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.

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.

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.

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, too general.

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.

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.

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.

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.