중앙대학교 인문콘텐츠연구소 비전공자를 위한 2022 여름 프로그래밍 강좌

Colab을 활용한 파이썬 텍스트 분석 입문

나이브 베이즈 분류
(Naïve Bayes Classification)

2022/08/26(금) 오후1시~오후5시@중앙대 303관 B101호

강사: 조 희 련 (중앙대 인문콘텐츠연구소 HK교수)

분류(Classification)

- 대상을 분류하는 행위는 대상을 인식하고 구분하는 행위
- 고양이와 강아지를 분류
- 사람의 얼굴을 분류, 사람의 목소리를 분류
- 과제물(리포트, 그림, 악기 연주 등)에 평어(A, B, C, D, 또는 F)를 부여
- 텍스트를 특정 카테고리로 분류
    - 스팸 메일 분류(spam detection)
    - 상품평의 감성 분석(sentiment analysis)
    - 저자 판별(authorship attribution)
    - 뉴스 기사의 카테고리 분류(document classification)

- 다양한 분류 문제를 기계학습을 이용하여 자동으로 분류(classification)

기계학습(Machine Learning)

- 기계학습은 인공지능의 한 분야로, (사람의 지시 없이) 컴퓨터가 데이터를 학습하여 문제를 해결하는 절차(알고리즘)을 연구하는 분야임
- 크게 지도학습(supervised learning), 비지도학습(unsupervised learning), 강화학습(reinforcement learning)으로 나뉨
- 오늘 배울 나이브 베이즈 분류는 지도학습의 한 종류
- 일반적으로 문서 분류와 같은 분류(classification) 문제는 지도학습으로 풂

- 물론 규칙(rule)을 나열하여 텍스트를 분류할 수도 있음(예: *고수익*, *대출*, *현금* → 스팸)
- 그러나 올바른 규칙을 정의하기 어렵고 망라하기도 힘듦

지도학습을 이용한 문서 분류 과정

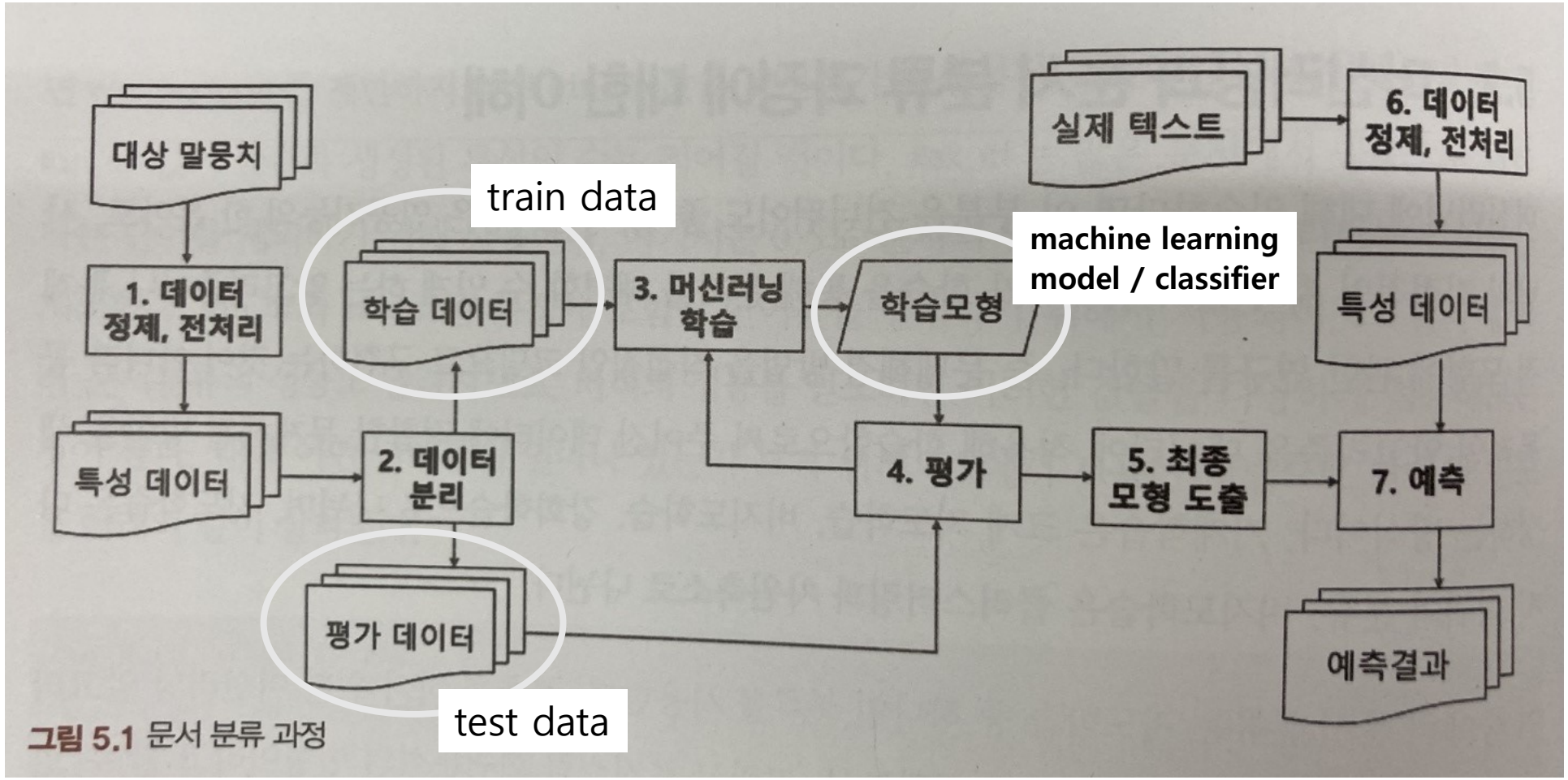

그림 5.1 문서 분류 과정

출처: 박상언, 강주영, 정석찬, "파이썬 텍스트 마이닝 완벽 가이드", p. 100.

지도학습을 이용한 문서 분류의 형식화

- 지도학습에서는 하나의 입력에 대하여 하나의 출력이 정의된 훈련 데이터(train set)를 이용하여 기계학습 모델을 학습함
- 예: '스팸'이라는 레이블이 달린 문자와 '정상'이라는 레이블이 달린 문자를 학습 데이터로 수집하여 스팸 문자 분류기를 학습

  - *Input:*
    - a set of N documents: d
    - a fixed set of classes: $C = \{c_1, c_2, \ldots, c_j\}$
    - a training set of n hand-labeled documents: $(d_1,c_1), (d_2,c_2), \ldots, (d_n,c_n)$

  - *Output:*
    - a learned classifier $f:d \rightarrow c$

Bag of Words 가정

- 단어들 간의 순서를 고려하지 않음
- 나이브 베이즈 분류기도 Bag of Words 모델을 가정함

- 출처: https://web.stanford.edu/~jurafsky/slp3/4.pdf (p. 3)

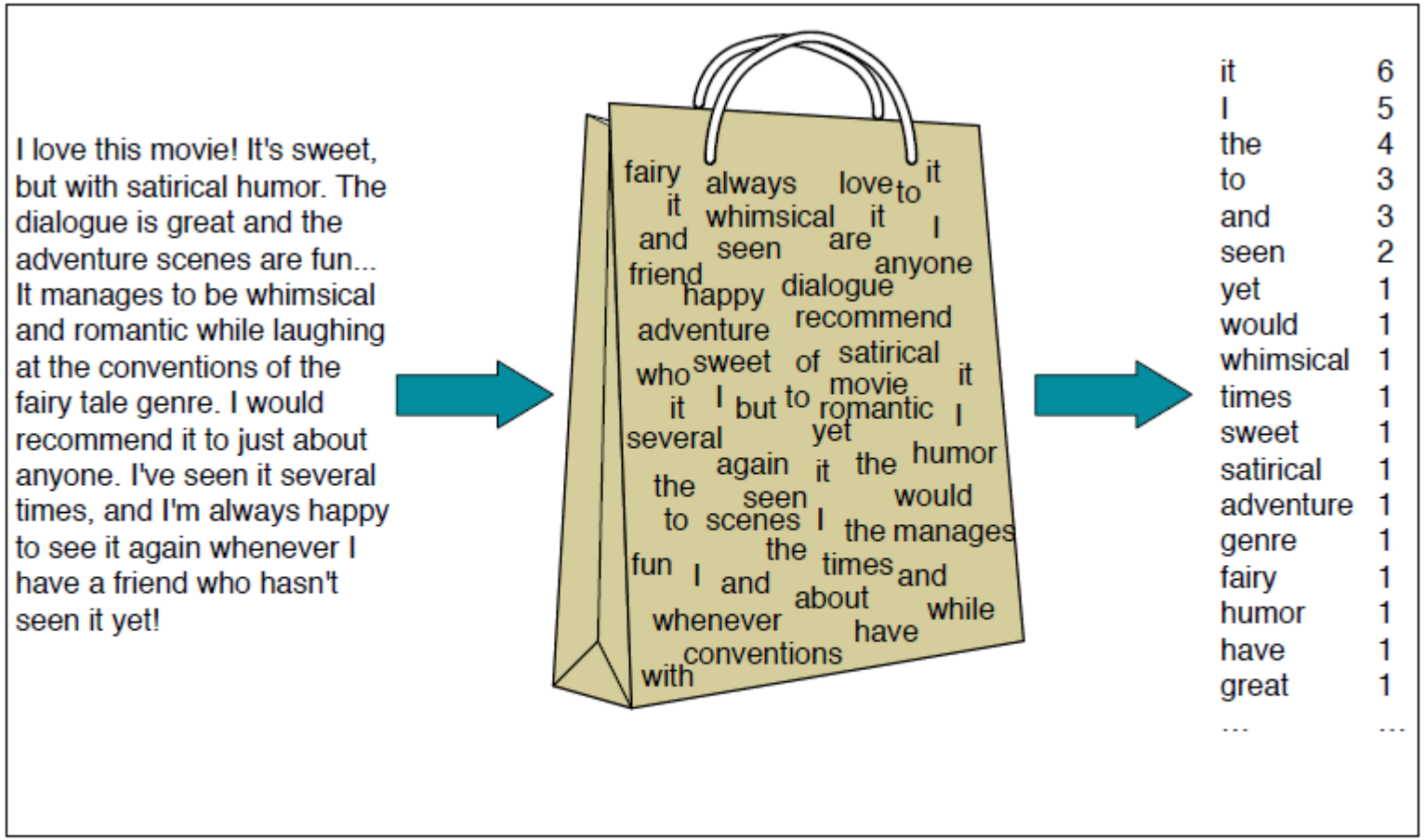

Figure 4.1 Intuition of the multinomial naive Bayes classifier applied to a movie review. The position of the words is ignored (the *bag of words* assumption) and we make use of the frequency of each word.

나이브 베이즈 알고리즘 개요(1/3)

- 한줄 요약: 베이즈 정리(Bayes' theorem)를 활용한 단순한(naïve) 자동 분류 기법

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)}$$

  - *X*가 (입력으로) 주어졌을 때, *Y* 일(또는 *Y* 를 출력할) 확률: *P*(*Y* | *X*) → 조건부확률
  - *P*(*X*) ≠ 0
  - *P*(*X*) 와 *P*(*Y*) 는 각각 *X* 와 *Y* 를 관측할 확률
  - P(*Y* | *X*) and P(*X* | *Y*) 는 조건부확률
  - *X* = 새로 수신한 문자, *Y* = {스팸문자, 정상문자}
- 나이브 베이즈

$$P(cat|doc) = \frac{P(cat)P(doc|cat)}{P(doc)} \propto P(cat)P(doc|cat)$$

- 텍스트를 Bag of words로 가정함(=단어 순서를 고려하지 않음); 주어진 카테고리 안에서 단어의 출현 확률은 서로 독립

$$P(doc|cat) = P(word_1 \wedge \cdots \wedge word_k|cat) = \prod_i P(word_i|cat)$$

나이브 베이즈 알고리즘 개요(2/3)

- 한줄 요약: 베이즈 정리(Bayes' theorem)를 활용한 단순한(naïve) 자동 분류 기법

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d)$$

MAP is "maximum a posteriori" = most likely class

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)}$$

Bayes Rule

$$= \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

Dropping the denominator

- 나이브 베이즈

$$P(cat|doc) = \frac{P(cat)P(doc|cat)}{P(doc)} \propto P(cat)P(doc|cat)$$

- 텍스트를 Bag of words로 가정함(=단어 순서를 고려하지 않음); 주어진 카테고리 안에서 단어의 출현 확률은 서로 독립

$$P(doc|cat) = P(word_1 \wedge \cdots \wedge word_k|cat) = \prod_i P(word_i|cat)$$

나이브 베이즈 알고리즘 개요(3/3)

- 한줄 요약: 베이즈 정리(Bayes' theorem)를 활용한 단순한(naïve) 자동 분류 기법

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\, P(c \mid d)$$

MAP is "maximum a posteriori" = most likely class

$$= \underset{c \in C}{\operatorname{argmax}}\, \frac{P(d \mid c)P(c)}{P(d)}$$

Bayes Rule

$$= \underset{c \in C}{\operatorname{argmax}}\, P(d \mid c)P(c)$$

Dropping the denominator

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\, P(x_1, x_2, \ldots, x_n \mid c)P(c)$$

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}}\, P(c_j) \prod_{x \in X} P(x \mid c)$$

(단어들이 서로 독립)

$$P(doc|cat) = P(word_1 \wedge \cdots \wedge word_k|cat) = \prod_i P(word_i|cat)$$

$$P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c)$$

나이브 베이즈 문서 분류 : 실제 계산 방법(1/3)

- 학습 데이터의 문서 빈도와 단어 빈도를 이용하여 확률을 계산

$$\hat{P}(c_j) = \frac{doccount(C = c_j)}{N_{doc}}$$

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

- 수식 위: 특정 카테고리일 확률
- 수식 아래: 단어 w_i가 c_j 카테고리를 가지는 문서들의 모든 단어 중에서 나타날 확률
- 이때 분모는 c_j 카테고리를 가지는 문서를 하나의 문서로 통합한 후,
- 통합문서 내 모든 단어의 빈도를 더함

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c_j) \prod_{x \in X} P(x \mid c)$$

나이브 베이즈 문서 분류 : 실제 계산 방법(2/3)

- 예를 들어 햄(정상) vs. 스팸으로 문자를 분류하는 문제를 생각해 봄
- 햄/스팸 카테고리 확률은 각각 0.5이고, 햄 문자와 스팸 문자에 출현하는 단어들의 확률들을 계산(학습)하여 나이브 베이즈 분류기를 구축한다고 가정함
- 새로운 문자X $(w_1, w_2, \ldots, w_n)$ 가 주어졌을 때, 해당 문자 속 단어들이 훈련 데이터에 출현했던 단어들이면 문제가 없음

$$P(w_1, w_2, \ldots w_n|Ham) = .90$$
$$P(w_1, w_2, .. w_n|Spam) = .10$$

- 그런데 훈련 데이터에는 없던 새로운 단어가 새로운 문자X에 출현한다면?
- 새로운 문자X $(w_1, w_2, \ldots, w_n, w_{n+1})$ 에는 분류기 학습 때 한 번도 안 나온 단어 w_{n+1} 가 포함되기 때문에, 이로 인해 계산이 0 (zero)이 되는 문제가 발생함

$$P(w_{n+1}|Ham) = P(w_{n+1}|Spam) = 0$$
$$P(w_1, w_2, \ldots w_n, w_{n+1}|Ham) = P(w_1, w_2, \ldots w_n|Ham) * P(w_{n+1}|Ham) = 0$$
$$P(w_1, w_2, .. w_n, w_{n+1}|Spam) = P(w_1, w_2, \ldots w_n|Spam) * P(w_{n+1}|Spam) = 0$$

$$c_{MAP} = \text{argmax}_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c)$$

나이브 베이즈 문서 분류 : 실제 계산 방법(3/3)

- 이렇게 미지의 단어로 인해 계산 결과가 0 이 되는 문제를 해결하기 위해
- Laplace (add-1) smoothing 을 적용함 (계산 결과가 zero가 되는 것을 방지)

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c)+1}{\sum_{w \in V} \left(count(w, c)+1\right)}$$

$$= \frac{count(w_i, c)+1}{\left(\sum_{w \in V} count(w, c)\right) + |V|}$$

- [참고사항] 실제 컴퓨터를 이용한 계산에서는 소수점 계산의 오차 문제를 회피하기 위해 로그 변환 후 더하기 연산을 함 → log(*xy*) = log(*x*) + log(*y*)

$$c_{MAP} = \text{argmax}_c \, \hat{P}(c) \prod_i \hat{P}(x_i \mid c) \quad \rightarrow \quad c_{NB} = \underset{c_j \in C}{\text{argmax}} \log P(c_j) + \sum_{i \in positions} \log P(x_i \mid c_j)$$

Laplace (add-one) Smoothing 계산 예시

- 훈련 데이터에서 특정 클래스를 가지는 문서에 특정 단어가 없는 경우

$$\hat{P}(w_i|c) = \frac{count(w_i,c)}{\sum_{w\in V} count(w,c)}$$

$$\hat{P}(w_i \mid c) = \frac{count(w_i,c)+1}{\sum_{w\in V}\left(count(w,c)+1\right)} = \frac{count(w_i,c)+1}{\left(\sum_{w\in V} count(w,c)\right) + |V|}$$

$$\hat{P}(\text{“fantastic”}|\text{positive}) = \frac{count(\text{“fantastic”},\text{positive})}{\sum_{w\in V} count(w,\text{positive})} = 0$$

$$\hat{P}(w_i|c) = \frac{count(w_i,c)+1}{\sum_{w\in V}\left(count(w,c)+1\right)} = \frac{count(w_i,c)+1}{\left(\sum_{w\in V} count(w,c)\right)+|V|}$$

나이브 베이즈 문서 분류 계산 예시

Table 13.1: Data for parameter estimation examples.

		docID	words in document	in c = China?	
	training set	1	Chinese Beijing Chinese	yes	
		2	Chinese Chinese Shanghai	yes	
		3	Chinese Macao	yes	
		4	Tokyo Japan Chinese	no	
	test set	5	Chinese Chinese Chinese Tokyo Japan	?	

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c)+1}{\sum_{w \in V} \left(count(w, c)+1\right)}$$

$$= \frac{count(w_i, c)+1}{\left(\sum_{w \in V} count(w, c)\right) + |V|}$$

$$\hat{P}(c) = 3/4 \qquad \hat{P}(\text{Chinese}|c) = (5+1)/(8+6) = 6/14 = 3/7$$

$$\hat{P}(\text{Tokyo}|c) = \hat{P}(\text{Japan}|c) = (0+1)/(8+6) = 1/14$$

$$\hat{P}(\overline{c}) = 1/4 \qquad \hat{P}(\text{Chinese}|\overline{c}) = (1+1)/(3+6) = 2/9$$

$$\hat{P}(\text{Tokyo}|\overline{c}) = \hat{P}(\text{Japan}|\overline{c}) = (1+1)/(3+6) = 2/9$$

$$\hat{P}(c|d_5) \propto 3/4 \cdot (3/7)^3 \cdot 1/14 \cdot 1/14 \approx 0.0003.$$

$$\hat{P}(\overline{c}|d_5) \propto 1/4 \cdot (2/9)^3 \cdot 2/9 \cdot 2/9 \approx 0.0001.$$

$$\hat{P}(t|c) = \frac{T_{ct}+1}{\sum_{t' \in V}(T_{ct'}+1)} = \frac{T_{ct}+1}{(\sum_{t' \in V} T_{ct'}) + B'}$$

$$c_{MAP} = \text{argmax}_c \, \hat{P}(c) \prod_i \hat{P}(x_i \mid c)$$

나이브 베이즈 감성 분석 계산 예시

②

$$\hat{P}(w_i|c) = \frac{count(w_i,c)+1}{\left(\sum_{w\in V} count(w,c)\right)+|V|}$$

①

$$\hat{P}(c) = \frac{N_c}{N_{doc}}$$

$$P(-) = \frac{3}{5} \qquad P(+) = \frac{2}{5}$$

$$P(\text{“predictable”}|-) = \frac{1+1}{14+20} \qquad P(\text{“predictable”}|+) = \frac{0+1}{9+20}$$

$$P(\text{“no”}|-) = \frac{1+1}{14+20} \qquad P(\text{“no”}|+) = \frac{0+1}{9+20}$$

$$P(\text{“fun”}|-) = \frac{0+1}{14+20} \qquad P(\text{“fun”}|+) = \frac{1+1}{9+20}$$

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer
Test	?	predictable with no fun

③

$$P(-)P(S|-) = \frac{3}{5} \times \frac{2\times 2\times 1}{34^3} = 6.1\times 10^{-5}$$

$$P(+)P(S|+) = \frac{2}{5} \times \frac{1\times 1\times 2}{29^3} = 3.2\times 10^{-5}$$

이후 Colab 실습 과정

- 나이브 베이즈 분류기를 하나부터 파이썬으로 직접 구현함
- scikit-learn 기계학습 라이브러리의 나이브 베이즈 분류기 객체를 사용하여 분류를 수행함
  - 위 두 과정은 영어로 된 스팸 문자 데이터 세트로 수행
- 실습: scikit-learn 기계학습 라이브러리로 한국어 영화평(네이버 영화평)을 긍정/부정 영화평으로 자동 분류함