# Algorithm details

## Algorithm intro

TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical statistic used in information retrieval and text mining to assess the importance of a term within a document or a collection of documents. It is a widely used algorithm for text analysis and plays a crucial role in various natural language processing tasks. TF-IDF is helpful for Page Rank, the ranking algorithm used by search engines. It assesses the relevance of web pages to search queries and guides keyword optimization. TF-IDF helps differentiate content, assess content quality, and improve the accuracy of Page Rank assessments.

## WORKING

The TF-IDF algorithm evaluates the importance of a term in a document or collection of documents. It works by calculating the term frequency (TF), which measures how often a term given by the user in query appears in a document, and the inverse document frequency (IDF), which penalizes terms that occur in many documents. The TF-IDF score for a term is obtained by multiplying its TF by its IDF. This algorithm helps identify important terms that are frequent in a document but relatively rare in the corpus. It is used in various text analysis tasks such as document similarity, keyword extraction, and text classification.

By using TF-IDF, SEO practitioners can identify the query words that are both frequently occurring in a document (TF) and relatively rare across the collection of documents (IDF). This helps target and optimize content to align with the specific query words, improving the chances of ranking well in search engine results for relevant queries.

## Why there is need?

TF-IDF is essential in the field of SEO because it enables search engines to understand the relevance and importance of specific terms within web pages. By analyzing term frequency and inverse document frequency, search engines can determine the relevance of a page to a given search query. TF-IDF helps optimize website content by identifying important keywords and incorporating them strategically. It also aids in detecting duplicate content and analyzing competitors' keyword usage. Overall, TF-IDF enhances search result relevance, improves content optimization, and provides insights for effective SEO strategies.

# DESIGN OF ALGORITHM

The design of the TF-IDF algorithm can be summarized in the following steps:

## Corpus Preparation:

Gather a collection of documents (corpus) and preprocess the text. This involve tasks like tokenization, removing stop words, and applying stemming or lemmatization and spell check of the

## Term Frequency (TF) Calculation:

For each document, calculate the term frequency of each term by counting its occurrences within the document. Normalize the count by dividing it by the total number of terms in the document to obtain the TF value for each term.

## Inverse Document Frequency (IDF) Calculation:

For each term in the corpus, calculate the inverse document frequency. Divide the total number of documents in the corpus by the number of documents that contain the term. Apply logarithmic scaling to dampen the effect of very high IDF values.

## TF-IDF Calculation:

Multiply the TF value of each term in a document by its IDF value to obtain the TF-IDF score for that term in the document.

## Ranking:

Use the TF-IDF scores to rank documents based on their relevance to specific queries or tasks. Higher TF-IDF scores indicate greater relevance and importance of terms within the document. It's important to note that while this represents the general design of the TF-IDF algorithm, there may be variations and additional considerations depending on specific implementations or use cases.

## ANALYSIS OF TF\_IDF INCLUDING ITS IMPLEMENTATION IN SEO:

The time and space complexity of the TF-IDF algorithm can vary depending on the implementation and the size of the corpus. Here is a general analysis of the complexity:

## Time Complexity:

### Corpus Preparation:

Preprocessing the corpus involves additional SEO-related steps, such as keyword extraction, spell check, highlighting the query syntax in content and content analysis. The time complexity depends on the complexity of these steps and the size of the corpus.

### Term Frequency (TF) Calculation:

Calculating TF for each document involves counting the occurrences of keywords. The time complexity is influenced by the number of documents and the average number of keywords per document.

### Inverse Document Frequency (IDF) Calculation:

Calculating IDF includes analyzing the occurrence of keywords across documents, which impacts search engine relevance calculations. The time complexity depends on the number of documents and the number of unique keywords.

### TF-IDF Calculation:

Multiplying TF by IDF to obtain TF-IDF scores has a time complexity of **O (1)** for each term in a document. The time complexity analysis in the context of SEO implementation remains similar to the general TF-IDF algorithm, with considerations for additional SEO-related steps.

## Space Complexity:

### Corpus Storage:

Storing the corpus involves considerations for SEO-specific data, such as metadata, URLs, and indexed information. The space complexity remains proportional to the number of terms in the corpus.

### Term Frequency (TF) Storage:

Storing TF values for each document requires additional space, considering the presence of SEO-specific data and keyword indices.

### Inverse Document Frequency (IDF) Storage:

Storing IDF values requires space proportional to the number of unique keywords and their occurrence across documents.

TF-IDF Storage:

Storing TF-IDF scores for each term in a document adds to the space complexity, considering the need for storing relevance-related information. The space complexity analysis includes considerations for additional SEO-related data storage and indexing requirements. In summary, the implementation of TF-IDF in SEO involves additional considerations for time and space complexity, primarily driven by SEO-specific tasks like corpus preparation, keyword extraction, and relevance calculations. The complexity analysis expands to accommodate the SEO-specific steps, but the core time and space complexity characteristics of the TF-IDF algorithm remain intact.