COMP 1859

Information Retrieval

3886 words

Tejesh Ramesh Bawa
001173097-8

Table of Contents

1. Introduction

In the dynamic environment of the healthcare system, the role of Information Retrieval (IR) system has become an integral part in improving patient care and providing for groundbreaking research. The aim of this report is to document on how an IR system is developed to assist and navigate through diverse data sources, such as structural Electronic Health Records (EHR), and unstructured clinical notes, medical images and research articles.

2. Part A: System Design

2.1 Designing an Innovative Information Retrieval System

As there are multiple requirements to fulfil by a healthcare information system. The system is required to be developed to be seamlessly integrate both the structured and unstructured data, which needs to be adapted to both medical professionals and researcher alike. This will not only bridge the gap between different data types, but also introduce to new features to improve the utilization of the developed system.

It is very important to identify a clear scope of objectives for this retrieval system. This allows to identify the specific type of information needed to retrieve from datasets. Additionally, this authorize the developer to understand who the target audience are and how to develop the system to their specific requirements.

It is important to highlight what types of retrieval methodologies will be used to develop the new retrieval system, such as Natural Language Processing (NLP), Machine learning algorithms and many others. It will also assist the designing process by understanding by which specifications does the data security and privacy are needed when developing this new retrieval system.

2.2 Architectural Components and Data Flow

The prioritized requirements for this system will be the ability to operate an effective and smooth flow of data, it also has cover and effectively utilize both the structured and unstructured data. Therefore, the system needs to be designed to implement a console-based search system to involve several key components to manage user interactions, data pre-processing, storage and displaying result.

2.2.1 User Interface and Interaction

As this is going to be a console-based system, it is necessary for the user interface to provide the user with a choice to either explore the EHR or the research articles. If the user chooses to select the EHR option, the system will prompt the user to input a name and date of birth to get specific records of that person. However, if the user opts for research articles, the system will prompt the user to search using keywords relating to the topic, in order to retrieve the relevant research articles. These interactions are very important for personalizing the ensuing search based on the information provided by the user.

2.2.2 Structured Data Storage and Handling

When developing the system, the structured EHR dataset will be stored onto a CSV file. This CSV file will contain the Name, Date of Birth, Gender, Symptoms, Causes, Disease and Medicine. To be able to manipulate and load the data the pandas library will be used. The EHR dataset will repeat to prioritize the user's pre-processed input text to identify the matching documentation to ensure that the textual data is represented in a standardize and comparable manner.

2.2.3 Unstructured Data Storage and Handling

The unstructured dataset consists of 3 CSV files which are the clinical notes, medical images and research articles. The formatting applied to the clinical notes and medical images is done in such a way that it is able to correspond to the searched name and data of birth by the user. This grants the user to be able to compare both the structured and unstructured dataset without any difficulties.

The retrieval system for the researched articles will be developed by making use of inverted index and comparing the user's keywords to the researched article titles. The system will also utilize the Term Frequency-Inverse Document Frequency (TF-IDF) to further process the search terms by ranking the article title according to their respective relevance to the searched query keywords.

2.2.4 Result Presentation

After retrieving relevant information depending on the user choice of either the EHR records or research articles, the results will be presented to the user through a detailed overview of relevant information (depending on the searched query) such as patients records, clinical notes, medical images, research articles and TF-IDF scores. This will guarantee that the user receives their desired set of results to fit the users search criteria.

2.3 Retrieval Methods, Machine Learning and NLP Techniques

2.3.1 Tokenization and Preprocessing

(Kashina et al., 2020) explains that Natural Language Processing (NLP) helps to analyse the problems of computer analysis and synthesis of natural languages. It is further explained that the NLP can help to extract valuable information from medical texts and electronic medical records. They go on to elaborate that pre-processing is an important method in retrieving data where it is used to pre-process text to clear text to bring it to a suitable format for the computer to process. This permits the system to determine the quality of tokenization of the data.

The system will make use of the 'preprocess_text' functionality when using NLP technique to tokenize and pre-process the data. The tokenization grants the system to break down the search query to individual words in order to tokenize them the system will make use of the NLTK's 'word_tokenize' method. By pre-processing the system will be able to not only tokenize the data but also remove any common stop words (such as 'is', 'a', 'the', etc) by applying the Porter Stemming Algorithm. This will create a more efficient analysis tool to only focus on important keywords.

2.3.2 Data Loading and Pandas

Pandas library will be used to effectively manage data to load up EHR records, clinical notes, and images from their respective CSV files into data frames. It is easier to work with data frames as it has a better structure for the system to retrieve searched query and manipulate data.

2.3.3 Document Retrieval-EHR Search

The retrieval process for the EHR dataset will be done by iterating by using pandas through each and every record using the 'iterrows' method. Indicating that whenever a user inputs a name and date of birth, the system will be able to pre-process the patient's name and date of birth within the dataset to compare it to what the user has searched for. Thereby allowing the system to find the suitable match which is relevant to the user's searched query.

2.3.4 Inverted Index-Research Articles Search

(Ilic et al., 2014) described inverted index into two different phases. The first phase is the index construction, where each and every text token is processed to build a list of posting for each term within the dataset in an incremental pattern. The second phase is when query processing happens.

In this phase the index that was built in phase one will be stored to be used to process search queries.

Whenever the user searches for research articles, the system will develop and construct an inverted index. After tokenization and pre-processing is completed by the system, it will draw the data structure to the titles of the research title within the dataset. This will improve the systems retrieval proficiency by making use of indexing techniques to quickly match and identify the searched terms to the articles without having to repeat within the entire dataset over and over again.

2.3.5 TF-IDF Calculation

(Ramos, n.d.) explains that TF-IDF is used to determine the words in a collection of documents. This is done by calculating a value for each word in a document through an inverse proportion of the frequency of the word in a specific document to the percentage of documents the word appears in. When the TF-IDF is high, it signifies that if the word appears in the query and document, that document is of high interest to the user.

'TfidVectorizer' will be used by the system to calculate the TF-IDF scores of the documents. This value will mirror the significance of the searched term within the group of documents. To compute the cosine similarities between the searched user query and the datasets, which will allow the system to measure the cosine of these two vectors and thereby providing a metric for the document relevance.

3. Part B: Development and Implementation

3.1 Tools, Technology and File Format

It is important to understand that the system should be able to provide relevant and appropriate information to its users, based on the combination of techniques integrated into the system, such as retrieval methods, algorithms and the natural language processing (NLP). To do this the system will need to be built on specific tools, technologies, and file formatting.

3.1.1 Tools

PyCharm

According to (DataScientest, 2023), PyCharm is an Integrated Development Environment (IDE) developed by JetBrains. It significantly simplifies the programming and development process. As it is a hybrid platform mostly used for python and has many supporting technologies and libraries used for information retrieval. It has numerous plugins and productivity shortcuts which are a valuable asset and it generally easier and faster to develop information retrieval system with PyCharm.

3.1.2 Technologies

Python

According to (Magnimind, n.d.), python is one of the most versatile programming languages because it mirrors human language as it has consistent syntax and many packages for the code reusability. Due to its semantics and syntax being transparent, it makes it a great choice for information retrieval system. Python also has many libraries regrading natural language processing, which is a great advantage as it is difficult to develop software which can handle natural language.

Pandas

(McIntire et al., n.d.) describes python pandas as a very important technology in the information retrieval system. Pandas is a very powerful tool when it comes to machine learning and visualization,

because of this it has now become the backbone of many data retrieval projects. It helps to explore, clean, transform and analyse dataset, which is a very important task in information retrieval.

Natural Language Toolkit (NLTK)

(TutorialFreak, n.d.) emphasizes that the NLTK is an important library used in information retrieval. This is because it is used to remove common stop words such as 'the', 'is', 'and' etc which do not contribute to the overall meaning of the text. They further explain that removing them can help improve accuracy and efficiency of text analysis and natural language processing task. This will help the system to focus on the essential keywords and improve the relevance of the results.

Porter Stemmer

(GeeksforGeeks, n.d.) describe stemming as a natural language processing technique used to reduce words to their root form. The porter stemmer is one of the most widely used algorithms and is based on a set of heuristics that are used to remove common suffixes from words. The stemmer is widely known for its speed, simplicity and less error rates compared to other stemmers; however, a major disadvantage is that the produced variants are not always real words when comparing to the content of the original text.

3.1.3 File Formatting

CSV Files

Comma Separated Value files (CSV Files) are files where data is actually input as data that is separated by commas as explained by (Lahar, 2020). CSV files are also editable, and changes are not locked, however users can lock certain sets of data from being edited which is very important in the case sensitive EHR dataset where the names and date of birth should not be changes, instead only the medicine and disease can be altered with passing time.

3.2 Security and Privacy Concerns

It is important to be aware, when dealing with sensitive healthcare information. The system will need to be efficient, consistent and safe to retrieve records from the EHR, images, clinical notes and research articles. Due its data sensitivity, the system will should be able to make the data security and privacy the number one priority. This can be accomplished by integrating vigorous measures such as encryption, access controls, training and policies.

(Rajkamal et al., 2021) explains that the conventional method of encrypting entire datasets leads to a higher computational complexity and reduced data usability to authorized users. It is emphasized that securing sensitive information depends on whether the data is structured or unstructured. In a structured dataset the encryption is operated on all sensitive attributes, while unstructured datasets, the encryption is only operated on the sensitive attribute retrieved.

For access controls, when the users (doctors) are prompted by the system to choose between 2 option the EHR or research articles. For the EHR, the system will prompt the user to enter the name and date of birth. This will keep confidentiality between the doctor and the patients as the doctors will know the name but not the date of birth which protects the information of the patient from unauthorized access.

The users who will be using the system will need to be trained on the data security and privacy policies. This is to guarantee that the staff are well aware of the importance of safeguarding patient information. By embracing all these protective measures there will be a trustworthy foundation between the doctor and their patients for responsible and ethical use of the digital healthcare date.

3.3 Functionality of the prototype

3.3.1 Choose Option

```
Choose an option:
1. Search the EHR of Patients
2. Search Different Types of Research Articles
Enter the number of your choice (1 or 2):
```

When the system will prompt the user to choose between 2 options to either search for patients through the HER dataset or to search for research articles by either pressing 1 or 2.

3.3.2 Invalid Option

```
Enter the number of your choice (1 or 2): 3
Invalid choice. Please enter 1 or 2.
Enter the number of your choice (1 or 2):
```

If the user enters anything beside the 1 or 2 the system will give back feedback, by explaining that their choice was invalid, and they should enter 1 or 2.

3.3.3 Opting for Option 1

```
Enter the number of your choice (1 or 2): 1
You Chose Search the EHR of Patients
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\tejes\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\tejes\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
Enter the name:
```

If the user opts to choose option 1, the system will download the punkt package using the NLTK to pre-process the text, and thereby removing the stop words. All this is done efficiently, quickly and smoothly in the background. After pre-processing and stemming the system will prompt the user to enter the name and data of birth.

3.3.4 Entering Details

```
Enter the name: John Doe
Enter the date of birth (DD/MM/YYYY): 15/05/1980
Matching documents from EHR:
Name: John Doe
Date of Birth: 15/05/1980
Gender: Male
Symptoms: Fever, Cough
Causes: Viral Infection
Medicine: Ibuprofen, Rest
Matching documents from Clinical Notes:
Clinical Notes: Difficulty falling asleep
Annotation: ['no able to fall asleep']
Matching documents from Images:
Images: Picture
```

When the user enters the correct information regrading the patient's name and data of birth, their full record will be displayed which includes the EHR, clinical notes and the medical images.

3.3.5 Entering Incorrect Details

```
Enter the name: test
Enter the date of birth (DD/MM/YYYY): 13/12/2021
No matching documents found for the given name and date of birth.
No matching documents found for the given name and date of birth.
No matching documents found for the given name and date of birth.
```

When an incorrect search is made the system will display that no matching documents were found with the give name and date of birth. This helps provide security for the patients as their data will not be accessible by the doctors unless they know both the full name and date of birth.

3.3.6 Choosing Option 2

```
Enter the number of your choice (1 or 2): 2
You Chose Search Different Types of Research Articles
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\tejes\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\tejes\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
Enter your query:
```

If the user chooses to option 2, then the system will run the same background NLTK to pre-process and tokenise the terms in the document. The system will then prompt the user to search for a query.

3.3.7 Entering a Query

```
Enter your query: heart
Search query terms heart
Search Results with Tf-IDF Scores for Query:
Document Title: liver function abnormalities in heart failu
Document Content: BACKGROUND    This study analyzed liver
RESULTS A post hoc analysis was conducted with the use of
RESULTS Liver function tests ( LFTs ) were measured at 7 ti
RESULTS Survival analyses were used to assess the associati
RESULTS The percentage of patients with abnormal LFTs decre
RESULTS When mean hemodynamic profiles were compared in pat
RESULTS Multivariable analyses revealed that patients with
CONCLUSIONS Abnormal LFTs are common in the ADHF populatio
CONCLUSIONS Elevated MELD-XI scores are associated with poo
TF-IDF Score: 0.40824829046386296
```

When the user search for the keyword heart the system retrieved the research article based on the searched term. The results show the document title, and content. The system also displays the TF-IDF score to show the relevance of the document retrieved. Although it the TF-IDF is shown in the console it will not be needed by the doctors.

3.3.8 Background Results

```
No matching documents found for title: endovascular aneurysm repair ( EVAR )
No matching documents found for title: social anxiety
No matching documents found for title: diesel exhaust causes inflammatory responses
No matching documents found for title: Depressive disorders
No matching documents found for the query.
```

There is also a background process operating, which compares the search term with each and every document to saw the relevance to the query.

3.3.9 When Documents are not Found for Query Keyword

```
Enter your query: cisco
No matching documents found for the query.
```

If the query does not have any relating documents, the system will provide feedback to the user by displaying no matching documents found for the query.

4. Part C: Evaluation and Optimization

4.1 Evaluation Plan

4.1.1 Precision, Recall and F1 Scores

The system will be evaluated using the precision, recall and F1-scores by making use of how well the system retrieves the desired data. This will be done by effectively using True Positives (TP), False Positives (FP) and False Negatives (FN).

The formulas used are:

- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

For the EHR dataset, these were the respective scores:

- Precision = 1.0
- Recall = 1.0
- F1 Score = 1.0

As seen all the scores suggest that the EHR retrieval is a highly accurate model. As much as the perfect scores are desirable, it is important to understand that it is highly likely that the system could be overfitting.

For the Research Articles dataset, these were the respective scores:

- Precision = 1.0
- Recall = 0.33333 (1/3)
- F1 Score = 0.5

A 1.0 for precision indicates that there are no false positive predicted by the system. However, a low recall score suggests that there is a high number of false negatives. Because of this, the F1 score is brought down to 0.5 but it showcases a balance between the precision and recall. To further optimize the system, the recall needs to be improved without sacrificing precision.

3.1.2 Testing

No.	What is Being Tested	How	Where	Expected Results	Date	Actual Results	Action Taken
1	Feedback when wrong value is entered when choosing options	Entering wrong values when choosing options	At the start, when the system prompts the user to choose options	Give feedback to the user of why it is wrong	1/12/2023	The system provided feedback	None
2	Choosing option 1	Entering the number 1 when choosing options	At the start, when the system prompts the user to choose options	Will take the user to option 1 to enter name and date of birth	2/12/2023	The system takes the user to the option 1 to enter name and date of birth	None
3	Correct retrieval of patient's records	Enter the correct name and date of birth	At option 1	Display correct records of the patient	2/12/2023	It retrieved information only for the EHR dataset and not for the notes and images.	There was a problem with how the dataset were structure and with the code function and methods
4	Feedback when incorrect information is entered for option 1	Entering incorrect name and date of birth	At option 1	Give feedback to the user that the system does not have the record	2/12/2023	The system provided feedback	None
5	Choosing option 2	Entering the number 2 when choosing options	At the start, when the system prompts	Will take the user to option 2 to search	3/12/2023	The system takes the user to the option 1 to enter name	None

			the user to choose options	their query		and date of birth	
6	Correct retrieval of research articles	Enter query keywords	At option 2	Display correct research articles	3/12/2023	Kept on displaying all the data	The problem was the functions, this affected the data retrieval
7	Feedback when there are no relating articles for option 2	Entry query which is not in the dataset	At option 2	Give feedback to the user that there are no relevant articles relating to the query keywords	3/12/2023	The system provided the appropriate feedback to the user	None
8	Display TF-IDF	The system will compute it and displayed when query is searched	At option 2	The TF-IDF score will be displayed	4/12/2023	The results were not appearing	There was in print functionality placed, therefore the system did not print it
9	Read CSV files	Is the system able to read the CSV file for it to carry out its computation	At the start of the coding phase	The system is able to read the CSV file without any errors	9/11/2023	There was an error when the system tried to read the CSV files	The CSV files was not stored in the same folder
10	Search using multiple keywords	Search for research articles using multiple keywords	At option 2	The system should retrieve relevant articles relating to the searched keywords	5/12/2023	The system was able to retrieve the relevant articles.	None

4.2 Optimization, Feedback and Technological Advancement

The developed system has been built to provide the basic framework for search through electronic health records (EHR) and different types of research articles depending on what the user has searched. However, all systems have short comings and can be developed to be better and more robust to fulfil the changing requirements in the real world.

The code itself will be needed to be broken down into smaller functions to improve the overall readability, reusability and maintainability. This will allow the system to be set according to the industrial standards. This will not only make it easier for other developers to understand the code easier, but it also helps when the system will be required to be updated on later dates. It will also be important for the developed system to take appropriate feedback from medical professionals to understand what is lacking in the system. This can be done in forms of questionnaires, interviews and user studies.

As this is a prototype, the system can be a console-based retrieval system. However, later when the system is required to be launched, the system will be required to improving user interface to make it more user friendly. This can be achieved by developing Graphical User Interface (GUI) or creating a web-based interface. This is important as the main user will be medical professionals, who are not well versed in technical console-based system. This will allow them to feel the system to be more user friendly and intuitive.

Considering this system is built for medical purposes, both the structured and unstructured data will be huge. Therefore, it will not be optimal to keep using CSV files to store data. The best approach would be for the data to be stored in databases such as MongoDB as it can store and operate all structured, semi-structured and unstructured data. It all has the ability to encrypt the data based on the system requirements.

The system will also require building a logging-based system to improve the security. This will allow for access controls and monitor which data is being accessed by which user. This allows the system to adhere to the industrial based security practices while also adhering to the GDPR. The system can also implement machine learning algorithms to learn from user interaction and enhance the system's capabilities to give the user a more personalized experience.

All these strategies will collectively contribute to improve the optimization and technological advancement by considering the feedback from the medical professionals. This will have improved the system and make the system more user centric, best security protocols and the most important better information retrieval of important and sensitive information in the field on healthcare.

5. Conclusion

In conclusion, the developed information retrieval system has 3 key phases. In Part A, it is emphasised on the system design and how can structured and unstructured healthcare data can be integrated seamlessly to retrieve information. This can be done by making use od retrieval methods, algorithms and natural processing techniques. Part B is the phase where the system is being developed and implemented, here technologies and tools are being selected to develop the retrieval system. In this phase data security and privacy measures are being enforced, as well as demonstrating the functionalities of the prototype. In Part C the system is being evaluated to bring the system to the next level and make it the industry's best standards. Overall, the system fulfils the desired retrieval methodology within the dynamic healthcare landscape.

6. URL to Prototype

6.1 Link to Google Colab

https://colab.research.google.com/drive/1Z2DYznpMdifUVtk-nInxktTZbo6sFgYb?usp=sharing

6.2 Link to GitHub repository

https://github.com/TejeshRameshBawa/COMP-1859-Information-Retrieval.git

7. References

DataScientest. (2023, February 20). *PyCharm: all about the most popular Python IDE*. https://datascientest.com/en/pycharm-all-about-the-most-popular-python-ide

GeeksforGeeks. (n.d.). *Introduction to Stemming - GeeksforGeeks*. Retrieved December 14, 2023, from https://www.geeksforgeeks.org/introduction-to-stemming/

Ilic, M., Spalevic, P., & Veinovic, M. (2014). *Inverted Index Search in Data Mining*.

Kashina, M., Lenivtceva, I. D., & Kopanitsa, G. D. (2020). Preprocessing of unstructured medical data: The impact of each preprocessing stage on classification. *Procedia Computer Science*, *178*, 284–290. https://doi.org/10.1016/j.procs.2020.11.030

Lahar, S. (2020, December 16). *What is a CSV File? Guide to Uses and Benefits*. https://flatfile.com/blog/what-is-a-csv-file-guide-to-uses-and-benefits/

Magnimind. (n.d.). *What Programming Languages Are Suitable For Natural Language Processing?* Retrieved December 14, 2023, from https://magnimindacademy.com/blog/what-programming-languages-are-suitable-for-natural-language-processing/

McIntire, G., Martin, B., & Washington, L. (n.d.). *Python Pandas Tutorial: A Complete Introduction for Beginners – LearnDataSci*. Retrieved December 14, 2023, from https://www.learndatasci.com/tutorials/python-pandas-tutorial-complete-introduction-for-beginners/

Rajkamal, M., Sumathi, M., Vijayaraj, N., Prabu, S., & Uganya, G. (2021). *Sensitive Data Identification and Protection in a Structured and Unstructured Data in Cloud Based Storage* (Vol. 25, Issue 2). http://annalsofrscb.ro

Ramos, J. (n.d.). *Using TF-IDF to Determine Word Relevance in Document Queries*.

TutorialFreak. (n.d.). *Remove Stop Words from String in Python (With NLTK & spaCy)*. Retrieved December 14, 2023, from https://www.tutorialsfreak.com/python-tutorial/examples/remove-stop-words-python-string