Evaluate tf-idf matrix for the following corpora

C1:

1. Tôi yêu Google

2. Tôi ghét Bing

3. Google là số một

4. Tôi yêu Puskin

5. Pushkin là số một

**Corpora C1:**

Tf matrix:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Tôi | Yêu | Google | Ghét | Bing | Là | Số | Một | Pushkin |
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 |
| 4 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 5 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |

Idf maxtrix: (with idf = log­2(5/df))

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Tôi | Yêu | Google | Ghét | Bing | Là | Số | Một | Pushkin |
|  | 0.74 | 1.32 | 1.32 | 2.32 | 2.32 | 1.32 | 1.32 | 1.32 | 1.32 |

Tf.Idf matrix:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Tôi | Yêu | Google | Ghét | Bing | Là | Số | Một | Pushkin |
| 1 | 0.74 | 1.32 | 1.32 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0.74 | 0 | 0 | 2.32 | 2.32 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 1.32 | 0 | 0 | 1.32 | 1.32 | 1.32 | 0 |
| 4 | 0.74 | 1.32 | 0 | 0 | 0 | 0 | 0 | 0 | 1.32 |
| 5 | 0 | 0 | 0 | 0 | 0 | 1.32 | 1.32 | 1.32 | 1.32 |

C2:

1. Tôi yêu Google. Tôi ghét Bing. Google là số một

2. Tôi yêu Pushkin. Pushkin là số một

**Corpora C2:**

Tf matrix:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Tôi | Yêu | Google | Ghét | Bing | Là | Số | Một | Pushkin |
| 1 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 0 |
| 2 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 2 |

Chuẩn hóa Tf matrix (chia mỗi phần tử cho phần tử lớn nhất = 2):

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Tôi | Yêu | Google | Ghét | Bing | Là | Số | Một | Pushkin |
| 1 | 1 | 0.5 | 1 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0 |
| 2 | 0.5 | 0.5 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 1 |

Idf maxtrix: (with idf = log­2(2/df))

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Tôi | Yêu | Google | Ghét | Bing | Là | Số | Một | Pushkin |
|  | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 |

Tf.Idf matrix:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Tôi | Yêu | Google | Ghét | Bing | Là | Số | Một | Pushkin |
| 1 | 0 | 0 | 1 | 0.5 | 0.5 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |