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. From the above knowledge will help us create a data management application for a medium-sized company.

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, 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, 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".

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, helps search documents by topic, manage documents. and easy and effective duplication checking.

- 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 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).

TF-IDF là một phương pháp dùng để đánh giá độ quan trọng của một từ hoặc cụm từ trong tập văn bản

To determine how relevant a particular page is, Google will analyze the pages contained in its indexed list based on some specific attributes it considers relevant to the query. .

Because most online content is text, most of these attributes may be the presence or absence of certain words or phrases on the page. In addition, the system also considers the prominence of those words on the page compared to other pages on the website.

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, too general. and does not have a specific meaning if standing alone) to yield a more accurate result.

First, we need to know that the index TF-IDF can be calculated using the following formula: TF-IDF = TF x IDF

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:

TF (Term Frequency): We denote TF for the frequency of occurrence of term t. The term t here can be interpreted as an n-grams token in a document d:

Quotient of the number of occurrences of 1 word in the text and the number of words in that text in the text

IDF ((Inverse Document Freequency): Calculate IDF to decrease the value of common words. To avoid words appearing multiple times without meaning

When you search the Internet, you 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]

1.1.1. Introduction Euclid algorithm

In mathematics, the Euclidean distance is the "normal" distance between two points that can be measured with a ruler, and is calculated using the Pythagorean formula. [1] Using this formula to calculate distances, an Euclidean space (or even any scalar space) becomes a metric space.

The Euclidean distance between points p and q is the length of the segment pq. In the Cartesian coordinate system, if p = (p1, p2, ..., pn) and q = (q1, q2, ..., qn) are two points in an Euclidean n-dimensional space, then the distance from p to q is equal to:

Example: Given two points A (1, 1) and B (2, 2) on a 2-dimensional plane. Based on the Pythagorean formula, the Euclidean distance between points A and B is:

How to calculate the distance between two vectors not in the same number of dimensions:

- Euclidean distance formula finds the distance between any two points in Euclidean space.
- A point in Euclidean space is also called an Euclidean vector.
- You can use the Euclidean distance formula to calculate distances between vectors of two different lengths.
- For vectors of different sizes, the same principle applies.
- Assume a vector of a lower dimension also exists in the higher dimensional space. You can then set all the missing elements in the lower dimensional vectors to 0 so that both vectors have the same direction. You will then use any of the mentioned distance formulas to calculate the distance.
- Calculate the distance between A and B?
- To represent A in R³, you would set its components to (a1, a2,0). Then it is possible to find the Euclidean distance d between A and B using the formula:

1.1.2. ReactJS framework

ReactJS được hiểu nôm na là một thư viện trong đó có chứa nhiều JavaScript mã nguồn mở và cha đẻ của ReactJS đó chính là một ông lớn với cái tên ai cũng biết đó chính là Facebook. Mục đích của việc tạo ra ReactJS là để tạo ra những ứng dụng website hấp dẫn với tốc độ nhanh và hiệu quả cao với những mã hóa tối thiểu. Và mục đích chủ chốt của ReactJS đó chính là mỗi website khi đã sử dụng ReactJS thì phải chạy thật mượt thật nhanh và có khả năng mở rộng cao và đơn giản thực hiện.

Nhìn chung tất cả những tính năng hay sức mạnh của ReactJS thường xuất phát từ việc tập trung vào các phần riêng lẻ chính vì điểm này nên khi làm việc trên web thay vì nó sẽ làm việc trên toàn bộ ứng dụng của website thì ReactJS cho phép developer có chức năng phá vỡ giao diện của người dùng từ một cách phức tạp và biến nó trở thành các phần đơn giản hơn nhiều lần có nghĩa là render dữ liệu không chỉ được thực hiện ở vị trí sever mà còn có thể thực hiện ở vị trí Client khi sử dụng ReactJS.[5]

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 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.

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.