**텍스트 데이터 분석 6주차 과제**

**정치외교학과 20180431 김수지**

**문서 구성**

D1 : 오늘의 날씨는 맑음

D2 : 내일의 날씨는 흐림

D3 : 너의 표정은 맑음

D4 : 나의 날씨는 기쁨

**문서단어행렬 (절대빈도)**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **기쁨** | **나의** | **너의** | **내일의** | **날씨는** | **맑음** | **오늘의** | **표정은** | **흐림** |
| **D1** | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 |
| **D2** | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| **D3** | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| **D4** | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

절대빈도로 확인해보았을 때 ‘날씨는’이 총 3개의 문서에서 언급되어 가장 높은 빈도수를 나타낸다. 그 다음 순서로 ‘맑음’이 2개, 나머지 단어들은 모두 1번씩 언급된다.

**문서단어행렬 (TF-IDF)**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **기쁨** | **나의** | **너의** | **내일의** | **날씨는** | **맑음** | **오늘의** | **표정은** | **흐림** |
| **D1** | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 |
| **D2** | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| **D3** | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| **D4** | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **IDF** | 0.69 | 0.69 | 0.69 | 0.69 | 0 | 0.29 | 0.69 | 0.69 | 0.69 |

IDF = log(N/df(t)+1)

↓

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **기쁨** | **나의** | **너의** | **내일의** | **날씨는** | **맑음** | **오늘의** | **표정은** | **흐림** |
| **D1** | 0 | 0 | 0 | 0 | 0 | 0.29 | 0.69 | 0 | 0 |
| **D2** | 0 | 0 | 0 | 0.69 | 0 | 0 | 0 | 0 | 0.69 |
| **D3** | 0 | 0 | 0.69 | 0 | 0 | 0.29 | 0 | 0.69 | 0 |
| **D4** | 0.69 | 0.69 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

TF-IDF로 확인해보았을 때, 앞서 절대빈도에서 가장 높은 빈도수를 가지던 ‘날씨는’이 0으로 처리되었다. 두번째로 높던 ‘맑음’도 다른 단어들이 0.69임에 비해 0.29로 낮아졌음을 확인할 수 있다.

**문서단어행렬 ( Normalized TF-IDF, L2 )**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **기쁨** | **나의** | **너의** | **내일의** | **날씨는** | **맑음** | **오늘의** | **표정은** | **흐림** | **제곱합** | **SQRT**  **(제곱합)** |
| **D1** | 0 | 0 | 0 | 0 | 0 | 0.29 | 0.69 | 0 | 0 | 0.5602 | 0.7485 |
| **D2** | 0 | 0 | 0 | 0.69 | 0 | 0 | 0 | 0 | 0.69 | 0.9522 | 0.9758 |
| **D3** | 0 | 0 | 0.69 | 0 | 0 | 0.29 | 0 | 0.69 | 0 | 1.0363 | 1.018 |
| **D4** | 0.69 | 0.69 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.9522 | 0.9785 |

↓

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **기쁨** | **나의** | **너의** | **내일의** | **날씨는** | **맑음** | **오늘의** | **표정은** | **흐림** |
| **D1** | 0 | 0 | 0 | 0 | 0 | 0.3874 | 0.9220 | 0 | 0 |
| **D2** | 0 | 0 | 0 | 0.7071 | 0 | 0 | 0 | 0 | 0.7071 |
| **D3** | 0 | 0 | 0.6778 | 0 | 0 | 0.2850 | 0 | 0.6778 | 0 |
| **D4** | 0.7052 | 0.7052 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Normalized TF-IDF 결과, 문서 D1의 ‘오늘의’가 0.9220을 가장 높게 측정되었다. 그 다음으로는 문서 D4의 ‘기쁨’과 ‘나의’가 0.7052로 높게 측정되었다.

절대빈도에서 가장 높게 측정되었던 ‘날씨는’과 ‘맑음’은 각각 0 , (0.3874, 0.2850)으로 다른 단어들에 비해 낮게 측정되었음을 알 수 있다.