나이브 베이즈 서술형 과제 풀이

문제

문서번호	주요단어	문서분류
1	fun, couple, love, love	comedy
2	fast, furious, shoot	action
3	couple, fly, fast, fun, fun	comedy
4	furious, shoot, shoot, fun	action
5	fly, fast, shoot, love	action

1.1 입력문서가 {fast, furious, fun} 만을 주요단어로 가질 때, 이 문서는 얼마의 확률로 어떤 문서로 분류되는가?
1.2 어떠한 문제점이 있고, 이를 해결하기 위해 어떻게 할 것인가? (방법론만 제시)
*식과 함께 답을 같이 제출해주세요~

나이브 베이즈 서술형 과제 풀이

*나이브 베이즈 함수식

$$f^*(x) = argmax_{Y=y}\mathrm{P}(\mathrm{X}=\mathrm{x}|\mathrm{Y}=\mathrm{y})\mathrm{P}(\mathrm{Y}=\mathrm{y}) \quad \approx \boldsymbol{argmax_{Y=y}} \boldsymbol{P(Y=y)}\, \boldsymbol{\Pi_{1\le i\le d}}\mathbf{P}(\mathbf{X}=\boldsymbol{x_i}|\mathbf{Y}=\boldsymbol{y})$$

1.1 p(comedy | x) = p(comedy) * p(fast | comedy) * p(furious | comedy) * p(fun | comedy)

$$= \frac{2}{5} * \frac{1}{9} * \frac{0}{9} * \frac{3}{9} = 0$$

9: Comedy 문서에 등장하는 총 단어의 수

p(action | x) = p(action) * p(fast | action) * p(furious | action) * p(fun | action)

$$= \frac{3}{5} * \frac{2}{11} * \frac{2}{11} * \frac{1}{11} = 0.0018$$

11: action 문서에 등장하는 총 단어의 수

따라서 입력문서는 사후확률이 보다 큰 action으로 분류된다.

1.2 ·문제점: 위 문제 1.1에서, comedy 문서에는 furious 단어의 빈도가 0이므로,
furious 단어를 포함하는 새로운 자료에 대한 사후확률은 항상 0이 되어 버린다.

·해결책: 이러한 문제점을 해결하기 위해, 작은 수를 더해주어 계산을 수행한다. (=라플라스 스무딩)