Universidad de La Laguna

Sistema de recomendación con Modelos basados en el contenido.

Análisis de resultados obtenidos

Carlos Ravina Morales
Javier Yanes de León

Resultados y análisis TF-IDF

Metodología

El análisis TF-IDF se realizó con el binario ./bin/tfidf, utilizando listas de stopwords y lemmas para normalizar el texto.Se calcularon las métricas TF, IDF, TF-IDF y la similitud coseno entre documentos.

- TF mide la frecuencia de un término en un documento.
- IDF refleja cuán raro es el término en el corpus.
- TF-IDF destaca palabras frecuentes en un documento pero infrecuentes en el conjunto total.
- La similitud coseno (0–1) evalúa la semejanza entre documentos según su vocabulario.

Fichero concretos_1_2_3.txt

Las similitudes son bajas (<0.08), lo que indica que los tres documentos tratan temas distintos con poco solapamiento léxico.

Contenido: análisis de tres documentos (document-01.txt, document-02.txt, document-03.txt).

Observaciones:

- Términos comunes (como a, the, and) tienen IDF = 0 ⇒ TF-IDF = 0.
- Términos únicos de un documento presentan IDF ≈ 1.0986 (caso de N = 3, df = 1).

Ejemplo:

you (doc-01): TF = 4, IDF = 1.0986 → TF-IDF ≈ 4.39

afraid (doc-01): TF = 2 → TF-IDF ≈ 2.19

Matriz de similitud coseno (3×3):

Par de documentos	Similitud
doc-01 vs doc-02	0.0633
doc-01 vs doc-03	0.0466

doc-02 vs doc-03 0.0731

Fichero new_docs.txt

Contenido: resultados para 10 documentos (doc01.txt … doc10.txt).

Observaciones:

TF-IDF identifica correctamente los términos temáticos de cada texto:

Ej.: gato (doc01) → TF = 2, IDF ≈ 2.3026 → TF-IDF ≈ 4.6052.

Algunos conectores (la, el, y) conservan IDF bajo pero no nulo (≈0.35), lo que indica una lista de stopwords incompleta.

Se observan tokens truncados o sin acentos (p. ej. cnicas, salud, sica), evidencia de una tokenización simple.

Matriz de similitud coseno (extracto):

- Valor máximo: 0.1499 (entre doc01 y doc04).
- Otros pares con valores medios: 0.127 (doc07–doc08), 0.064 (doc01–doc06).
- La mayoría de pares <0.05.

Los documentos son en general disímiles, con pequeñas coincidencias de vocabulario entre algunos pares.

Los términos con TF-IDF alto representan adecuadamente el tema de cada texto, aunque algunas palabras funcionales persisten por un filtrado de stopwords incompleto.

Las similitudes entre documentos son en general muy bajas (<0.05), con pocos pares moderadamente similares (≈0.15).

Esto confirma que los textos son mayoritariamente independientes y con vocabulario diferenciado.

Fichero salida_ejs.txt

Contenido: análisis completo del corpus ejemplos (document-01..10 y otros).

Observaciones:

- Algunos términos funcionales (the, a, i, to) presentan TF-IDF elevados (p. ej. the TF=87, IDF≈0.26 → TF-IDF≈22.8).
- Esto indica que el fichero de stopwords no se aplicó correctamente o no contenía todas las formas.
- Palabras raras (IDF ≈ 2.56) mantienen TF-IDF alto → correcto funcionamiento para términos específicos.
- Las matrices de similitud muestran valores moderados/bajos, coherentes con documentos diversos.

Aparecen palabras muy frecuentes (the, a, i) con TF-IDF alto, lo que sugiere que el fichero de stopwords no se aplicó correctamente o no incluía todas las formas.

Aun así, los términos menos comunes (IDF alto) se identifican correctamente como característicos de cada documento.

Las similitudes entre textos son bajas, coherentes con un conjunto heterogéneo de temas.

Conclusiones generales

El modelo TF-IDF funciona correctamente para identificar términos distintivos cuando se aplica una normalización adecuada.

La calidad de los resultados depende del preprocesamiento: es esencial usar una lista de stopwords completa y una tokenización robusta.

Las matrices de similitud coseno muestran poca relación entre documentos, lo que indica diversidad temática y baja redundancia.