International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 04 April-2026 | Impact Factor: 3.5

The Document Similarity & Deduplication Tool

Mrs. B Saritha 1, Pogula Sandhya 2, P Vamshi Goud 3, Banothu Narahari 4, Chittoju Koushik Avinash 5

[1]Assitant Professor, Department of CSE (Data Science), ACE Engineering College, Hyderabad, Telangana, India

[2,3,4,5] III B.Tech. Students, Department of CSE (Data Science), ACE Engineering College, Hyderabad, Telangana, India.

How to Cite this Article:

Saritha, B., Sandhya, P., Goud, P. V., Narahari, B. & Avinash, C. K. (2026). The Document Similarity & Deduplication Tool. International Journal of Creative and Open Research in Engineering and Management, 02(04). https://doi.org/10.55041/ijcope.v2i4.163

https://doi.org/ 10.55041/ijcope.v2i4.163

Abstract

The Document Similarity and Deduplication Tool is made to solve the problem of having much repeated data in digital storage. Every day more and more documents are being. When there are duplicate or almost duplicate files it takes up space that is not needed and makes it hard to manage the data. The Document Similarity and Deduplication Tool uses Natural Language Processing techniques to prepare the text so the documents can be compared and analyzed correctly. The Document Similarity and Deduplication Tool system uses ways to measure how similar things are, like Cosine Similarity, Jaccard Index and Euclidean Distance to figure out how similar the documents are to each other. When the Document Similarity and Deduplication Tool finds documents it gives the user choices about what to do with them. The user can merge the documents put them in an archive or delete them depending on what the user wants and this makes the data cleaner and uses storage space. The Document Similarity and Deduplication Tool has a web interface that uses HTML, CSS and JavaScript. It lets the user upload files see how similar they are and manage the duplicates easily all in one place.

1. Introduction

The Document Similarity and Deduplication Tool proves its worth when you're swamped with digital documents. Daily, individuals. Store a huge number of files. This leads to a lot of duplicated and almost duplicated documents. As a result we waste storage space. It gets confusing and hard to find the data we need. The old way of doing things looks at the names of the files how big they are or special codes. This only finds exact matches. It does not find documents that're similar but not exactly the same.

The Document Similarity and Deduplication Tool solves this problem by using techniques from Python and Natural Language Processing, such as breaking down words removing common words and making everything the same. This helps us compare the text in a way that makes sense. Then it uses math like Cosine Similarity, Jaccard Index and Euclidean Distance to figure out how similar the documents are and find the duplicates.

The Document Similarity and Deduplication Tool has a website where people can upload their documents see how similar they are and get rid of the duplicates. This website is made with HTML, CSS and JavaScript. This system helps us use storage space better makes the data more accurate reduces duplicates and keeps the data clean.
The Document Similarity and Deduplication Tool is a solution to make managing documents easier save time and keep everything organized whether it is, for personal or work use.

2. Literature review

Research on document similarity and duplicate removal has led to various methods for managing redundant copies in digital storage systems. On people used methods like hashing and looking at files to remove duplicates, which worked well for exact copies but did not catch similar meanings. Researchers such as S. Patel and others in the year 2023 tried using TF-IDF with Cosine Similarity, which was better for matching text but still not great at understanding the meaning of the document similarity. In 2024, R. Kumar and his team introduced faster methods like MinHash and Locality Sensitive Hashing. However, these methods had trouble when dealing with documents that had many changes. Document similarity methods that used BERT and transformers like the ones used by A. Gupta and others in the year 2024 and J. Lee and others in the year 2025 understood the context better.

Other studies looked at removing duplicates in areas like genome sequencing, where methods based on similarity and special encoding like the ones used by Vinicius Vielmo Cogo and others in the year 2023 saved storage space but did not compress as well as methods made just for that area and used for document similarity. Some researchers, like Xingjun Zhang and Runting Zhao in the year 2021 focused on finding files in a scalable way but had performance issues with very big sets of data and document similarity. While old methods are great, at finding duplicates new methods using natural language processing and machine learning can find similar documents at a semantic level but they are more complex. This shows that we need systems that balance speed with understanding the meaning of documents to handle all kinds of document collections and document similarity effectively.

3. Existing System

Existing systems for document deduplication mostly use methods like comparing file names, matching file sizes or checking for exact duplicates using hashes. These methods are good at finding files but they do not find documents that are almost the same with just a few small differences. Some storage solutions use block-level deduplication, which finds repeated blocks of data. This only works for exact matches. There are also tools that compare documents line by line but they do not understand what the content means. So these systems often miss documents that're similar in meaning and the problem of redundancy is not fully solved. Overall these systems only provide solutions and have trouble handling large collections of documents.

The problems with these existing systems become clear when dealing with varied sets of data. Since they cannot find documents that're similar in meaning documents with content that is reworded or slightly changed are treated as unique which wastes storage space. People often have to check for documents that are almost the same which takes a lot of time and is not efficient. Also these systems are not good for data environments, where millions of documents need to be processed quickly. The fact that they do not use natural language processing makes it harder for them to understand the context, which makes them unsuitable for document deduplication tasks. As a result while current systems reduce redundancy to some extent they are not good enough to provide intelligent and efficient document management solutions, for document deduplication.

4. Proposed System

The Document Similarity and Deduplication Tool is a system that helps fix the problems of methods. It uses Natural Language Processing techniques and special algorithms to compare documents in a way. First it gets the text ready by breaking it down into parts removing common words and making everything consistent. This way, the Document Similarity and Deduplication Tool can compare documents in a way that makes sense. The system then uses things like Cosine Similarity, Jaccard Index and Euclidean Distance to figure out how similar documents are and find overlapping or redundant content. When it finds duplicates the Document Similarity and Deduplication Tool gives users options to merge documents, archive files that are not needed or delete duplicate entries. This helps save space and makes the data more accurate.

The Document Similarity and Deduplication Tool has a web-based interface that's easy to use. Users can upload documents see how similar they are and manage duplicates with a dashboard. This means that anyone can use the Document Similarity and Deduplication Tool even if they are not good with technology. The Document Similarity and Deduplication Tool does not just remove files it also makes sure the data is accurate and reliable. By using Natural Language Processing and special algorithms the Document Similarity and Deduplication Tool is a practical and efficient way to manage documents. It is good for both organizational use. Other systems can remove some files but they do not do as good of a job, as the Document Similarity and Deduplication Tool at managing documents in a smart and efficient way.

5. Methodology

The document similarity and deduplication tool works by getting the text ready for comparison. It does this by breaking down the text into parts removing common words that do not mean much and making sure everything is in the same format. Then it uses ways to measure how similar the documents are to each other like cosine similarity, jaccard index and euclidean distance. The document similarity and deduplication tool uses these measurements to figure out how similar the documents are. The document similarity and deduplication tool then looks at these scores to find documents that're duplicates or very similar. It gives the user options to combine these documents store them away or get rid of them. The user can do all of this with a web-based interface that was built with HTML, CSS and JavaScript. This interface lets the user upload documents look at the results. Manage duplicates in a dashboard that is easy to use.

The document similarity and deduplication tool helps to make sure that storage space is used efficiently that the data is accurate and that the repositories are organized. The document similarity and deduplication tool is not too heavy it can handle a lot of data. It is practical to use. This makes the document similarity and deduplication tool a choice, for many different types of applications whether they are big or small.

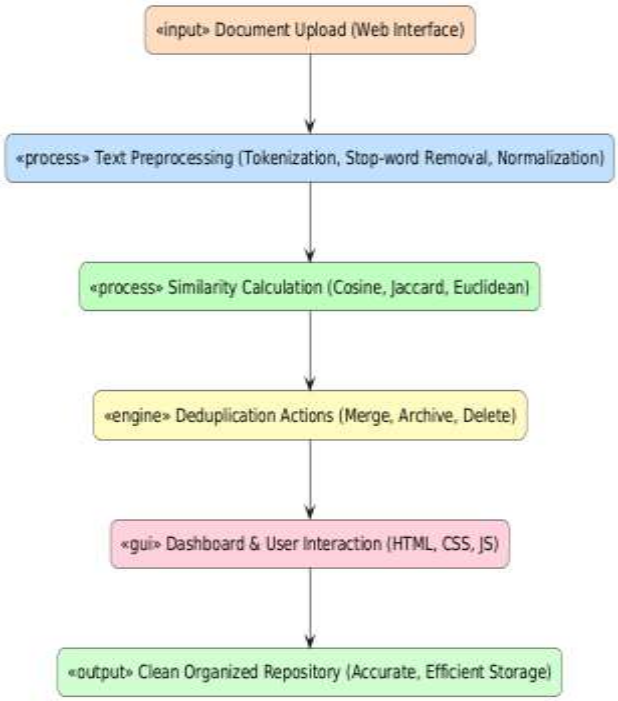

Figure 1 Methodology

5.1 System Architecture

System architecture is the whole architecture of a system's components and how those components behave, and interact with one another, whether the system is a software system, a computer system, or a complex system of systems. System architecture provides a high-level perspective on the structure of the system, and how the components of the system, such as hardware, software, data storage, processing unit, communication protocols, and user interfaces, communicate, work together, and performs functions. In a software system, the architecture explains how the modules or services are separated, and how they communicate (with APIs, message queues, etc.) as well as how the data goes from place to place in the system. In a hardware system, the architecture encompasses how the processors, memory units, input and output devices are designed, connected, and interoperate with one another. System architecture also considers scalability (how to support growth in number of users or amount of data), security (i.e. protection of data and operations), maintainability (i.e. how difficult it is to push an update or fix as a result of a bug), and performance (e.g. responsiveness and speed).

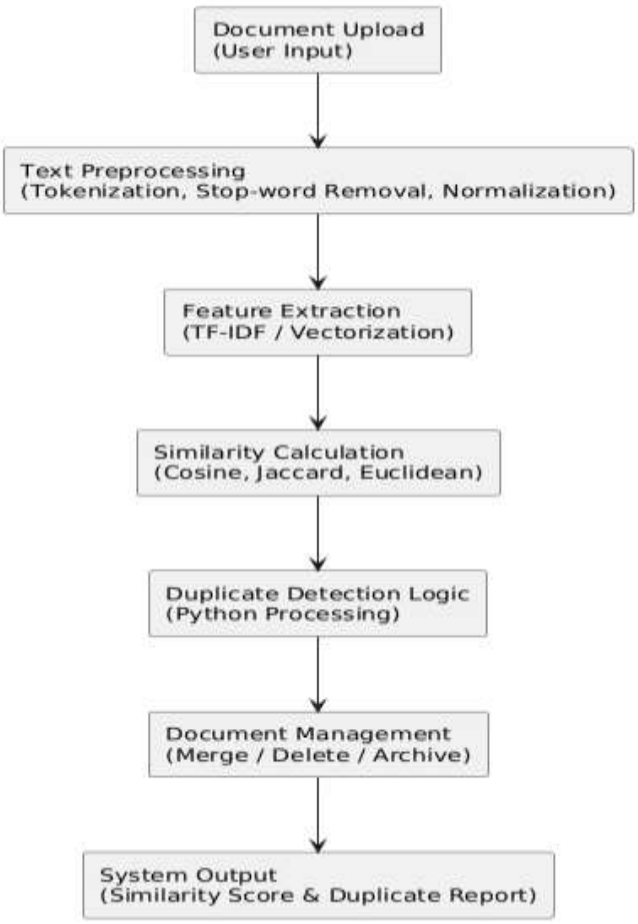

Figure 2 System Architecture

5.1.1 Input Acquisition and User Interaction Layer

The user interface layer is the main point at which a user will be able to access the system. Any document that users wish to upload can be accomplished via an HTML, CSS, and JavaScript web-based interface, making the interface easy to use and accessible; the user interface also validates the format of the input, supports drag & drop functionality, and provides feedback (i.e., progress bar or error messages). Through the interactive dash-board, a user can view all similarity scores and manage duplicate documents easily; thus, providing a mechanism for user interface and back-end communication and usability.

5.1.2 Data Collection and Parsing Layer

The layer is responsible for collecting all the submitted documents and converting them into formats suitable for analysis, as well as extracting unstructured text from numerous file types, defining structure to the extracted content. The layer uses NLP preprocessing methods such as tokenizing, removing stop words, and normalizing the data for it to be consistent across all documents and to provide a common base on which to measure simi-larity among documents accurately. In effect, this layer provides a basis for processing through its conversion of unstructured content into meaningful data.

5.1.3 AI Powered Summarization Layer

- Uses NLP models to generate concise summaries of documents.
- Identifies key sentences and important themes for quick understanding.
- Reduces redundancy by highlighting unique content.
- Improves efficiency in reviewing large document sets.
- Provides semantic-level insights beyond basic text matching.

5.1.4 API integration and Communication Layer

This level provides seamless connections for all parts of your system and outside services. The API will allow information to move between the web interface, processing modules in the backend, and storage systems to allow for document uploads, calculating similarity, and performing deduplication to all occur in real-time. By establishing standard protocols, this layer improves interoperability and scalability of services provided by this layer.

The key points for this layer include:

1. Uses APIs to facilitate the transfer of information between the different modules within the document process.
2. Provides for real-time communication between all entities performing the above document process; thus, allowing for faster processing.
3. Provides for scalability and the ability to connect to greater/extensive outside services than those traditionally considered.

5.1.5 Security and Access Control Layer

- Implements user authentication through login credentials.
- Provides role-based access to restrict sensitive actions.
- Encrypts data during upload and storage for safety.
- Monitors system activity to detect unauthorized access.
- Ensures compliance with data privacy and security standards.

5.1.6 Output Visualization and Download Layer

The similarity analysis and deduplication results are presented in a format that is easy for the user to understand (an executive-level overview). The dashboard contains: similarity scores; summary statistics regarding the number of duplicates; options for managing duplicate documents; and the ability to view data trends using charts/graphs. Users can also download a cleaned-up version of the dataset for their further use. This layer will enhance the efficiency of making decisions by providing clarity and being interactive.

Core elements of this layer include:

1. Show similarity scores and summary statistics of how many duplicates exist.
2. Provide graphical or chart-based data visualization.
3. Enable the user to download organized data sets that have been cleaned up.

5.1.7 Testing and Evaluation

The final stage of application of a system is to validate the system's functioning as well as its accuracy. The validation process consists of testing the preprocessing, measure of similarity, and deduplication engine with various different data sets. Evaluation metrics (e.g., precision, recall, efficiency) are used to measure the effectiveness of the system and its components with respect to the system's requirements. Continuous testing helps to establish.

1. How well an application works
2. How efficiently the application works with a variety of different types of data, and
3. How reliable and scalable the application is when utilized in the real world.

6. Results

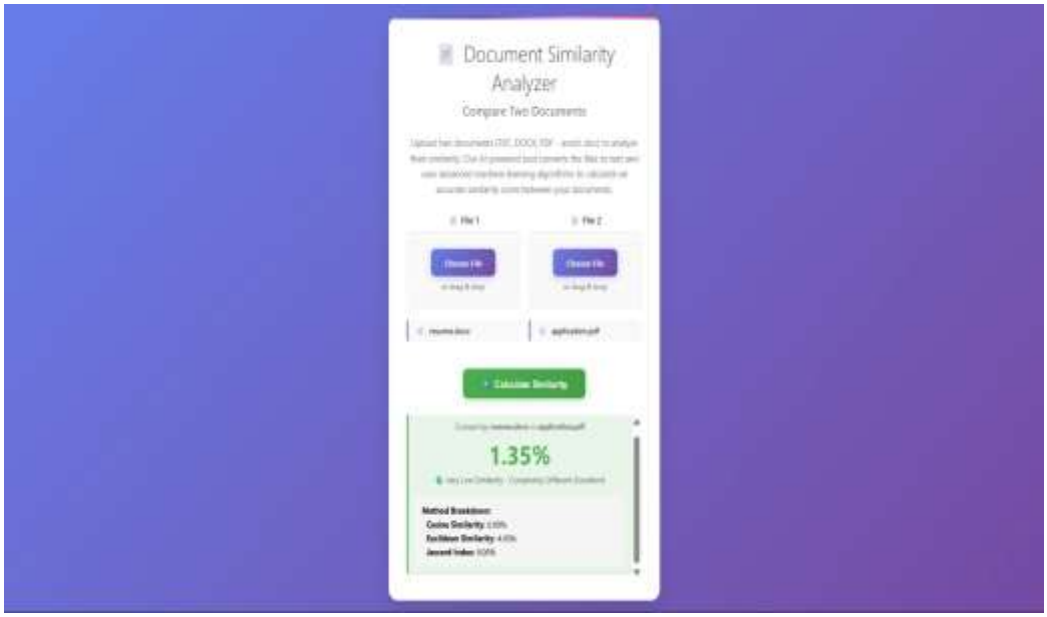

Figure 3

Figure 4

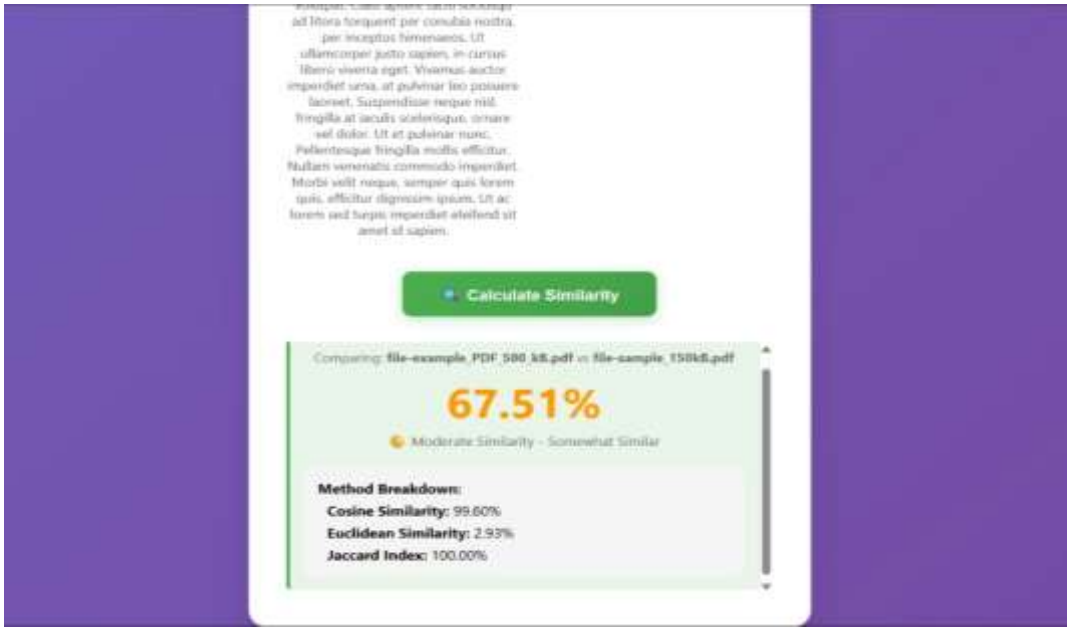

Figure 5

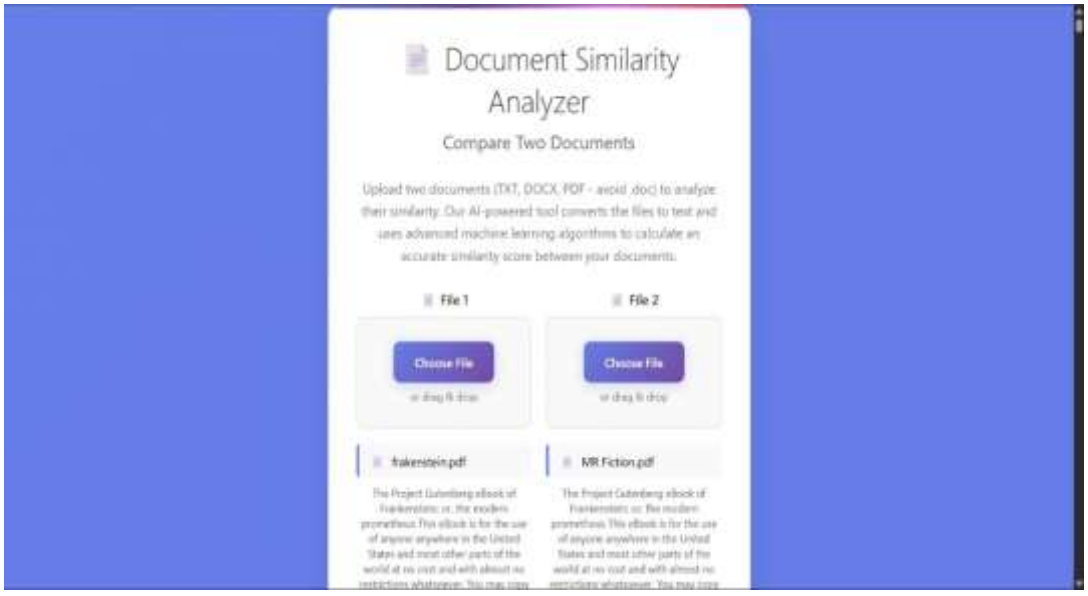

Figure 6

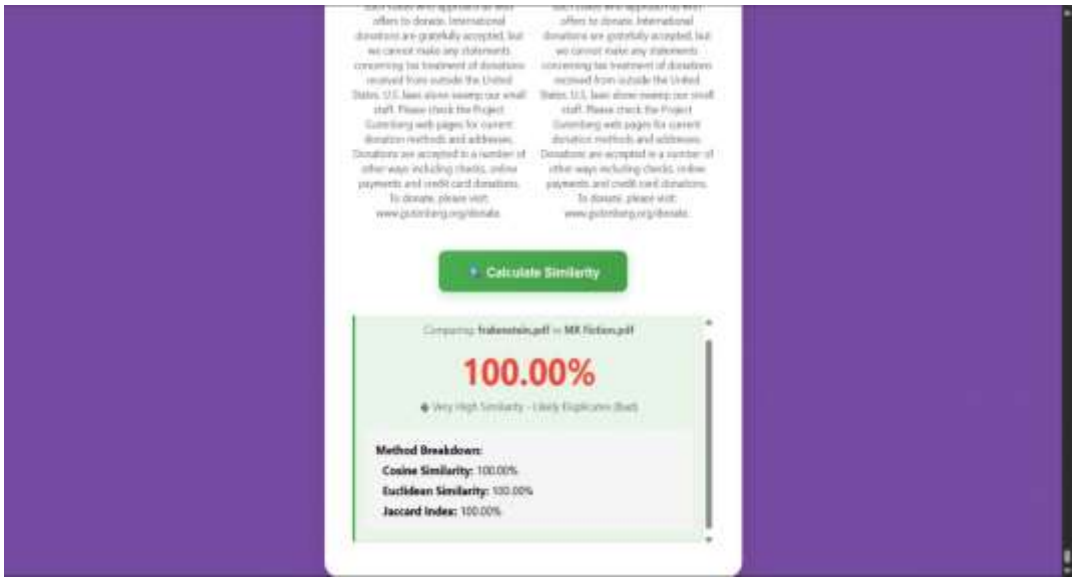

Figure 7

Conclusion

Document Similarity and Deduplication Tools (DSDD) provide data redundancy management solutions. Using Natural Language Processing methods (e.g., Cosine, Jaccard, Euclidean distance), DSDD tools are used to look through documents to determine similarity in content. The DSDD tools architecture is designed for easy use and comprehension of users. The network-based user interface for the DSDD is user-friendly and highly interactive. The DSDD Tools resolve all issues with existing DSDD systems. By enabling efficient storage of data and easing retrieval of the correct data, DSDD tools reduce the amount of manual processing required for duplicate detection. Improvements can be made in the future to DSDD tools. DSDD tools enable robust datasets, resource optimization, and effective document management through DSDD use.

References

- A Comparative Study of TF-IDF & Cosine Similarity for Document Matching - S. Patel et al (2023)
- Shingling Algorithm for Near-Duplicate Detection - Andrei Broder et al.(2022)
- Data De-duplication on Similar File Detection - Xingjun Zhang, Runting Zhao (2021)
- Near-Duplicate Detection in Web App Model Inference - Rahulkrishna Yandrapally et al.(2020)
- Efficient Similarity Joins for Near-Duplicate Detection - Chuan Xiao et al.(2020)