Text Retrieval and Search Engines

Assignment 2

Submitters:

Ariel Suller: 324369412

Avi Ferdman: 316420132

Vector Space Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | d1 | d2 | d3 | d4 |
| a | 0 | 1 | 1 | 1 |
| b | 1 | 2 | 0 | 1 |
| c | 2 | 0 | 0 | 0 |
| d | 0 | 0 | 0 | 0 |
| e | 1 | 0 | 1 | 1 |
| f | 7 | 5 | 7 | 2 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  | tf | wf | df | Idf | qi = wf \* idf |
| a | 3 | 2.09861229 | 3 | -1.0986123 | -2.3055612 |
| b | 4 | 2.38629436 | 3 | -0.8697417 | -2.0754597 |
| c | 2 | 1.69314718 | 1 | -0.526589 | -0.8915927 |
| d | 0 | 0 | 0 | 0 | 0 |
| e | 3 | 2.09861229 | 3 | -0.7412763 | -1.5556516 |
| f | 21 | 4.04452244 | 4 | -1.3973635 | -5.651668 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | d1 |  |  |
|  | tf | wf | di |
| a | 0 | 0 | 0 |
| b | 1 | 1 | 0.27171698 |
| c | 2 | 1.69314718 | 0.46005684 |
| d | 0 | 0 | 0 |
| e | 1 | 1 | 0.27171698 |
| f | 7 | 2.94591015 | 0.80045381 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | d2 |  |  |
|  | tf | wf | di |
| a | 1 | 1 | 0.30595074 |
| b | 2 | 1.69314718 | 0.51801964 |
| c | 0 | 0 | 0 |
| d | 0 | 0 | 0 |
| e | 0 | 0 | 0 |
| f | 5 | 2.60943791 | 0.38322429 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | d3 |  |  |
|  | tf | wf | di |
| a | 1 | 1 | 0.30593202 |
| b | 0 | 0 | 0 |
| c | 0 | 0 | 0 |
| d | 0 | 0 | 0 |
| e | 1 | 1 | 0.30593202 |
| f | 7 | 2.94591015 | 0.90124825 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | d4 |  |  |
|  | tf | wf | di |
| a | 1 | 1 | 0.41274558 |
| b | 1 | 1 | 0.41274558 |
| c | 0 | 0 | 0 |
| d | 0 | 0 | 0 |
| e | 1 | 1 | 0.41274558 |
| f | 2 | 1.69314718 | 0.69883902 |

0.713309932

0.447508076

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | tf | wf | df | Idf | qi = wf \* idf |
| a | 3 | 2.09861229 | 3 | -1.0986123 | -2.3055612 |
| b | 4 | 2.38629436 | 3 | -0.8697417 | -2.0754597 |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| f | 21 | 4.04452244 | 4 | -1.3973635 | -5.651668 |

Term Weighting

Short-documents bias happens because cosine similarity focuses on the direction of the document vector and doesn't consider the document's length. As a result, short documents usually have fewer terms, and the few terms they have can have a big impact on their similarity score. If two short documents share even a few words, their similarity score can be much higher than it should be, compared to longer documents with more diverse content.

This bias can be an issue because:

* **Limited context in short documents:** Short documents often don't have enough content to make meaningful comparisons, but they can still score highly if they share a few terms.
* **Sparse term representation:** When short documents share terms, these terms might be given too much importance, which doesn't reflect how relevant the documents are overall.

**Given the weighted tf function:**

Where:

* is a constant between 0 and 1
* is the maximum raw term frequency in document d,

**Why is this weighted tf function useful?**

This function adjusts raw term frequency by:

1. **Normalizing term frequencies**: Dividing​ by ensures that term importance is scaled relative to the most frequent term in the document. This avoids overemphasizing documents with very high raw term frequencies for a few terms.
2. **Addressing term-frequency imbalance**: The constant ensures that even low-frequency terms receive a baseline weight, preventing their importance from being completely diminished in the normalization process.

**What issue might arise?**

1. **Sensitivity to parameter α\alphaα**: The choice of can significantly affect the weighting. For example:
   * If is too high, all terms will have similar weights, reducing the impact of term frequency and hurting relevance.
   * If is too low, rare terms might receive too little weight, potentially skewing the model.
2. **Favoring high-term-frequency documents**: Documents with high might still dominate the similarity score because terms with high raw frequencies retain relatively high weights after normalization.