CSC 820

Natural Language Technologies

HW 11 Report

Khalid Mehtab Khan

SFSU ID: 923673423

Calculating the similarity between two sentences, documents, or contents is incredibly useful in various real-life applications. Every day, text queries help people find web pages, research papers, songs, pictures, and more. When using cosine similarity to compare feature vectors, it's crucial to consider how each term is weighted. Ensuring the most relevant results are returned is vital. In this assignment, we are exploring multiple techniques for assigning values to feature vectors, such as TF (Term Frequency) and TF-IDF (Term Frequency-Inverse Document Frequency) models. These methods can significantly impact the relevancy of each document. For our analysis, we are examining the descriptions of 10 TED talks.

Documents: 10 TED talk descriptions

Queries:

1) "the presented data dazzles how politics has changed in recent years"

2) "science fiction movies emotionally charged the audience"

3) "humanity and humor earn more than half of the population by New York Times"

Query 1 Results: the presented data dazzles how politics has changed in recent years

TF-IDF Model: [0, 0.265, 0, 0.507, 0.653, 0.252, 0.188, 0.436, 0.188, 0.265]

TF Model: [0, 0.426, 0, 0.492, 0.492, 0.405, 0.302, 0.473, 0.302, 0.426]

For the first query about recent political changes, the TF-IDF model really zeroes in on Document 5. It thinks this document's got special terms that aren't found much elsewhere, making it super relevant. The TF model, though, isn’t picky about how unique the words are; it cares more about how often words from the query show up. That's why it likes Document 4 and 5 a lot.

Query 2 Results: science fiction movies emotionally charged the audience

TF-IDF Model: [0, 0.221, 0, 0.794, 0.221, 0.221, 0, 0.469, 0, 0.221]

TF Model: [0, 0.378, 0, 0.655, 0.378, 0.378, 0, 0.458, 0, 0.378]

For Query 2, which is about the emotional impact of science fiction movies, the TF-IDF model rates document 4 the highest, likely due to unique terms specifically discussing emotions in movies. The TF model similarly values document 4, but it shows slightly higher scores for other documents like document 8, indicating a general relevance based on the frequency of common terms rather than their uniqueness.

Query 3 Results: humanity and humor earn more than half of the population by New York Times

TF-IDF Model: [0.402, 0.573, 0.696, 0.209, 0.268, 0.199, 0.193, 0.230, 0, 0.241]

TF Model: [0.378, 0.535, 0.598, 0.378, 0.436, 0.359, 0.267, 0.378, 0, 0.378]

For the third query about humanity and humor as covered by the New York Times, the TF-IDF model points to Document 3 as the top match because it likely has unique content that matches the query well. The TF model shows more even spread, suggesting a bunch of documents throw around terms from the query often.

The TF-IDF model is better at picking out documents that directly answer the query with unique content, making it great for more detailed searches. On the other hand, the TF model is good for when you want to know which documents talk about your topic a lot, even if they're not using unique language. This means the right choice between TF and TF-IDF depends on whether you need to find documents that are uniquely relevant or just broadly relevant.