TF-IDF

Định nghĩa

1. tfidf là tích của hai thống kê, term frequency và inverse document frequency. Có nhiều cách xác định giá trị của hai thống kê này.
2. tfidf là công thức xác định tầm quan trọng của từ khóa hoặc cụm từ đối với một văn bản trong kho văn bản.

Công thức

- $tf(t, d)$ là số lần từ t xuất hiện trong văn bản d.
- $idf(t, D) = log(N/n_t)$ đo lượng thông tin từ đó cung cấp.

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D)$$

Ví dụ

Document 1: "this is a sample"

Document 2: "this is another example"