In our specialized essay, we will study a new part of data management. The essential thing that makes this essay is the bag of word model combined with the tf-idf algorithm to compare the similarities of the two documents.. Both file's that are passed are to be of file type and not directory. length in coherently according to each specific topic bytes should not be the same. From the above knowledge will help us create a data management application for a medium-sized company. Both are different files and not one and the same.

In the current 4.0 era, with the increasing demand for document storage, the question is how can we store documents effectively and in a long term? Since then to serve our document storage needs clearly and coherently according to each specific topic. Then compare the contents. especially for a large number of users in a long way like in companies. , schools, digital libraries, ... There is also a concern that is the duplication of data when storing such a large amount of documents. The size can differ, also for same content. Depending on several factors.from the above needs, our team would like to consider thought of the project "Building a document management system and checking duplicate documents on a common data warehouse".If you realy want to compare the content, then an easy check is to make a checksum of both files and compare them.

Create a website system that can store a large amount of data in a long term, which can be accessed and used by a quite large number of people. You can use md5 on the bytearray of the files. Also an compare of the bytearrays can be used.helps search documents by topic, manage documents. and easy and effective duplication checking.

Learn about Mongodb and how to information retrieval and text mining transform data through mongoose. I'm curious under what circumstances two files with the same byte content would have different

The bag-of-words model is a simplified representation used in natural language processing and information retrieval.

The magnetic pocket model is often used in document classification methods in which the occurrence (frequency) of each word is used as a characteristic to train classifiers.

TF-IDF (short for term frequency - inverse document frequency) is a statistical method commonly used in information retrieval and text mining to evaluate the level of the importance of a phrase to a particular document in a set that includes many documents.

For example, TF-IDF has long been a part of Google's ranking mechanism. Google uses TF-IDF to determine which terms are topically relevant (or irrelevant) by analyzing how often a phrase appears average frequency of usage for on a page (term frequency - TF) and how often. The estimate occurs per page on average, within a larger set of inverse document frequency (IDF).

This is a necessary condition for storing the same content. But then I'd like to listen to your approaches. If the two files are stored on the same hard drive (like in most of my cases) it's probably not the best way to jump too many times.

Because most online content is text, most of these attributes may be the presence or absence of certain words or phrases on the page. And this is when the TF-IDF algorithm becomes necessary. It will calculate the average frequency of usage for a specific phrase across the website as well as set a benchmark for stop words (which are considered too common. In addition, the system also considers the prominence of those words on the page compared to other pages on the website.

where the quantity TF represents the number of occurrences of a phrase in a certain document, while the quantity IDF will be a quantity calculated by the log function. All computations will be performed by computers. However, you should understand that the TF-IDF value is not based solely on keyword density. Here's the formula for these metrics:

When you search the Internet, you have an account to log into the website can see that there are many different TF formulas, but each variation is built on the basis: The more a word appears in a file, the more correlation there is. relation, and the contribution of TF to the relevance of a document is essentially a sub-linear function.[2].

Express is a popular unopinionated web framework, written in JavaScript and hosted within the Node.js runtime environment. This module explains some of the key benefits of the framework, how to set up your development environment and how to perform common web development and deployment tasks.[6]

MongoDB is a document-oriented NoSQL database used for high volume data storage. Instead of using tables and rows as in the traditional relational databases, MongoDB makes use of collections and documents. Documents consist of key-value pairs which are the basic unit of data in MongoDB. Collections on the website's data warehouse contain sets of documents and function which is the equivalent of relational database tables. MongoDB is a database which came into light around the mid-2000s.[7]

A data management website for a small company, to be able to use each employee must have an account to log into the website, this account will be registered by the admin until that employee comes to work. company.

For each account that can log into the website, users can view documents, download documents on the website's data warehouse. In addition, users can also upload documents to a data warehouse for archiving, especially when uploading documents, the system will check if the document is already in the common data warehouse to avoid the fields. merge upload too many identical or similar documents. In addition, users can create themes to explicitly section document types.

As for the admin account, there are functions of registering new accounts for new employees in the company and modifying passwords of old accounts according to employees' requirements.

Users must authenticate to the system and use according to their authority

To be able to use the system, the user must log in to the system with the account issued by the company when working. Each account will have certain features.

Users can view documents by topic documents in the system or add, edit, and delete topics and upload document files according to those topics. The system modifying passwords of will automatically detect 2 duplicate document files and notify the user. In addition, users can search for documents in the system.

Users must authenticate to the system and use according to their authority

To be able to use the system, the user must log in to the system with the account issued by the company when working. Each account will have certain features.

Users can view documents by topic or add, edit, and delete topics and upload document files according to those topics. The system will automatically detect 2 duplicate document files and notify the user. In addition, users can search for documents in the system.