Tecnológico de Monterrey

Desarrollo de aplicaciones avanzadas

Métodos Cuantitativos

Actividad 4.4 Comparación entre técnicas de semejanza

María Fernanda Moreno Gómez A01708653

En este reporte se presentan los resultados entre las técnicas de vectorización para el análisis de similitud textual: Bag of Words (BOW), TF-IDF y cadenas de Markov. El análisis se realizó comparando un documento original con 10 textos de diferentes niveles de similitud.

Los niveles de similitud se clasificaron según los siguientes rangos:

- **Alta similitud:** Valores de coseno entre 0.85 y 1.00
- **Similitud media:** Valores de coseno entre 0.45 y 0.85
- **Baja similitud:** Valores de coseno entre 0 y 0.45

Resultados por técnica

- **Bag of Words (BOW):** 6/10 casos correctos (60% de precisión)
- **TF-IDF:** 4/10 casos correctos (40% de precisión)
- **Cadenas de Markov:** 3/10 casos correctos (30% de precisión)

La distribución de los textos analizados eran 4 con alta similitud, 3 con media similitud y 3 con baja similitud.

Nombre del archivo	Similitud esperada	BOW (coseno)	BOW correcto	TF-IDF (coseno)	TF-IDF correcto	Markov (coseno)	Markov correcto
high_00.txt	High	0.9014	✓	0.6766	✗	0.2144	✗
high_01.txt	High	0.8189	✗	0.5740	✗	0.1089	✗
high_02.txt	High	0.7578	✗	0.5139	✗	0.1366	✗
high_03.txt	High	0.6084	✗	0.3654	✗	0.0450	✗
low_01.txt	Low	0.5745	✗	0.3060	✓	0.0299	✓
low_02.txt	Low	0.3743	✓	0.2260	✓	0.0315	✓
low_03.txt	Low	0.3801	✓	0.1737	✓	0.0445	✓
moderate_01.txt	Medium	0.7274	✓	0.4511	✓	0.0697	✗
moderate_02.txt	Medium	0.6448	✓	0.3883	✗	0.0503	✗
moderate_03.txt	Medium	0.4551	✓	0.2540	✗	0.0203	✗

En la técnica de BOW se mostró un mejor rendimiento global con un 60% de precisión, resaltando que identificó correctamente todos los textos de similitud media, que detectó solamente 2 de 3 de baja similitud y detectó 1 de 4 de alta similitud.

La técnica de TF-IDF fue la segunda mejor, con una precisión del 40% en esta técnica, resaltando que detectó correctamente todos los textos de baja similitud (3 de 3), acertó 1 de los 3 de similitud media y no identificó ningún texto de alta similitud.

La técnica de Markov obtuvo la menor precisión con el 30% de los archivos, fue efectiva para detectar únicamente los textos de baja similitud y falló en todos los casos de media y alta similitud.

Por lo tanto, BOW fue la más efectiva para la tarea de clasificación de similitud textual, con un 60% de los textos. En general, los textos mejor identificados fueron los de baja similitud y los que les tocó más trabajo fue los de alta similitud (la única técnica que pudo identificar alguna de alta similitud fue BOW). Por lo tanto, BOW es la mejor técnica para medir similitudes de textos en general.