**Mail 분류 해보기**

다음과 같은 Mail 분류기가 있다고 가정하자.

분류 되어 있는 2개의 Mail에 대하여 Mail 제목이 "무료 배송"인 Mail이 어떤 Mail인지 분류해보도록 한다.

1번째 Mail은 '광고'성 Mail이다.

|  |  |
| --- | --- |
| 분류 : "광고", mail 제목 : "세일 무료 배송 할인" | |
| word\_dict Dict(사전) | words 집합 내용 |
| {'광고':  {'세일': 1, '무료': 1, '배송': 1, '할인': 1}  } | {'무료', '세일', '할인', '배송'} |
| category\_dict 사전 내용 |  |
| {'광고': 1} |  |

2번째 Mail은 '일반' Mail이다.

|  |  |
| --- | --- |
| 분류 : "일반", mail 제목 : "일정 확인" | |
| word\_dict Dict(사전) | words 집합 내용 |
| {  '광고':  {'세일': 1, '무료': 1, '배송': 1, '할인': 1},  '일반':  {'일정': 1, '확인': 1}  } | {'할인', '확인', '일정', '무료', '세일', '배송'} |
| category\_dict 사전 내용 |  |
| {'광고': 1, '중요': 1} |  |

Mail 제목이 "무료 배송"인 Mail이 어떤 Mail인가요?

**Mail 분류 방식**

Mail 제목이 "무료 배송"이다.

각각의 단어로 나누면 '무료'와 '배송'이다.

이 각 단어들이 '광고'와 '일반' 카테고리에 들어 있는 점수(확률)의 합이 큰 쪽으로 분류를 하면 된다.

|  |  |
| --- | --- |
| 항목 | 설명 |
| 카테고리 비율 | 해당카테고리도수/총카테고리도수 |
| 분자 | word\_dict[category]에 있으면 읽어 와서 더하기 1  없으면 0 더하기 1 |
| 분모 | sum(word\_dict[category]) + len( 워드 집합의 총 갯수) |
| 각 신규 단어의 비율 | 분자 / 분모 |
| 점수(확률)의 합 | log( 카테고리비율 ) + Σ( log( 각 신규 단어의 비율 )) |

**Mail 분류 연산하기**

각각의 단어로 나누면 '무료'와 '배송'이다.

이 각 단어들이 '광고'와 '일반' 카테고리에 들어 있는 점수(확률)의 합이 큰 쪽으로 분류를 하면 된다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 항목 | '광고' Mail이라고 가정하면 | | '일반' Mail이라고 가정하면 | |
|  | 무료 | 배송 | 무료 | 배송 |
| 카테고리 비율 | 2/3 | 2/3 | 2/3 | 2/3 |
| 분자 | 1 + 1 | 1 + 1 | 0 + 1 | 0 + 1 |
| 분모 | 6 + 6 | 6 + 6 | 4 + 6 | 4 + 6 |
| 각 신규 단어의 비율 | 2/12 | 2/12 | 1/10 | 1/10 |
| 점수(확률)의 합 | log(2/3) + ( log(2/12) + log(2/12) )  = -1.73239376 | | log(2/3) + ( log(1/10) + log(1/10) )  = -2.176091259 | |

카테고리의 수에 상관 없이 점수(확률)의 합이 가장 큰 것을 선택하면 된다.

예시에서는 '광고'의 점수의 합(-1.73239376)이 더 크다.

따라서, Mail 제목이 "무료 배송"인 Mail은 '광고'성 Mail이다.