Manual de algoritmos de similitud (Req2)

Descripción

Este documento resume la base matemática, relación con el análisis bibliométrico e interpretación de los valores (0–1) de los algoritmos implementados para comparar pares de textos (títulos/abstracts).

Rango de valores

En todas las métricas: 0.0 indica sin similitud y 1.0 indica textos idénticos. Valores intermedios reflejan similitud parcial.

Levenshtein (similitud)

Fórmula

sim = 1 - (distancia_levenshtein(a,b) / max(len(a), len(b))). Mide el mínimo número de ediciones (inserción/eliminación/sustitución) para transformar a→b.

Relación con el análisis

Útil para detectar versiones casi idénticas de títulos o pequeñas variaciones tipográficas. Sensible a longitud.

Interpretación

Cercano a 1: textos casi iguales. Cercano a 0: textos muy diferentes.

Jaccard (palabras)

Fórmula

J(A,B) = |A ∩ B| / |A ∪ B|, donde A y B son conjuntos de palabras normalizadas.

Relación con el análisis

Capta solapamiento de vocabulario clave en títulos/abstracts. No considera frecuencia ni orden.

Interpretación

0: vocabulario disjunto. 1: mismo conjunto de palabras.

Dice (palabras)

Fórmula

Dice(A,B) = 2|A ∩ B| / (|A| + |B|). Métrica similar a Jaccard pero con diferente ponderación.

Relación con el análisis

Comparación de vocabularios con énfasis en coincidencias comunes.

Interpretación

Valores más altos indican mayor solapamiento léxico.

Cosine TF-IDF

Fórmula

cos(v1,v2) = (v1·v2) / (||v1||·||v2||), donde v1 y v2 son vectores TF-IDF (palabras o n-gramas de caracteres).

Relación con el análisis

Mide similitud basada en términos discriminativos. Para textos muy cortos usamos n-gramas de caracteres (heurística).

Interpretación

Alto cuando los términos relevantes son compartidos en proporciones similares.

Jaccard (caracteres n-gram)

Fórmula

J(Cn(a), Cn(b)) = |Cn(a) ∩ Cn(b)| / |Cn(a) ∪ Cn(b)|, con Cn(x) los n-gramas de caracteres de x.

Relación con el análisis

Robusto en textos cortos o con pequeñas variaciones morfológicas/ortográficas. n se ajusta por heurística según longitud.

Interpretación

Mayor valor indica mayor superposición de n-gramas de caracteres.

Dice (caracteres n-gram)

Fórmula

$Dice(Cn(a),Cn(b)) = 2|Cn(a) \cap Cn(b)| / (|Cn(a)| + |Cn(b)|)$.

Relación con el análisis

Complementa Jaccard a nivel de caracteres; útil para títulos cortos.

Interpretación

Cercano a 1 cuando los fragmentos de caracteres coinciden fuertemente.

Semantic SBERT

Base

Embeddings de frases (Sentence-BERT). Similitud coseno entre vectores de alta dimensión que capturan semántica.

Relación con el análisis

Detecta similitud semántica más allá de palabras exactas; muy útil para abstracts.

Interpretación

>0.8 típicamente indica alta cercanía semántica; 0.5–0.8 similaridades temáticas parciales; <0.5 escasa relación.