Limitaciones del Análisis de Similitud con TF-IDF en Chatbots Educativos

Cuando se desarrolla un chatbot educativo, es común utilizar técnicas como TF-IDF (Term Frequency-Inverse Document Frequency) para calcular la similitud entre preguntas y respuestas. Aunque esta técnica es sencilla y útil, tiene limitaciones importantes que pueden llevar a resultados engañosos si se interpreta de manera aislada.

1. Frases con similitud 0 pueden ser coherentes

Una similitud de 0 en TF-IDF indica que no hay coincidencia léxica directa entre la frase y la respuesta, pero no implica necesariamente que sean incoherentes. Algunas razones incluyen:

- Uso de sinónimos o paráfrasis: TF-IDF no detecta equivalencias semánticas.
- Frases implícitas o contextuales: pueden tener sentido dentro de una conversación aunque no compartan palabras clave.
- Frases emocionales o expresivas: emojis o expresiones breves pueden ser entendidas correctamente por humanos, pero no por el algoritmo.
- TF-IDF no interpreta: solo cuenta frecuencia de palabras, sin entender intención ni contexto.

2. Por qué no es recomendable usar stemming con TF-IDF en educación

El stemming reduce las palabras a su raíz, lo cual puede eliminar información importante en contextos educativos. Ejemplo:

- 'estudiante', 'estudiar' y 'estudio' podrían ser reducidos a 'estud', lo que confunde el sentido de la frase.

Esto es problemático porque en entornos educativos las variaciones verbales, sustantivas y los matices gramaticales son relevantes para entender la intención del estudiante.

3. Limpieza recomendada para chatbots educativos

En lugar de un stemming agresivo, se recomienda lo siguiente:

- Convertir a minúsculas para uniformidad.
- Eliminar signos repetitivos o de ruido, pero conservar puntuación relevante.
- Mantener emojis: aportan valor emocional y ayudan a entender la intención.
- Normalizar abreviaturas estudiantiles: 'tqm' → 'te quiero mucho'.
- Lematización (opcional): puede ser útil si se usan modelos semánticos más avanzados como

embeddings.
- No eliminar palabras funcionales ('cómo', 'dónde', 'cuándo') si la intención depende de ellas.

4. Ejemplos de frases con similitud 0 pero respuestas coherentes

- Frase: Ya me rajé 😢

Respuesta: No te preocupes, puedes recuperar con el examen final.

Similitud estimada: 0.0

Comentario: Aunque no comparten ninguna palabra, la respuesta es totalmente coherente con la intención emocional de la frase.

- Frase: ¿A qué hora es?

Respuesta: La charla empieza a las 9 am.

Similitud estimada: 0.0

Comentario: La frase es corta y ambigua, pero en el contexto adecuado, la respuesta es clara y útil.

- Frase: 😢 perdí

Respuesta: Hay otra oportunidad para presentar, no te desanimes.

Similitud estimada: 0.0

Comentario: El emoji comunica la intención. Aunque no hay coincidencia textual, la respuesta es empática y adecuada.

- Frase: ¿Se puede cancelar la materia?

Respuesta: Sí, puedes cancelarla antes de la semana 10 en Secretaría Académica.

Similitud estimada: 0.25 (baja)

Comentario: Aunque la estructura no coincide completamente, la respuesta es perfectamente pertinente.

5. Fórmula para calcular la similitud con TF-IDF y Coseno

La similitud entre dos textos (por ejemplo, una frase y una respuesta) puede calcularse utilizando la **similitud del coseno** sobre los vectores generados por TF-IDF.

$$\text{Similitud coseno} = \cos(\theta) = \frac{\vec{A} \cdot \vec{B}}{||\vec{A}|| \times ||\vec{B}||}$$

Donde:

- $\vec{A}$ es el vector TF-IDF de la primera frase.
- $\vec{B}$ es el vector TF-IDF de la segunda frase.
- $\vec{A} \cdot \vec{B}$ es el **producto punto** entre los dos vectores.
- $||\vec{A}||$ es la **norma** (o magnitud) del vector A.
- $||\vec{B}||$ es la **norma** (o magnitud) del vector B.

- El resultado está entre 0 (sin relación) y 1 (idénticos).

Esta métrica mide el **ángulo** entre los vectores. Si el ángulo es 0 (vectores iguales), la similitud es 1. Si los vectores son ortogonales (completamente diferentes), la similitud es 0.

En palabras simples:

- **Numerador**: mide qué tan alineados están los vectores (qué tan similares son).
- **Denominador**: normaliza por el tamaño de los vectores (para que la similitud no dependa de qué tan largos son los textos).

El valor de la similitud estará siempre entre **-1** y **1**, pero como los vectores TF-IDF no tienen valores negativos (porque TF-IDF siempre es ≥0), **la similitud coseno estará entre 0 y 1**:

- **1** → los textos son *idénticos* en términos de TF-IDF.
- **0** → los textos son *completamente diferentes*.

Un pequeño ejemplo numérico:

Supongamos que después de calcular el TF-IDF, tienes:

$$\vec{A} = (0.3, 0.5, 0.2)$$

$$\vec{B} = (0.4, 0.4, 0.1)$$

Entonces:

1. **Producto punto**:

$$\vec{A} \cdot \vec{B} = (0.3)(0.4) + (0.5)(0.4) + (0.2)(0.1) = 0.12 + 0.20 + 0.02 = 0.34$$

2. **Norma de A**:

$$||\vec{A}|| = \sqrt{(0.3)^2 + (0.5)^2 + (0.2)^2} = \sqrt{0.09 + 0.25 + 0.04} = \sqrt{0.38} \approx 0.616$$

3. **Norma de B**:

$$||\vec{B}|| = \sqrt{(0.4)^2 + (0.4)^2 + (0.1)^2} = \sqrt{0.16 + 0.16 + 0.01} = \sqrt{0.33} \approx 0.574$$

4. **Similitud coseno**:

$$\cos(\theta) = \frac{0.34}{0.616 \times 0.574} \approx \frac{0.34}{0.353} \approx 0.963$$

Así que la similitud coseno sería aproximadamente **0.963** → ¡muy alta! (los textos son muy similares).
