1.1 입력문서가 {fast, furious, fun}만을 주요단어로 가질 때, 이 문서는 얼마의 확률로 어떤문서로 분류되는가?

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 단어 | fun | couple | love | fast | furious | shoot | fly | 분류 |
| 1 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | Comedy |
| 2 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | Action |
| 3 | 2 | 1 | 0 | 1 | 0 | 0 | 1 | Comedy |
| 4 | 1 | 0 | 0 | 0 | 1 | 2 | 0 | Action |
| 5 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | Action |
| test | 1 | 0 | 0 | 1 | 1 | 0 | 0 | ? |

계산에 필요한 확률

test일 때 comedy와 action으로 각각 분류될 확률

Test는 action으로 분류될 확률이 높기 때문에 action으로 분류된다

1.2 어떠한 문제점이 있고, 이를 해결하기 위해 어떻게 할 것인가?

문제점 : Test가 comedy가 될 확률이 0 이 된다.

해결 : 라플라스 스무딩

분자에 1, 분모에 7을 더하여 계산한다

test일 때 comedy와 action으로 각각 분류될 확률

Test는 action으로 분류될 확률이 높기 때문에 action으로 분류된다

* 1. 입력문서가 {fast, furious, fun}만을 주요단어로 가질 때, 이 문서는 얼마의 확률로 어떤문서로 분류되는가?

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 단어 | fun | couple | love | fast | furious | shoot | fly | 분류 |
| 1 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | Comedy |
| 2 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | Action |
| 3 | 2 | 1 | 0 | 1 | 0 | 0 | 1 | Comedy |
| 4 | 1 | 0 | 0 | 0 | 1 | 2 | 0 | Action |
| 5 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | Action |
| test | 1 | 0 | 0 | 1 | 1 | 0 | 0 | ? |

계산에 필요한 확률

P( Y=comedy)=0.4 P( Y=action)=0.6

P( X=fun | Y=comedy)=3/9 P( X=fun | Y=action)=1/11

P( X=fast | Y=comedy)=1/9 P( X=fast | Y=action)=2/11

P( X=furious | Y=comedy)=0/9 P( X=furious | Y=action)=2/11

test일 때 comedy와 action으로 각각 분류될 확률

P( Y=comedy | X=test)=0.4\* 3/9\*1/9\*0/9=0

P( Y=action | X=test)=0.6\* 1/11\*2/11\*2/11=0.00180

Test는 action으로 분류될 확률이 높기 때문에 action으로 분류된다

1.2 어떠한 문제점이 있고, 이를 해결하기 위해 어떻게 할 것인가?

문제점 : Test가 comedy가 될 확률이 0 이 된다.

해결 : 라플라스 스무딩

분자에 1, 분모에 7을 더하여 계산한다

test일 때 comedy와 action으로 각각 분류될 확률

P( Y=comedy | X=test)=0.4\* 4/16\*2/16\*1/16=0.00078

P( Y=action | X=test)=0.6\* 2/18\*3/18\*3/18=0.00185

Test는 action으로 분류될 확률이 높기 때문에 action으로 분류된다