Assignment #1

CS 455 (Introduction to NLP)

Information Retrieval System for Judgments

Submitted By**: Muhammad Tanzeel Saleem**

Registration No.**: 04071913026**

Submitted To**: Dr. Akmal Saeed Khattak**

Introduction:

This project applies NLP techniques to analyze legal judgments from the Supreme Court of Pakistan. An Information Retrieval System has been developed to assist lawyers in finding relevant judgments quickly. NLP simplifies the process of searching for information and extracting insights from unstructured text. The system enhances efficiency and effectiveness in legal research by automating these tasks. It benefits legal professionals in analyzing and interpreting the judgments more efficiently.

Overview:

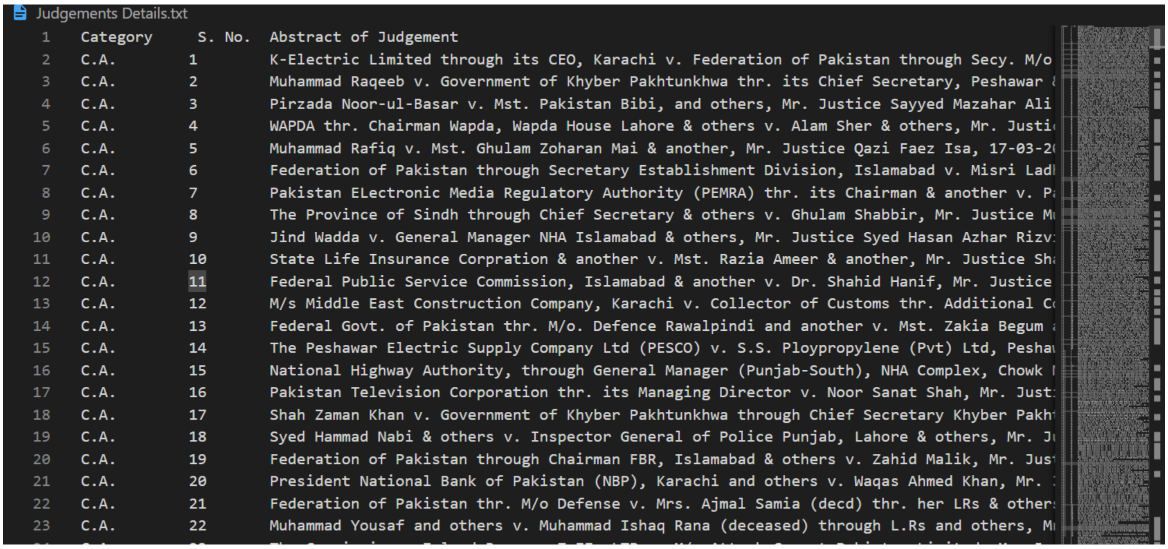
1. Data Collection:

In the initial phase of the project, the website of the Supreme Court of Pakistan was scraped to collect a substantial collection of legal documents for NLP analysis. Selenium, a popular framework for web automation, was utilized to navigate the website and extract the required data. The acquired dataset consisted of legal judgments categorized into six different categories.

I. C.A.
II. C.M.A.
III. C.P.
IV. Const.P.
V. Crl.A.
VI. Crl.P.

1047 judgments were recorded, along with two text files: one containing

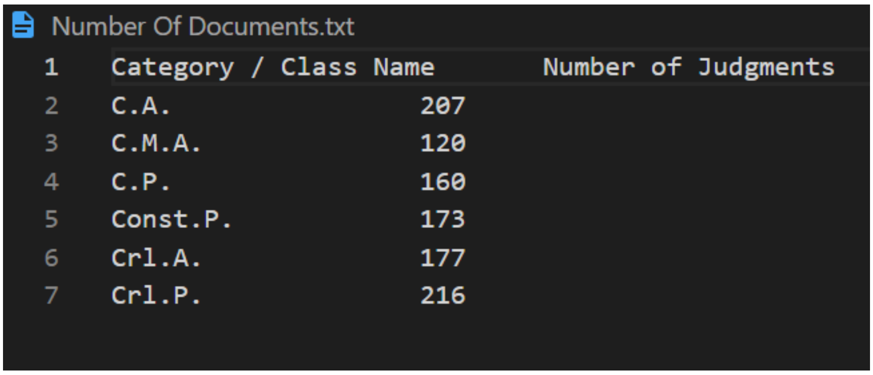
Document Abstracts



Judgements Details.txt

```
Category     S. No.   Abstract of Judgement
C.A.         1        K-Electric Limited through its CEO, Karachi v. Federation of Pakistan through Secy. M/o
C.A.         2        Muhammad Raqeeb v. Government of Khyber Pakhtunkhwa thr. its Chief Secretary, Peshawar
C.A.         3        Pirzada Noor-ul-Basar v. Mst. Pakistan Bibi, and others, Mr. Justice Sayyed Mazahar Ali
C.A.         4        WAPDA thr. Chairman Wapda, Wapda House Lahore & others v. Alam Sher & others, Mr. Justi
C.A.         5        Muhammad Rafiq v. Mst. Ghulam Zoharan Mai & another, Mr. Justice Qazi Faez Isa, 17-03-2
C.A.         6        Federation of Pakistan through Secretary Establishment Division, Islamabad v. Misri Lad
C.A.         7        Pakistan ELectronic Media Regulatory Authority (PEMRA) thr. its Chairman & another v. P
C.A.         8        The Province of Sindh through Chief Secretary & others v. Ghulam Shabbir, Mr. Justice M
C.A.         9        Jind Wadda v. General Manager NHA Islamabad & others, Mr. Justice Syed Hasan Azhar Rizv
C.A.         10       State Life Insurance Corpration & another v. Mst. Razia Ameer & another, Mr. Justice Sh
C.A.         11       Federal Public Service Commission, Islamabad & another v. Dr. Shahid Hanif, Mr. Justice
C.A.         12       M/s Middle East Construction Company, Karachi v. Collector of Customs thr. Additional C
C.A.         13       Federal Govt. of Pakistan thr. M/o. Defence Rawalpindi and another v. Mst. Zakia Begum
C.A.         14       The Peshawar Electric Supply Company Ltd (PESCO) v. S.S. Ployproplene (Pvt) Ltd, Pesha
C.A.         15       National Highway Authority, through General Manager (Punjab-South), NHA Complex, Chowk
C.A.         16       Pakistan Television Corporation thr. its Managing Director v. Noor Sanat Shah, Mr. Just
C.A.         17       Shah Zaman Khan v. Government of Khyber Pakhtunkhwa through Chief Secretary Khyber Pakh
C.A.         18       Syed Hammad Nabi & others v. Inspector General of Police Punjab, Lahore & others, Mr. J
C.A.         19       Federation of Pakistan through Chairman FBR, Islamabad & others v. Zahid Malik, Mr. Jus
C.A.         20       President National Bank of Pakistan (NBP), Karachi and others v. Waqas Ahmed Khan, Mr.
C.A.         21       Federation of Pakistan thr. M/o Defense v. Mrs. Ajmal Samia (decd) thr. her LRs & other
C.A.         22       Muhammad Yousaf and others v. Muhammad Ishaq Rana (deceased) through L.Rs and others, M
```

and another indicating **document counts by category**

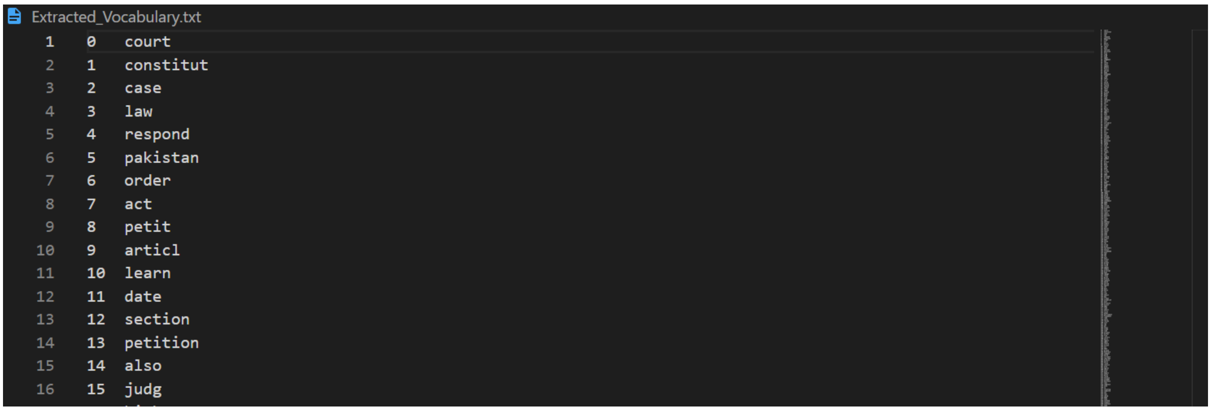


Number Of Documents.txt

```
Category / Class Name         Number of Judgments
C.A.                    207
C.M.A.                  120
C.P.                    160
Const.P.                173
Crl.A.                  177
Crl.P.                  216
```

2. Vocabulary Extraction:

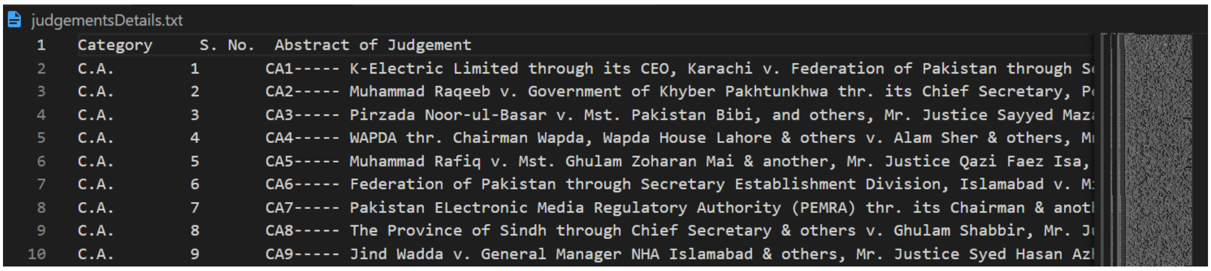
In Task 2 Using the Natural Language Toolkit (nltk) in Python, the documents underwent preprocessing steps, including tokenization, punctuation removal, stop word removal, filtering out short words, stemming, and lemmatization. The resulting unique words, totaling 43,214 were stored in a specific format within the vocabulary file, with each line representing a word and its corresponding index.



Extracted_Vocabulary.txt

```
0   court
1   constitut
2   case
3   law
4   respond
5   pakistan
6   order
7   act
8   petit
9   articl
10  learn
11  date
12  section
13  petition
14  also
15  judg
```

During this stage of the project, a document index file was generated. The file included the names of the documents and their respective indices. This file served as a reference point for the project, providing information about each document.

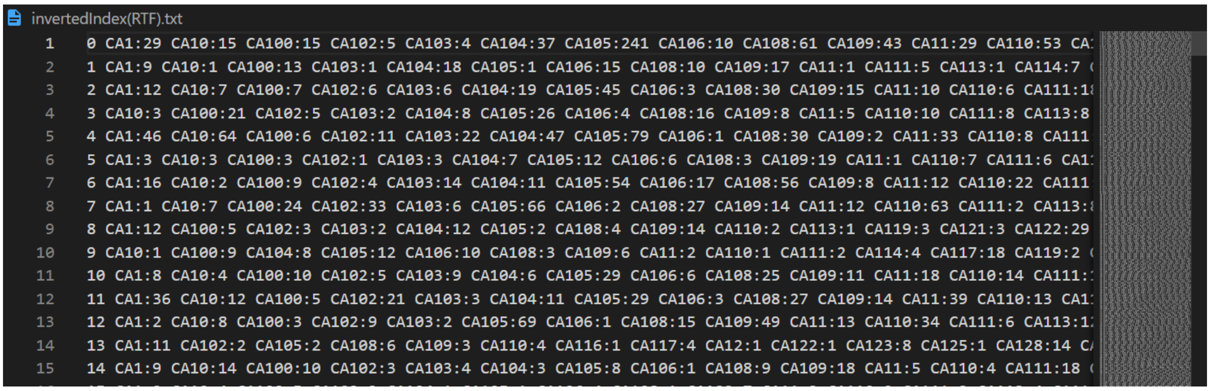


judgementsDetails.txt

```
Category      S. No.  Abstract of Judgement
C.A.          1       CA1----- K-Electric Limited through its CEO, Karachi v. Federation of Pakistan through S
C.A.          2       CA2----- Muhammad Raqeeb v. Government of Khyber Pakhtunkhwa thr. its Chief Secretary, P
C.A.          3       CA3----- Pirzada Noor-ul-Basar v. Mst. Pakistan Bibi, and others, Mr. Justice Sayyed Maz
C.A.          4       CA4----- WAPDA thr. Chairman Wapda, Wapda House Lahore & others v. Alam Sher & others, M
C.A.          5       CA5----- Muhammad Rafiq v. Mst. Ghulam Zoharan Mai & another, Mr. Justice Qazi Faez Isa,
C.A.          6       CA6----- Federation of Pakistan through Secretary Establishment Division, Islamabad v. M
C.A.          7       CA7----- Pakistan ELectronic Media Regulatory Authority (PEMRA) thr. its Chairman & anot
C.A.          8       CA8----- The Province of Sindh through Chief Secretary & others v. Ghulam Shabbir, Mr. J
C.A.          9       CA9----- Jind Wadda v. General Manager NHA Islamabad & others, Mr. Justice Syed Hasan Az
```

3. Generation of Inverted Index:

The Task 3 involved constructing an inverted index, which efficiently maps terms to the documents they appear in. By utilizing term frequencies, the index enables fast searching of legal documents based on query terms. This step greatly enhances the project's natural language processing capabilities.



```
0 CA1:29 CA10:15 CA100:15 CA102:5 CA103:4 CA104:37 CA105:241 CA106:10 CA108:61 CA109:43 CA11:29 CA110:53 CA1
1 CA1:9 CA10:1 CA100:13 CA103:1 CA104:18 CA105:1 CA106:15 CA108:10 CA109:17 CA11:1 CA111:5 CA113:1 CA114:7
2 CA1:12 CA10:7 CA100:7 CA102:6 CA103:6 CA104:19 CA105:45 CA106:3 CA108:30 CA109:15 CA11:10 CA110:6 CA111:1
3 CA10:3 CA100:21 CA102:5 CA103:2 CA104:8 CA105:26 CA106:4 CA108:16 CA109:8 CA11:5 CA110:10 CA111:8 CA113:8
4 CA1:46 CA10:64 CA100:6 CA102:11 CA103:22 CA104:47 CA105:79 CA106:1 CA108:30 CA109:2 CA11:33 CA110:8 CA111
5 CA1:3 CA10:3 CA100:3 CA102:1 CA103:3 CA104:7 CA105:12 CA106:6 CA108:3 CA109:19 CA11:1 CA110:7 CA111:6 CA1
6 CA1:16 CA10:2 CA100:9 CA102:4 CA103:14 CA104:11 CA105:54 CA106:17 CA108:56 CA109:8 CA11:12 CA110:22 CA111
7 CA1:1 CA10:7 CA100:24 CA102:33 CA103:6 CA105:66 CA106:2 CA108:27 CA109:14 CA11:12 CA110:63 CA111:2 CA113:
8 CA1:12 CA100:5 CA102:3 CA103:2 CA104:12 CA105:2 CA108:4 CA109:14 CA110:2 CA113:1 CA119:3 CA121:3 CA122:29
9 CA10:1 CA100:9 CA104:8 CA105:12 CA106:10 CA108:3 CA109:6 CA11:2 CA110:1 CA111:2 CA114:4 CA117:18 CA119:2
10 CA1:8 CA10:4 CA100:10 CA102:5 CA103:9 CA104:6 CA105:29 CA106:6 CA108:25 CA109:11 CA11:18 CA110:14 CA111:
11 CA1:36 CA10:12 CA100:5 CA102:21 CA103:3 CA104:11 CA105:29 CA106:3 CA108:27 CA109:14 CA11:39 CA110:13 CA1
12 CA1:2 CA10:8 CA100:3 CA102:9 CA103:2 CA105:69 CA106:1 CA108:15 CA109:49 CA11:13 CA110:34 CA111:6 CA113:1
13 CA1:11 CA102:2 CA105:2 CA108:6 CA109:3 CA110:4 CA116:1 CA117:4 CA12:1 CA122:1 CA123:8 CA125:1 CA128:14 C
14 CA1:9 CA10:14 CA100:10 CA102:3 CA103:4 CA104:3 CA105:8 CA106:1 CA108:9 CA109:18 CA11:5 CA110:4 CA111:18
```

Log Term Frequency:

$$1 + \ log_{10}tf_{t,d} \quad where\ tf_{t,d}\ is\ term\ frequency\ of\ term\ t\ in\ document\ d$$

Inverted Index Frequency:

$$log_{10}\left(\frac{N}{df_t}\right) \quad where\ df_t\ is\ the\ number\ of\ documents\ d\ that\ contain\ term\ t$$

TF-IDF Weighting:

$$(1 + \ log_{10}tf_{t,d}) * \ log_{10}\left(\frac{N}{df_t}\right)$$

BM25 Weighting:

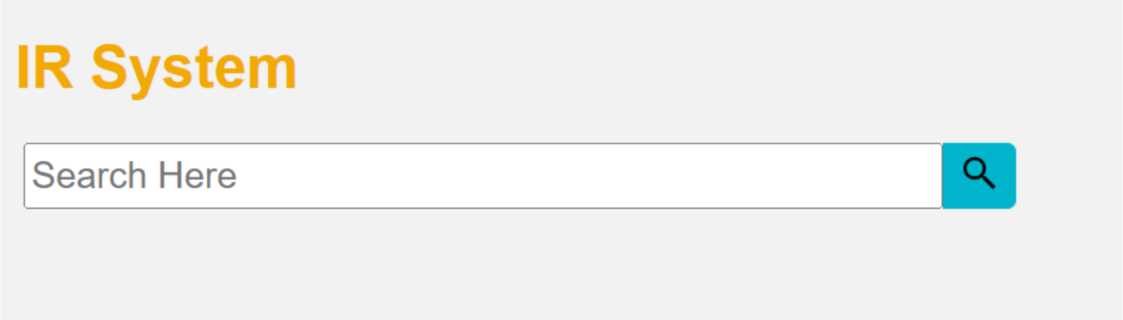
$$c(w,q)\frac{(k+1)c(w,d)}{c(w,d)+k}log\frac{M+1}{df(w)} \quad where\ k \geq 0$$

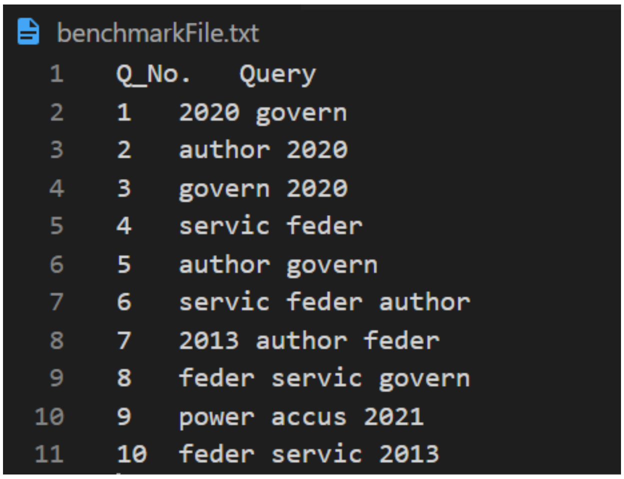
Using the above formulas in inverted indexed file, the four files were generated respectively.

4. Queries Benchmark and Web Interface:

This task involved creating a user-friendly web interface using Python Flask. The interface allows users to input queries and evaluate their relevance to the collected legal documents from the Supreme Court of Pakistan website. It features a search bar for query input and retrieval of relevant documents, ensuring a straightforward and efficient user experience.

To evaluate the system's performance, we created ten queries using the top terms from tf-idf and bm25 text files. These queries included both two-term and three-term combinations. By comparing the system's output with the expected results, we established a benchmark for system performance.





```
Q_No.   Query
1   2020 govern
2   author 2020
3   govern 2020
4   servic feder
5   author govern
6   servic feder author
7   2013 author feder
8   feder servic govern
9   power accus 2021
10  feder servic 2013
```

5. Cosine Similarity and Ranking:

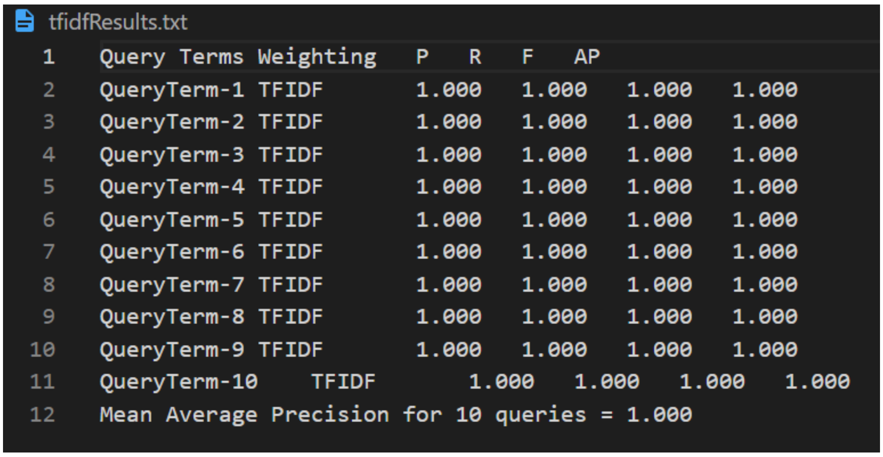
In this task, I computed the similarity between documents and queries using the cosine similarity method. The process involved the following steps:

I. I Computed the tf-idf and bm25 values for both documents and queries.
II. I Normalized the tf-idf and bm25 values of documents and queries.
III. I Calculated the cosine similarity between each document and query.
IV. I Ranked the documents based on their similarity scores.
V. I Selected the top ten documents for each query.
VI. I Wrote the selected documents to plain text files, one file per query.

The goal was to find the most relevant documents for each query using two weighting schemes (tf-idf and bm25) and store the results in plain text files.

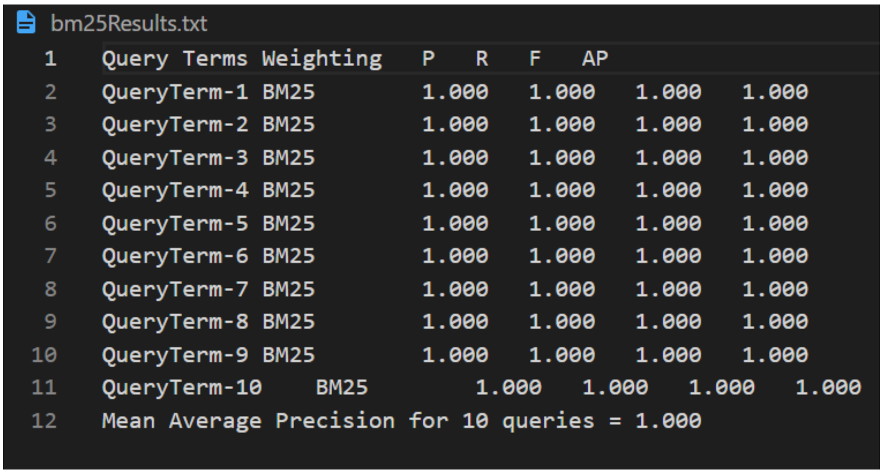
6. Evaluation:

I evaluated this information retrieval system using precision, recall, f1-measure and average precision.



tfidfResults.txt

```
Query Terms Weighting   P   R   F   AP
QueryTerm-1 TFIDF       1.000   1.000   1.000   1.000
QueryTerm-2 TFIDF       1.000   1.000   1.000   1.000
QueryTerm-3 TFIDF       1.000   1.000   1.000   1.000
QueryTerm-4 TFIDF       1.000   1.000   1.000   1.000
QueryTerm-5 TFIDF       1.000   1.000   1.000   1.000
QueryTerm-6 TFIDF       1.000   1.000   1.000   1.000
QueryTerm-7 TFIDF       1.000   1.000   1.000   1.000
QueryTerm-8 TFIDF       1.000   1.000   1.000   1.000
QueryTerm-9 TFIDF       1.000   1.000   1.000   1.000
QueryTerm-10    TFIDF       1.000   1.000   1.000   1.000
Mean Average Precision for 10 queries = 1.000
```

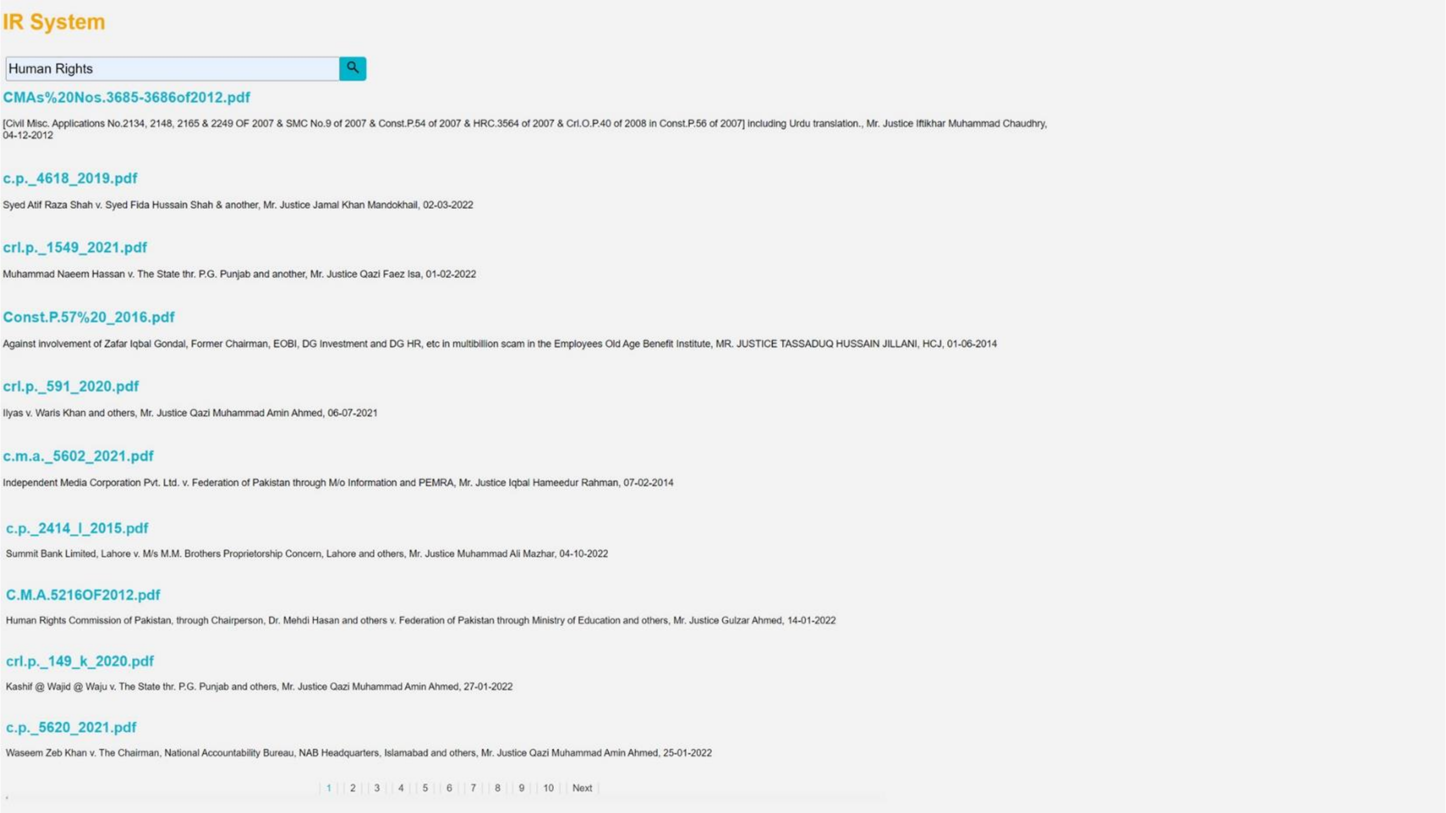


bm25Results.txt

```
Query Terms Weighting   P   R   F   AP
QueryTerm-1 BM25        1.000   1.000   1.000   1.000
QueryTerm-2 BM25        1.000   1.000   1.000   1.000
QueryTerm-3 BM25        1.000   1.000   1.000   1.000
QueryTerm-4 BM25        1.000   1.000   1.000   1.000
QueryTerm-5 BM25        1.000   1.000   1.000   1.000
QueryTerm-6 BM25        1.000   1.000   1.000   1.000
QueryTerm-7 BM25        1.000   1.000   1.000   1.000
QueryTerm-8 BM25        1.000   1.000   1.000   1.000
QueryTerm-9 BM25        1.000   1.000   1.000   1.000
QueryTerm-10    BM25        1.000   1.000   1.000   1.000
Mean Average Precision for 10 queries = 1.000
```

7. Representation:

This task improved the user interface of the search engine developed in Task 4. It retrieves documents with high cosine similarity to the user's query and displays them using pagination, with each page showing 10 documents. The word cloud feature helps users quickly understand the primary topics addressed in the displayed documents. Overall, this task enhances the usability of the search engine by providing a more intuitive and effective way for users to interact with it.



IR System

Human Rights

CMAs%20Nos.3685-3686of2012.pdf

[Civil Misc. Applications No.2134, 2148, 2165 & 2249 OF 2007 & SMC No.9 of 2007 & Const.P.54 of 2007 & HRC.3564 of 2007 & Crl.O.P.40 of 2008 in Const.P.56 of 2007] including Urdu translation., Mr. Justice Iftikhar Muhammad Chaudhry, 04-12-2012

c.p._4618_2019.pdf

Syed Atif Raza Shah v. Syed Fida Hussain Shah & another, Mr. Justice Jamal Khan Mandokhail, 02-03-2022

crl.p._1549_2021.pdf

Muhammad Naeem Hassan v. The State thr. P.G. Punjab and another, Mr. Justice Qazi Faez Isa, 01-02-2022

Const.P.57%20_2016.pdf

Against involvement of Zafar Iqbal Gondal, Former Chairman, EOBI, DG Investment and DG HR, etc in multibillion scam in the Employees Old Age Benefit Institute, MR. JUSTICE TASSADUQ HUSSAIN JILLANI, HCJ, 01-06-2014

crl.p._591_2020.pdf

Ilyas v. Waris Khan and others, Mr. Justice Qazi Muhammad Amin Ahmed, 06-07-2021

c.m.a._5602_2021.pdf

Independent Media Corporation Pvt. Ltd. v. Federation of Pakistan through M/o Information and PEMRA, Mr. Justice Iqbal Hameedur Rahman, 07-02-2014

c.p._2414_l_2015.pdf

Summit Bank Limited, Lahore v. M/s M.M. Brothers Proprietorship Concern, Lahore and others, Mr. Justice Muhammad Ali Mazhar, 04-10-2022

C.M.A.5216OF2012.pdf

Human Rights Commission of Pakistan, through Chairperson, Dr. Mehdi Hasan and others v. Federation of Pakistan through Ministry of Education and others, Mr. Justice Gulzar Ahmed, 14-01-2022

crl.p._149_k_2020.pdf

Kashif @ Wajid @ Waju v. The State thr. P.G. Punjab and others, Mr. Justice Qazi Muhammad Amin Ahmed, 27-01-2022

c.p._5620_2021.pdf

Waseem Zeb Khan v. The Chairman, National Accountability Bureau, NAB Headquarters, Islamabad and others, Mr. Justice Qazi Muhammad Amin Ahmed, 25-01-2022

3. Conclusion:

This project successfully developed a search engine for retrieving relevant PDF documents. It involved data collection, preprocessing, and constructing an inverted index. The system utilized cosine similarity and ranking techniques to provide accurate results. A user-friendly web interface was created, and the system's performance was evaluated using various metrics. Overall, the project achieved its goals of building an effective search engine with a user-friendly interface.