ESCUELA MILITAR DE INGENIERÍA

MCAL. ANTONIO JOSÉ DE SUCRE

“BOLIVIA”

“TF-IDF”

ESTUDIANTE: RODRIGO HAROLD MENDEZ PRADO

DOCENTE: ING. DIEGO CLAROS

MATERIA: TALLER DE TICS Y SOFTWARE

SEMESTRE: NOVENO SEMESTRE

CARRERA: INGENIERIA DE SISTEMAS

FECHA: 18/08/2024

TF-IDF (Term Frequency-Inverse Document Frequency) es una técnica utilizada en procesamiento de texto para evaluar la relevancia de una palabra en un documento dentro de un conjunto de documentos (corpus). A continuación, te guiaré a través del proceso de aplicar TF-IDF a un conjunto de documentos.

1. Selección de Documentos

Supongamos que elegimos los siguientes extractos para trabajar:

Documento 1: Un artículo sobre avances tecnológicos en inteligencia artificial.

Documento 2: Un blog sobre la historia de la arquitectura gótica.

Documento 3: Un extracto de una revista sobre recetas de cocina italiana.

Aquí están los textos:

Documento 1: "La inteligencia artificial ha transformado la tecnología moderna, permitiendo avances sin precedentes en diversas industrias."

Documento 2: "La arquitectura gótica se caracteriza por sus altos arcos y su enfoque en la altura y la luz, algo revolucionario en su tiempo."

Documento 3: "La cocina italiana es famosa por su simplicidad, centrada en ingredientes frescos y de alta calidad."

2. Identificación de Palabras Clave

Para cada documento, seleccionamos algunas palabras clave:

Documento 1: "inteligencia", "tecnología", "avances"

Documento 2: "arquitectura", "gótica", "altura"

Documento 3: "cocina", "italiana", "simplicidad"

3. Cálculo de Frecuencia de Término (TF)

El TF se calcula como la frecuencia de una palabra clave en un documento dividida por el número total de palabras en ese documento.

Documento 1:

Total de palabras: 13

TF("inteligencia") = 1/13 ≈ 0.077

TF("tecnología") = 1/13 ≈ 0.077

TF("avances") = 1/13 ≈ 0.077

Documento 2:

Total de palabras: 19

TF("arquitectura") = 1/19 ≈ 0.053

TF("gótica") = 1/19 ≈ 0.053

TF("altura") = 1/19 ≈ 0.053

Documento 3:

Total de palabras: 13

TF("cocina") = 1/13 ≈ 0.077

TF("italiana") = 1/13 ≈ 0.077

TF("simplicidad") = 1/13 ≈ 0.077

4. Cálculo de Frecuencia Inversa de Documento (IDF)

El IDF se calcula como el logaritmo del número total de documentos dividido por el número de documentos que contienen la palabra clave.

IDF = log(N / df), donde N es el número total de documentos, y df es el número de documentos que contienen la palabra clave.

Para las palabras clave seleccionadas:

IDF("inteligencia", "tecnología", "avances") = log(3/1) ≈ 0.477 (cada palabra aparece en 1 documento).

IDF("arquitectura", "gótica", "altura") = log(3/1) ≈ 0.477.

IDF("cocina", "italiana", "simplicidad") = log(3/1) ≈ 0.477.

5. Aplicación de TF-IDF

Finalmente, combinamos TF e IDF para obtener el valor TF-IDF:

Documento 1:

TF-IDF("inteligencia") = 0.077 * 0.477 ≈ 0.037

TF-IDF("tecnología") = 0.077 * 0.477 ≈ 0.037

TF-IDF("avances") = 0.077 * 0.477 ≈ 0.037

Documento 2:

TF-IDF("arquitectura") = 0.053 * 0.477 ≈ 0.025

TF-IDF("gótica") = 0.053 * 0.477 ≈ 0.025

TF-IDF("altura") = 0.053 * 0.477 ≈ 0.025

Documento 3:

TF-IDF("cocina") = 0.077 * 0.477 ≈ 0.037

TF-IDF("italiana") = 0.077 * 0.477 ≈ 0.037

TF-IDF("simplicidad") = 0.077 * 0.477 ≈ 0.037

Interpretación

Los valores más altos de TF-IDF indican que esas palabras son importantes y características en su respectivo documento, mientras que los valores más bajos podrían señalar palabras menos distintivas. Por ejemplo, las palabras "tecnología" y "inteligencia" en el Documento 1 tienen valores más altos en comparación con su presencia en otros documentos, lo que subraya su importancia en ese contexto específico.