TF-IDF được tính bằng tích số của hai giá trị: Term Frequency (TF) và Inverse Document Frequency (IDF).

Term Frequency (TF): là tần suất xuất hiện của một từ “t” trong một tài liệu “d”

Công thức:

Inverse Document Frequency (IDF): đo lường tầm quan trọng của một từ dựa trên tần suất xuất hiện của nó trong tập hợp các tài liệu. Một cách dễ hiểu hơn IDF đo lường mức độ phổ biến của một từ để xem xét từ đó có giá trị đặc biệt hay không.

Công thức:

1 được cộng vào để tránh việc chia cho 0 khi từ đó không xuất hiện trong bất kỳ tài liệu nào.

Khi IDF thấp, nghĩa là từ này xuất hiện ở rất nhiều tài liệu, có thể từ này không mang lại nhiều giá trị trong văn bản vì nó xuất hiện quá phổ biến. Ví dụ như những từ nối, giới từ, đại từ chỉ định, đây là những từ không quan trọng

Khi IDF cao, nghĩa là từ này ít khi xuất hiện trong tập các tài liệu, có khả năng mang lại các thông tin quan trọng và đặc trưng, giúp cho việc phân loại tài liệu hiệu quả hơn.