표지판, 시계이(가) 표시된 사진

자동 생성된 설명

2020년 2학기

tf-idf

|  |
| --- |
| 과목: 데이터사이언스개론 |
| 담당교수: 신효섭 |
| 제출일: 2020/10/31 |
| 학과: 기술경영학과 |
| 학번: 201512082 |
| 성명: 백 승 호 |

|  |
| --- |
| [ 목 차 ]  1. 각 문서를 tf-idf에 기반한 벡터로 표현  2. 문서쌍 고려시 similarity가 높은 쌍부터 낮은 쌍 순으로 나열 |

**1. 각 문서를 tf-idf에 기반한 벡터로 표현**

먼저 알파벳 순으로 df를 구해보면

|  |  |
| --- | --- |
| **Forecast** | 1 |
| **Home** | 4 |
| **In** | 3 |
| **Increase** | 1 |
| **July** | 2 |
| **New** | 2 |
| **November** | 1 |
| **Rise** | 2 |
| **Sales** | 4 |
| **Top** | 1 |

idf = log(n/df)

따라서 idf는

|  |  |
| --- | --- |
| **Forecast** | Log(4) |
| **Home** | 0 |
| **In** | Log(4/3) |
| **Increase** | Log(4) |
| **July** | Log(2) |
| **New** | Log(2) |
| **November** | Log(4) |
| **Rise** | Log(2) |
| **Sales** | 0 |
| **Top** | Log(4) |

문서별 tf를 구한뒤 이를 토대로 tf-idf를 구해보면

[문서 1]

|  |  |  |
| --- | --- | --- |
|  | **tf** | **Tf-idf** |
| **Forecast** | 1 | Log(4) |
| **Home** | 1 | 0 |
| **In** | 0 | 0 |
| **Increase** | 0 | 0 |
| **July** | 0 | 0 |
| **New** | 1 | Log(2) |
| **November** | 0 | 0 |
| **Rise** | 0 | 0 |
| **Sales** | 1 | 0 |
| **Top** | 1 | Log(4) |

[문서 2]

|  |  |  |
| --- | --- | --- |
|  | **tf** | **Tf-idf** |
| **Forecast** | 0 | 0 |
| **Home** | 1 | 0 |
| **In** | 1 | Log(4/3) |
| **Increase** | 0 | 0 |
| **July** | 1 | Log(2) |
| **New** | 0 | 0 |
| **November** | 0 | 0 |
| **Rise** | 1 | Log(2) |
| **Sales** | 1 | 0 |
| **Top** | 0 | 0 |

[문서 3]

|  |  |  |
| --- | --- | --- |
|  | **Tf** | **Tf-idf** |
| **Forecast** | 0 | 0 |
| **Home** | 1 | 0 |
| **In** | 2 | 2log(4/3) |
| **Increase** | 1 | Log(4) |
| **July** | 1 | Log(2) |
| **New** | 0 | 0 |
| **November** | 0 | 0 |
| **Rise** | 0 | 0 |
| **Sales** | 1 | 0 |
| **Top** | 0 | 0 |

[문서 4]

|  |  |  |
| --- | --- | --- |
|  | **Tf** | **Tf-idf** |
| **Forecast** | 0 | 0 |
| **Home** | 1 | 0 |
| **In** | 1 | Log(4/3) |
| **Increase** | 0 | 0 |
| **July** | 0 | 0 |
| **New** | 1 | Log(2) |
| **November** | 1 | Log(4) |
| **Rise** | 1 | Log(2) |
| **Sales** | 1 | 0 |
| **Top** | 0 | 0 |

**결과**

문서 1

|  |
| --- |
| Log(4) |
| 0 |
| 0 |
| 0 |
| 0 |
| Log(2) |
| 0 |
| 0 |
| 0 |
| Log(4) |

문서 2

|  |
| --- |
| **Tf-idf** |
| 0 |
| 0 |
| Log(4/3) |
| 0 |
| Log(2) |
| 0 |
| 0 |
| Log(2) |
| 0 |
| 0 |

문서 3

|  |
| --- |
| **Tf-idf** |
| 0 |
| 0 |
| 2log(4/3) |
| Log(4) |
| Log(2) |
| 0 |
| 0 |
| 0 |
| 0 |
| 0 |

문서 4

|  |
| --- |
| **Tf-idf** |
| 0 |
| 0 |
| Log(4/3) |
| 0 |
| 0 |
| Log(2) |
| Log(4) |
| Log(2) |
| 0 |
| 0 |

**2. 문서쌍 고려시 similarity가 높은 쌍부터 낮은 쌍 순으로 나열**

similarity 계산시 소수점 둘째자리까지 표현 했을 때 각 문서 쌍 a – b의 similarity는

1 – 2 는 0

1 – 3 은 0

1 – 4 는 0.13

2 – 3 은 0.4

2 – 4 는 0.32

3 – 4 는 0.05

이다

따라서 크기순 정렬하면

**(2 – 3) , (2 – 4) , (1 – 4), (3 – 4), (1 – 2), (1 – 3)** 로 나타낼수있다.