Propuestas

Función de similitud original:
cuenta las palabras en común entre cada par de oraciones.

$$Similitud(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)}$$

Subcadena común más larga

Devuelve la longitud de la subcadena más larga presente en las dos oraciones.

Similitud coseno

Usando el modelo TF-IDF representa cada oración como un vector y computa el resultado como el coseno del ángulo que comprenden.

BM25

Implementa una variación de la famosa función BM25 usada para recuperación de información; que no permite haya resultados que tomen valores negativos.