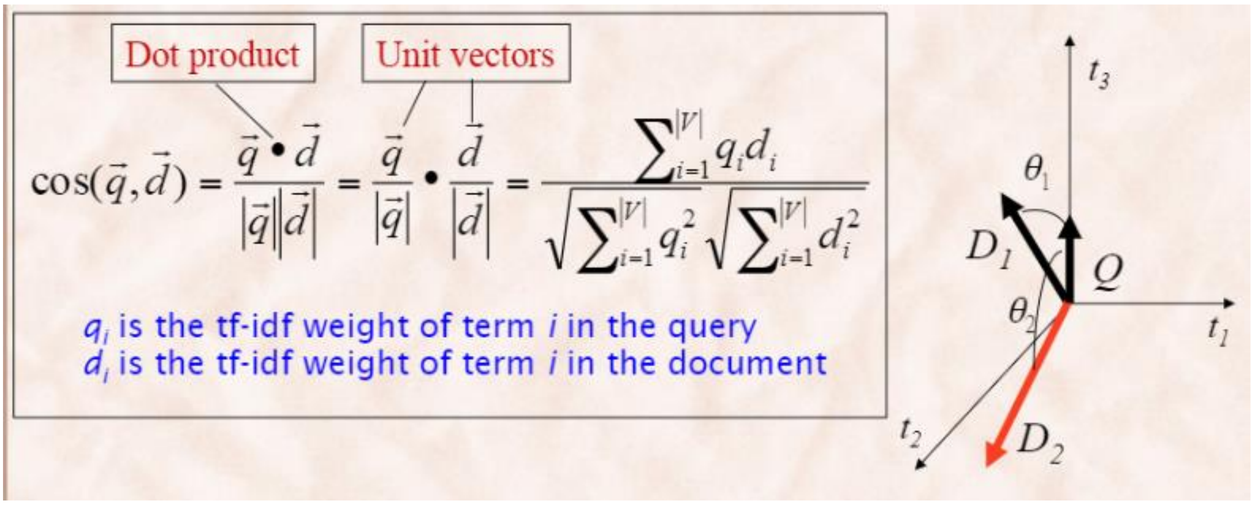
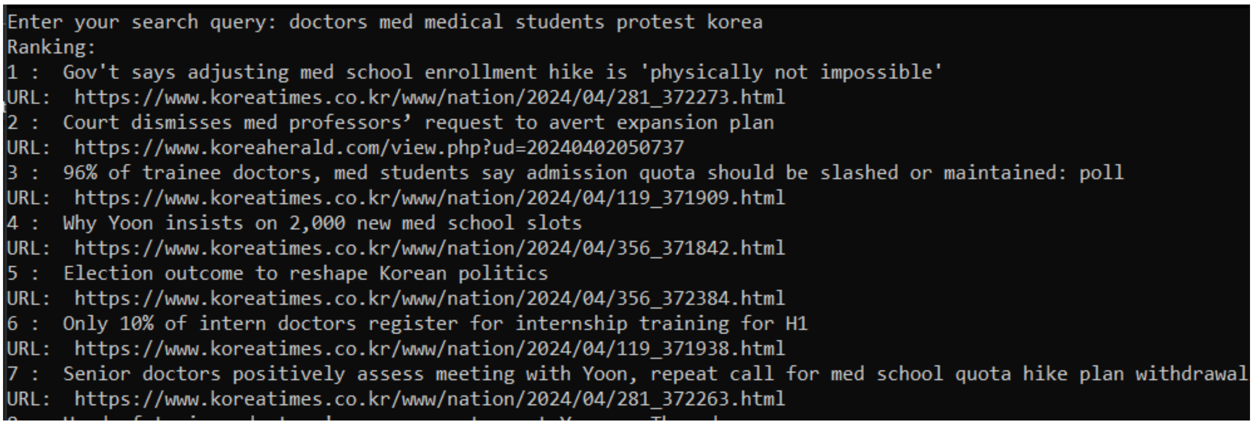

Dot product
Unit vectors
$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\bullet\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\bullet\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$
q_i is the tf-idf weight of term i in the query
d_i is the tf-idf weight of term i in the document
t_3
θ_1
D_1
Q
θ_2
t_1
t_2
D_2

Figure 5 Cosine similarity equation

5.3 Execution and evaluation

5.3.1 Execution I

For the first execution, I wanted to retrieve documents that had information about the situation with doctors in Korea. Recently, more than 10,000 South Korean junior doctors have resigned as a protest against the government’s new plan to recruit more medical students. This has been very important and there have been lots of news about it, this means that in my dataset I have plenty of documents about it.



```
Enter your search query: doctors med medical students protest korea
Ranking:
1 :  Gov't says adjusting med school enrollment hike is 'physically not impossible'
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/281_372273.html
2 :  Court dismisses med professors’ request to avert expansion plan
URL:  https://www.koreaherald.com/view.php?ud=20240402050737
3 :  96% of trainee doctors, med students say admission quota should be slashed or maintained: poll
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/119_371909.html
4 :  Why Yoon insists on 2,000 new med school slots
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/356_371842.html
5 :  Election outcome to reshape Korean politics
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/356_372384.html
6 :  Only 10% of intern doctors register for internship training for H1
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/119_371938.html
7 :  Senior doctors positively assess meeting with Yoon, repeat call for med school quota hike plan withdrawal
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/281_372263.html
```

Out of the total 950 documents, 57 where about this topic and I wanted to retrieve them and see the precision and recall at every *k* documents. My Information Retrieval system ranks all the documents which cosine similarity with the query is greater than 0, in this case 670 documents.

The query used for retrieving this information was “*doctors med medical students protest korea*”.

Until retrieving all the documents that had to do with the doctors situation, whether explaining it or consequences of it, 75 documents had been retrieved having 18 false negatives.

Precision

Using R-Studio I plotted the precision curve and the recall curve for the *k=75* positions until all documents were retrieved.

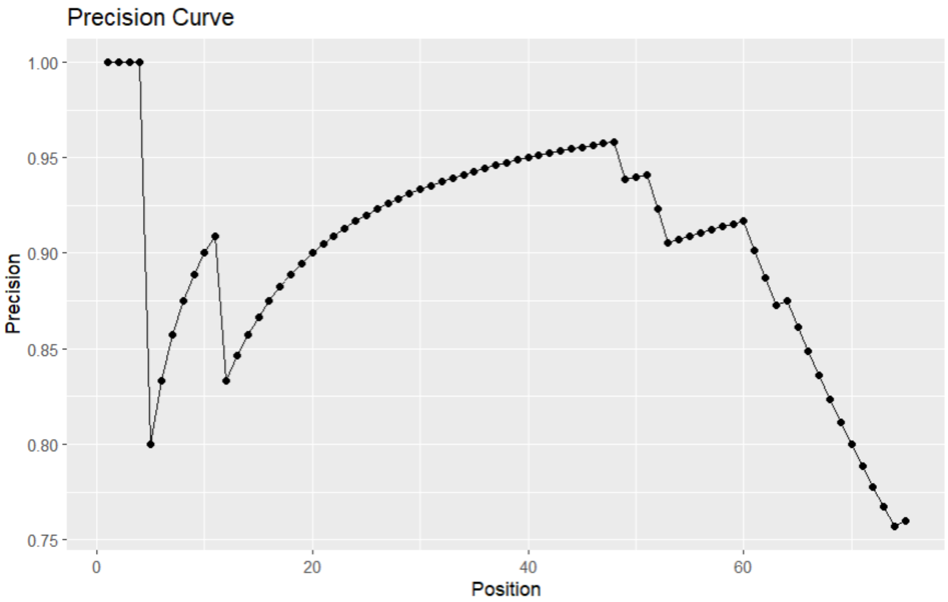


Figure 6 Precision curve for each k position

The result talking about precision is very good and very stable when the k is higher. It may not look like it at first sight, but the Y axis varies only from 76% to 100% precision, being those two precision values the lowest and highest I have while retrieving results.

The best precision I get with a decent K value if of **95.83%** at **K = 48**. This means that after retrieving 48 out of the 57 documents I had 46 true positives and 2 false negatives.

After retrieving all the documents, I had a precision of 76% due to an outlier that my code didn't retrieve earlier.

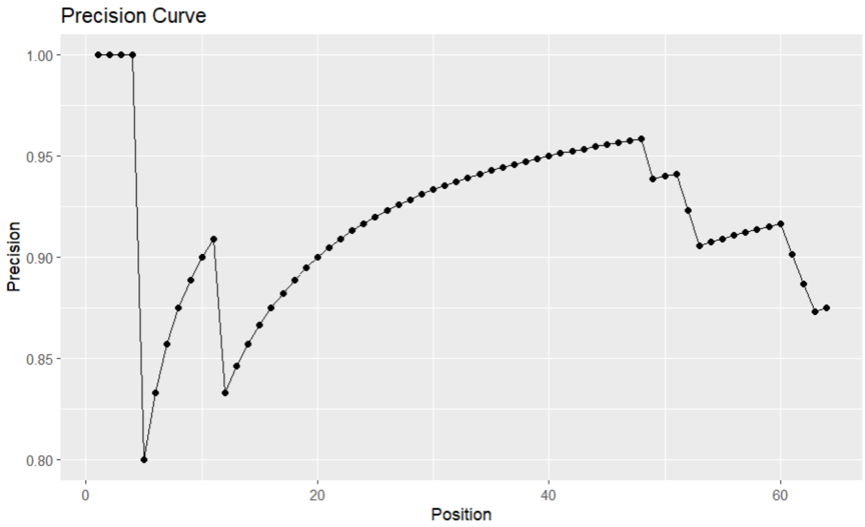


Figure 7 Precision curve without outlier

If we remove the outlier the resulting precision curve would be like shown below:

Recall

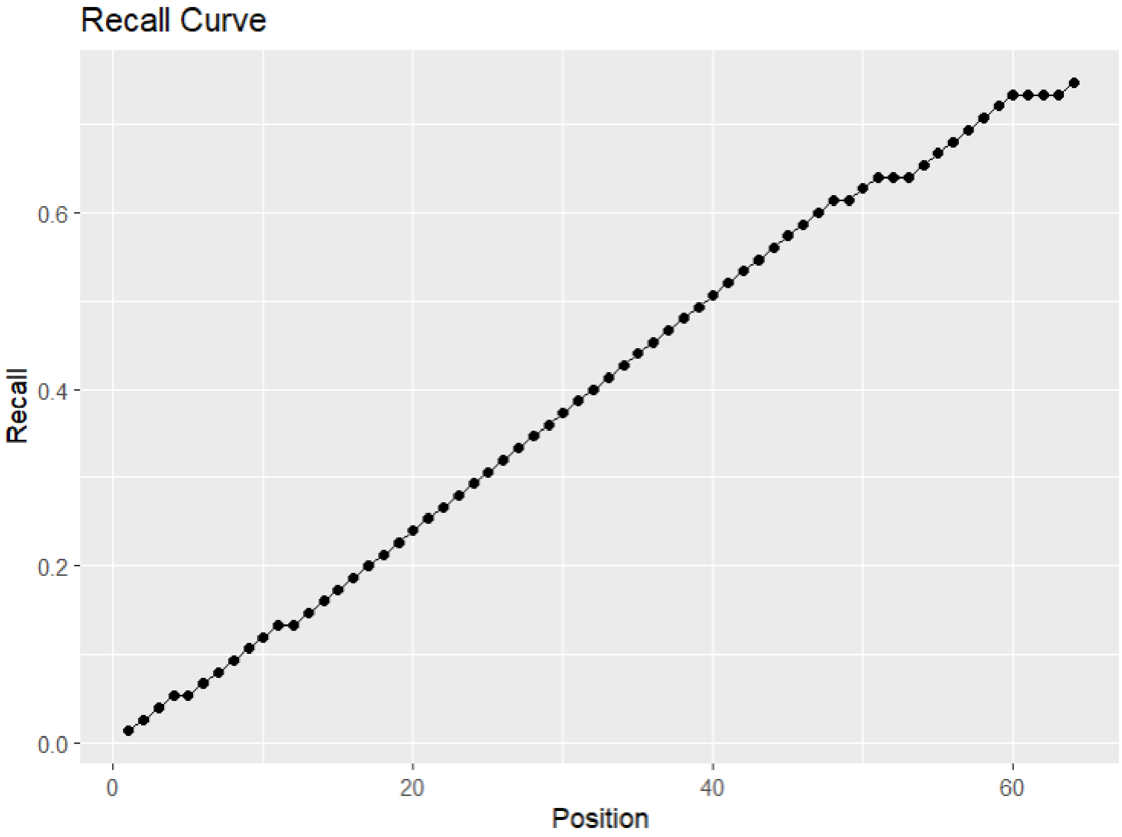


Figure 8 Recall curve per K position

We can observe that the recall curve augments in a linear way. This means that the model performance is proportional to the number of samples retrieved. As the model retrieves more samples, it also captures more true positives proportionally. This means that my model ability to detect true positives is consistent across different retrieval scenarios.

Precision-Recall curve

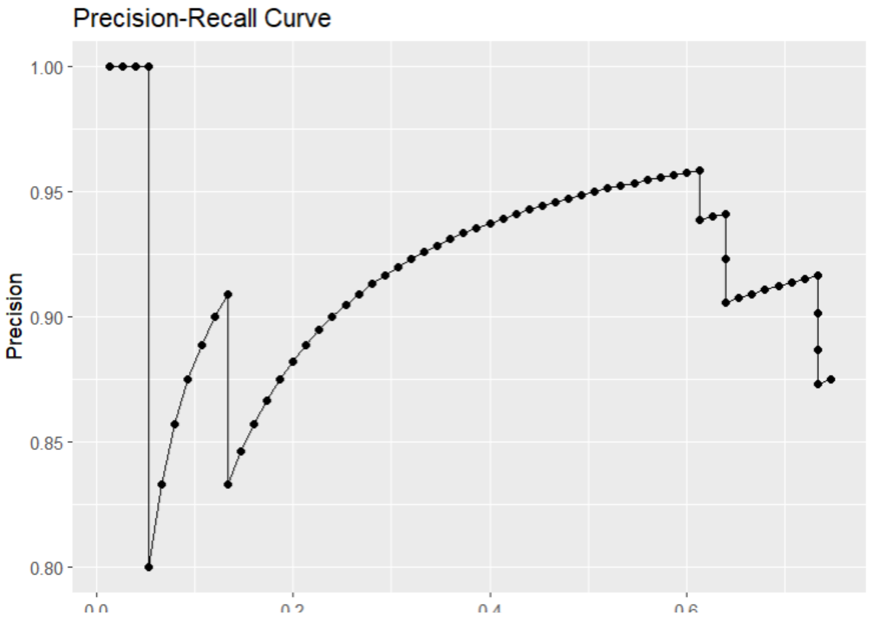
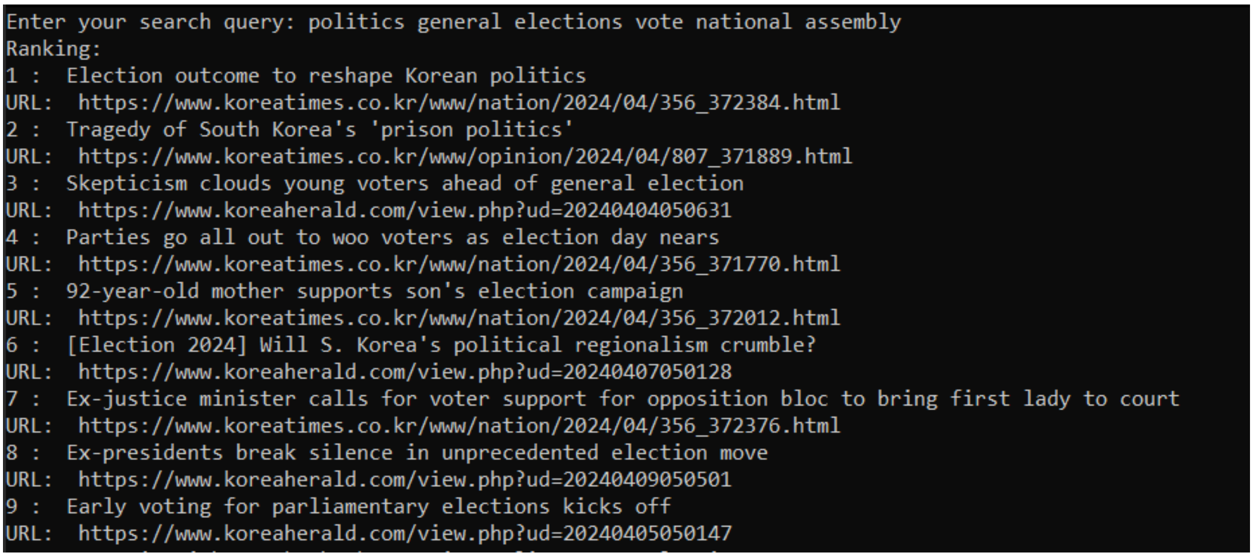


Figure 9 Precision-recall curve

We can observe that the precision-recall curve is really similar to the precision curve. This means that the dataset was balanced, in other words, in this case had lots of documents about the specified query.

5.3.2 Execution II

For the second execution I wanted to retrieve the documents related to the elections that will take place in Korea the 10th of April. Since this is also a very important topic, my dataset has lots of documents about it, 49 in total.



```
Enter your search query: politics general elections vote national assembly
Ranking:
1 :  Election outcome to reshape Korean politics
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/356_372384.html
2 :  Tragedy of South Korea's 'prison politics'
URL:  https://www.koreatimes.co.kr/www/opinion/2024/04/807_371889.html
3 :  Skepticism clouds young voters ahead of general election
URL:  https://www.koreaherald.com/view.php?ud=20240404050631
4 :  Parties go all out to woo voters as election day nears
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/356_371770.html
5 :  92-year-old mother supports son's election campaign
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/356_372012.html
6 :  [Election 2024] Will S. Korea's political regionalism crumble?
URL:  https://www.koreaherald.com/view.php?ud=20240407050128
7 :  Ex-justice minister calls for voter support for opposition bloc to bring first lady to court
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/356_372376.html
8 :  Ex-presidents break silence in unprecedented election move
URL:  https://www.koreaherald.com/view.php?ud=20240409050501
9 :  Early voting for parliamentary elections kicks off
URL:  https://www.koreaherald.com/view.php?ud=20240405050147
```

When typing the query, 364 documents were retrieved, but like before, only the top K documents are relevant and we want to try and find the optimal K. The query I used for retrieving the documents is "*politics general elections vote national assembly*".

When all the documents about the topic were retrieved, the 49 of them, 66 topics were retrieved having only 17 false positives, which is a high precision

Precision

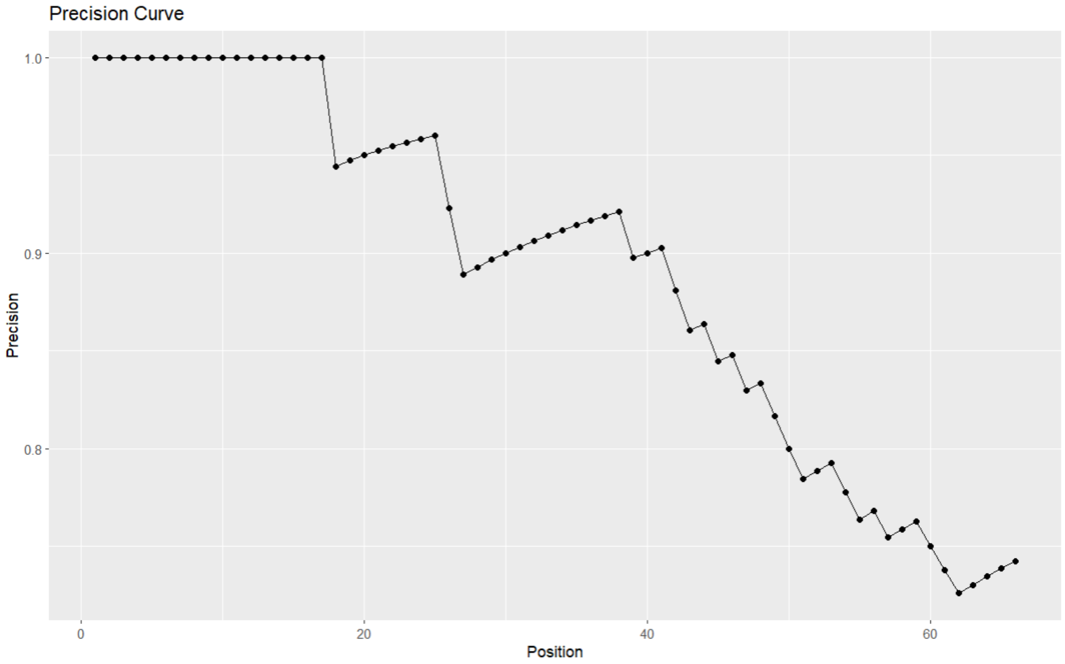


Figure 10 Precision curve

As before and as expected, the precision curve starts high and starts loosing precision when k gets bigger. This is very logic because the more retrieved documents about the topic, the less documents about the topic there are in the rest of them.

In despite of that a good precision was accomplished, starting with 100% precision for the first 17 retrieved documents and going down gradually until ending up with 74.24% precision after having retrieved all the 49 documents about the elections topic.

Recall

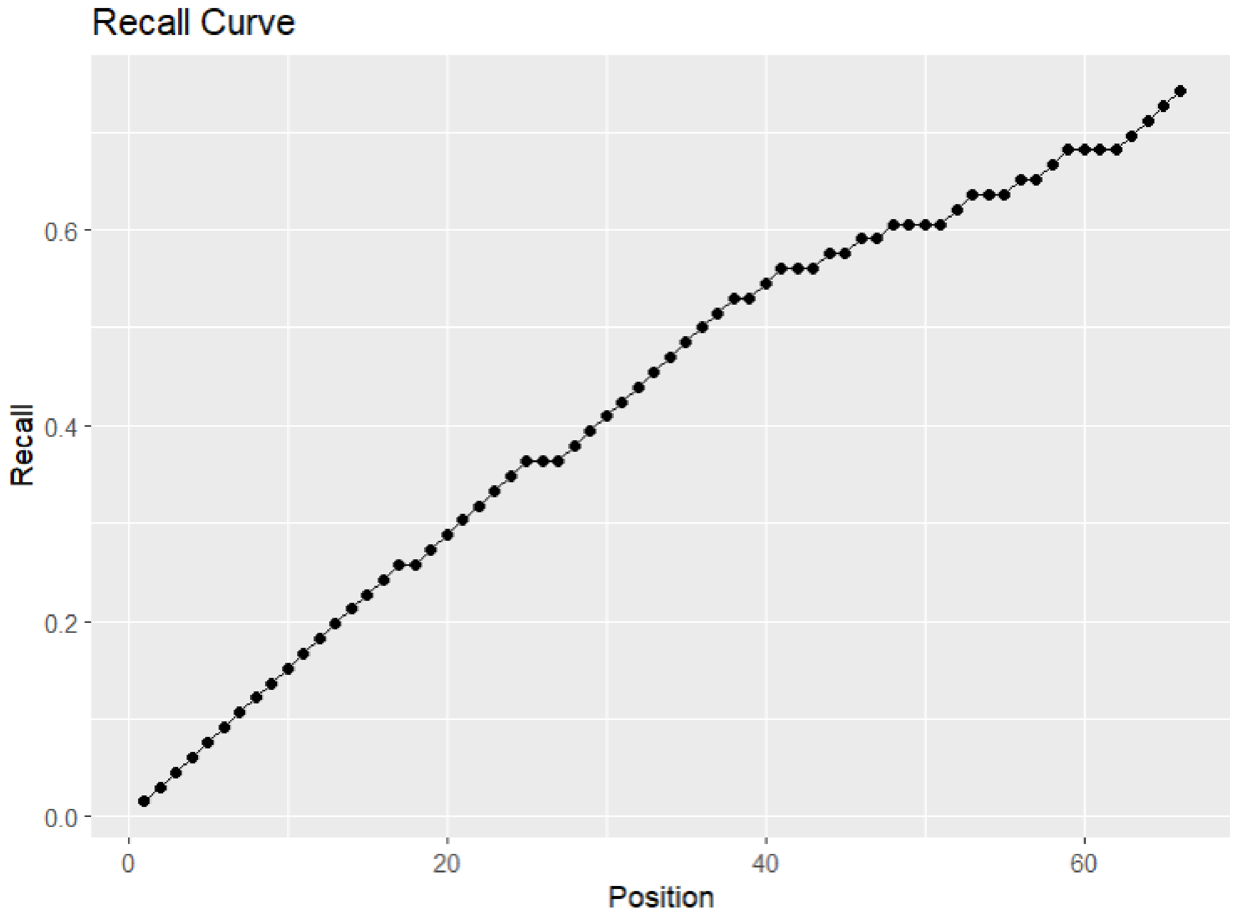


Figure 11 Recall curve

Same as before, we have a consistent linear increment, as more documents are retrieved the recall increases. Follows the same logic as *Execution 1*.

Precision-Recall curve

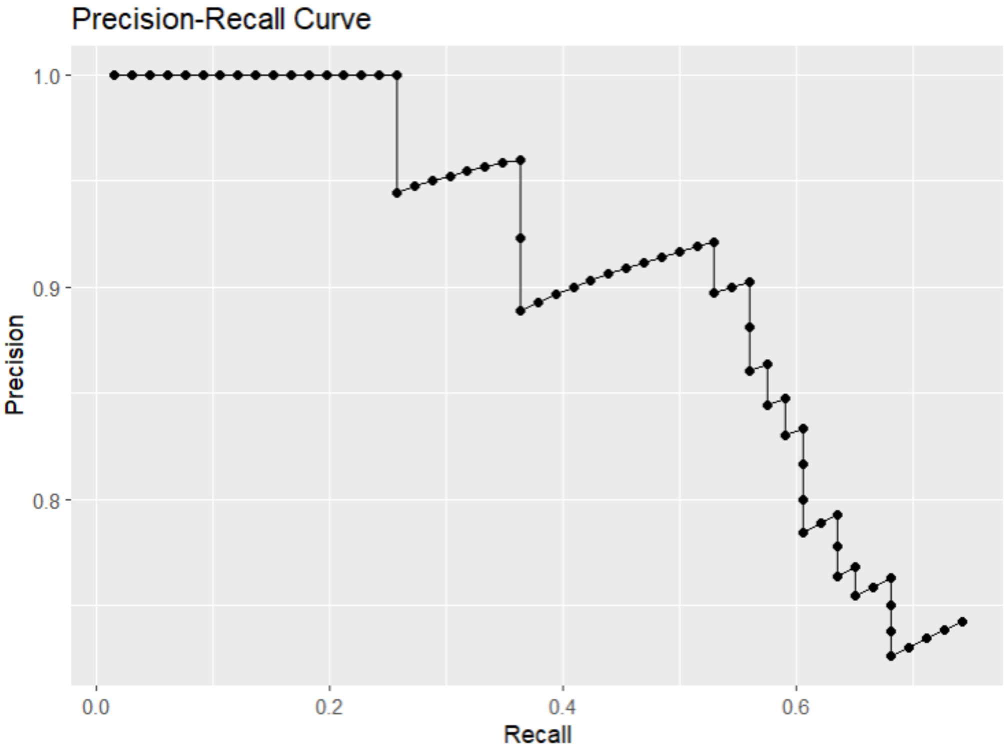


Figure 12 Precision-Recall curve

Same situation like *Execution 1*, we have a very similar precision-recall curve to the precision curve. This is because the dataset is balanced which makes sense, because we have a good amount of documents about the topic we searched for.

5.4 Conclusion

The Information Retrieval System created using cosine similarity works very well when working with balanced datasets. All outputs are as expected with precision decreasing at the same time that recall increases with *K* documents retrieved.

The precisions accomplished are very good, having at a decent number of *k* documents in *execution 1* the precision was about **95%** when having retrieved 48 documents. For execution 2 the precision is also really good with **100%** at 17 documents retrieved or **92.307%** at K = 26N (26 documents retrieved).

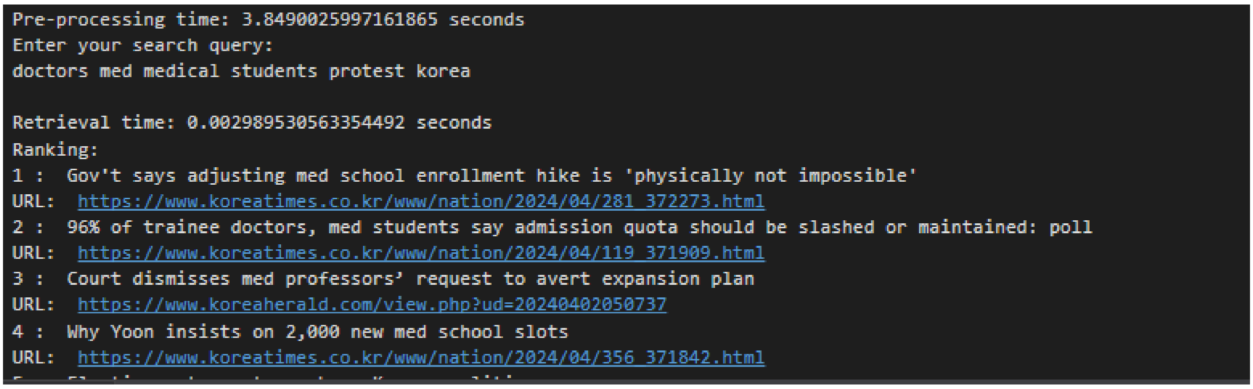
6. Time performance

A characteristic that makes a good Information Retrieval System is its time performance and time complexity. The faster we retrieve good results, the happier the user is. For making this evaluation we are going to look at two things, first pre-processing time, how much time do we need for calculating all the weights of the documents. Then we will calculate the retrieval time; how much time passes since the user makes a query until the results are obtained (we neglect printing time).

6.1 Getting time performance from the query in "Execution 1"

We are going to evaluate time performance based on the first query for retrieving documents about the medical situation in Korea, *Execution 1* point 5.3.1, using the same query.

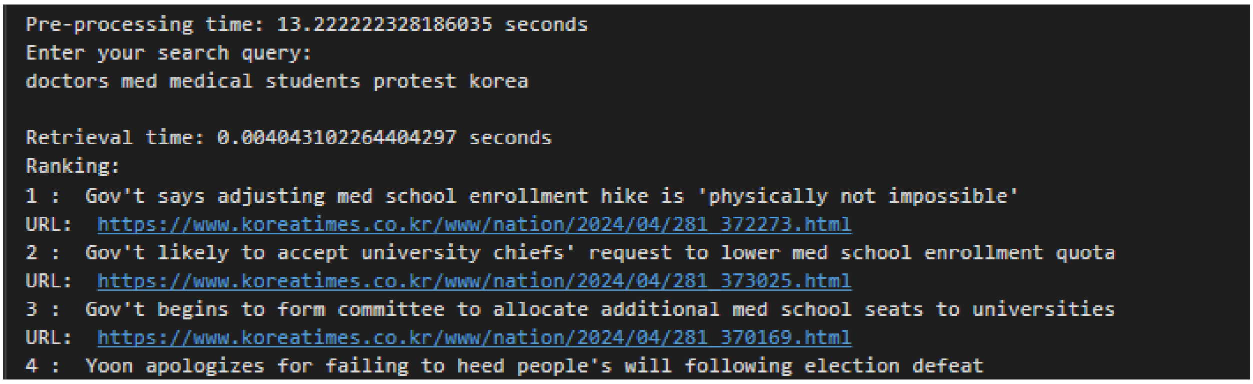
First we are going to evaluate the time for the dataset we used for evaluating our systems, consisting in 950 documents.



```
Pre-processing time: 3.8490025997161865 seconds
Enter your search query:
doctors med medical students protest korea

Retrieval time: 0.002989530563354492 seconds
Ranking:
1 :  Gov't says adjusting med school enrollment hike is 'physically not impossible'
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/281_372273.html
2 :  96% of trainee doctors, med students say admission quota should be slashed or maintained: poll
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/119_371909.html
3 :  Court dismisses med professors' request to avert expansion plan
URL:  https://www.koreaherald.com/view.php?ud=20240402050737
4 :  Why Yoon insists on 2,000 new med school slots
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/356_371842.html
```

We have a pre-preoccessing time of 3.8490025997161865 seconds that would give us 246 documents per second and a retrieval time of 0.002989530563354492 seconds.

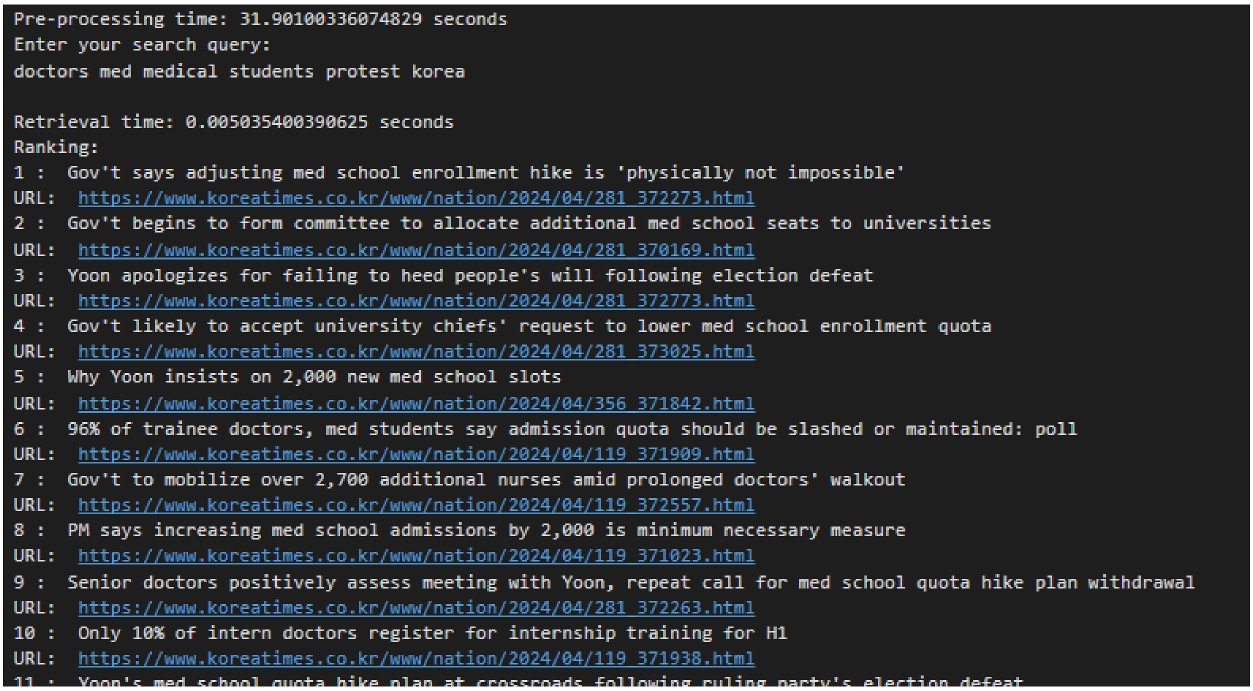


```
Pre-processing time: 13.222222328186035 seconds
Enter your search query:
doctors med medical students protest korea

Retrieval time: 0.004043102264404297 seconds
Ranking:
1 :  Gov't says adjusting med school enrollment hike is 'physically not impossible'
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/281_372273.html
2 :  Gov't likely to accept university chiefs' request to lower med school enrollment quota
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/281_373025.html
3 :  Gov't begins to form committee to allocate additional med school seats to universities
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/281_370169.html
4 :  Yoon apologizes for failing to heed people's will following election defeat
```

Secondly I tried with 1907 documents. This gives us a pre-processing time of 13.222222328186035 seconds adn a retrieval time of 0.004043102264404297 seconds. Pre-processing 144 documents per second.

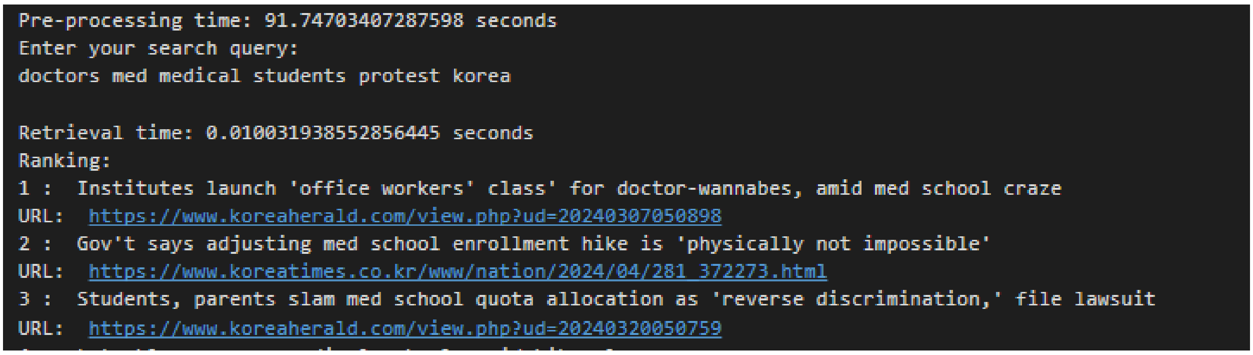
Then I evaluated time performance having a database consisting on 2816 documents, this documents are news from “The Korea Times” that range from the 19th of April 2024 until the 28th of February.



```
Pre-processing time: 31.90100336074829 seconds
Enter your search query:
doctors med medical students protest korea

Retrieval time: 0.005035400390625 seconds
Ranking:
1 :  Gov't says adjusting med school enrollment hike is 'physically not impossible'
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/281_372273.html
2 :  Gov't begins to form committee to allocate additional med school seats to universities
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/281_370169.html
3 :  Yoon apologizes for failing to heed people's will following election defeat
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/281_372773.html
4 :  Gov't likely to accept university chiefs' request to lower med school enrollment quota
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/281_373025.html
5 :  Why Yoon insists on 2,000 new med school slots
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/356_371842.html
6 :  96% of trainee doctors, med students say admission quota should be slashed or maintained: poll
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/119_371909.html
7 :  Gov't to mobilize over 2,700 additional nurses amid prolonged doctors' walkout
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/119_372557.html
8 :  PM says increasing med school admissions by 2,000 is minimum necessary measure
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/119_371023.html
9 :  Senior doctors positively assess meeting with Yoon, repeat call for med school quota hike plan withdrawal
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/281_372263.html
10 :  Only 10% of intern doctors register for internship training for H1
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/119_371938.html
11 :  Yoon's med school quota hike plan at crossroads following ruling party's election defeat
```

As before, until K doesn’t get high the precision is really accurate. Looking to pre-processing time we have 31.90100336074829, which would be 88 documents per seocond. As of retrieval time, we have only 0.005035400390625 seconds.



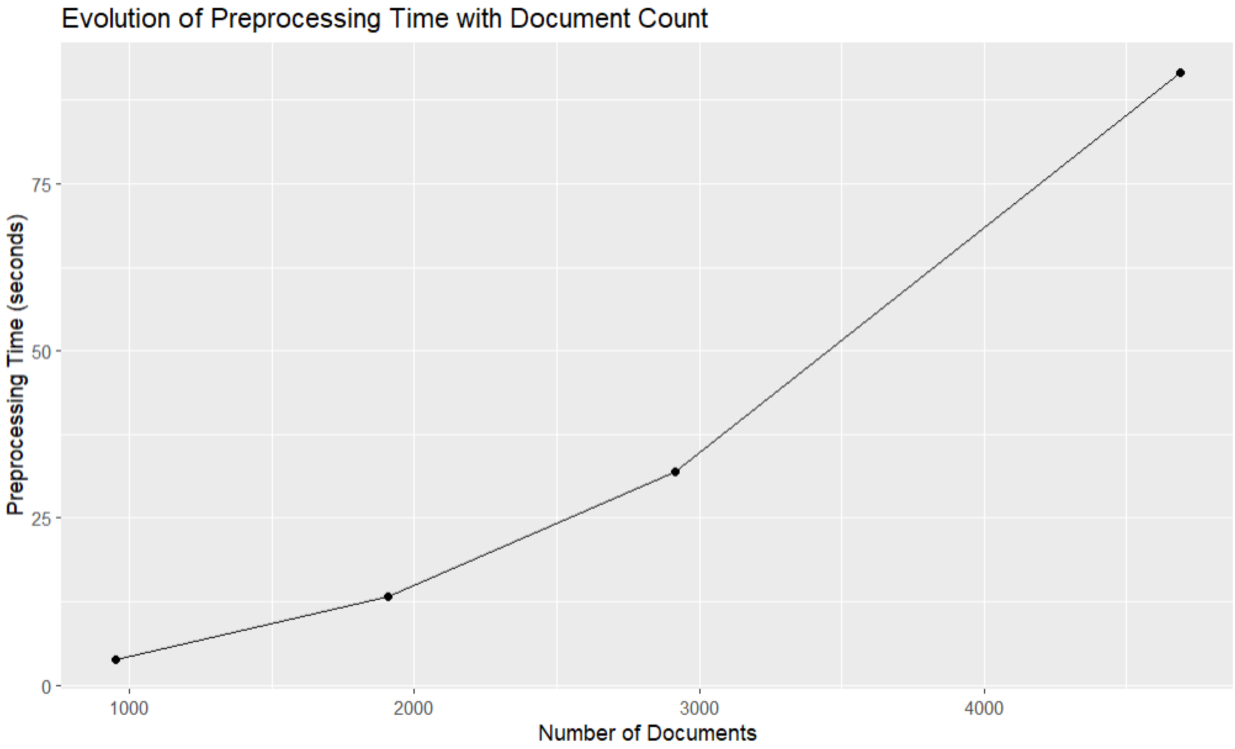
```
Pre-processing time: 91.74703407287598 seconds
Enter your search query:
doctors med medical students protest korea

Retrieval time: 0.010031938552856445 seconds
Ranking:
1 :  Institutes launch 'office workers' class' for doctor-wannabes, amid med school craze
URL:  https://www.koreaherald.com/view.php?ud=20240307050898
2 :  Gov't says adjusting med school enrollment hike is 'physically not impossible'
URL:  https://www.koreatimes.co.kr/www/nation/2024/04/281_372273.html
3 :  Students, parents slam med school quota allocation as 'reverse discrimination,' file lawsuit
URL:  https://www.koreaherald.com/view.php?ud=20240320050759
```

Lastly I tried an execution containing 4689 documents, this include all news from “The Korea Herald” and from “The Korea times” from 19th of April until the 28th of February. The pre-processing time for getting all the weights was 91.74703407287598 seconds and the retrieval time was 0.010031938552856445 seconds.

6.2 Time performance conclusion

With al the information gathered I made a plot in R-Studio to see how time complexity grows.



As we can observe, as it is normal time increases with the numebr of documents. The thing is we can start to see an exponential tendency with time. This may be due to different reasons.

As we have more documents, more words and indexes have to be created as well as number of complex operations. This operations need more memory and processing power and for average computers like mine it may take more time. Overall, I would say that even having an exponential tendency for time, my system Works well for a reasonable amount of documents.

7. Things to improve

My code isn't perfect and there is a couple things that could be made for improving the precision in the results and also time.

The more detailed the query is the better it works with my system. For improving this issue and avoid the need to write similar words in the same query, dictionaries could be created that contain synonyms for words. For example, if a query has the word “doctor” in it, it would also consider words like “doc”, “medic”…

For improving time and the experience for the user, what could be done is pre-processing all the documents in different moments. Like this, documents could be pre-processed only once and saved into memory and making the code so that if more documents are added, there is no need to also pre-process the ones already in memory.

8. Final conclusion

My information systems work as intended. The Boolean System works perfectly and the Vectorial system has a really high precision and is pretty well optimized. It works well for a good amount of documents and gives the results with precision as intended.