07

비정형데이터분석

텍스트 데이터의 전처리(1)

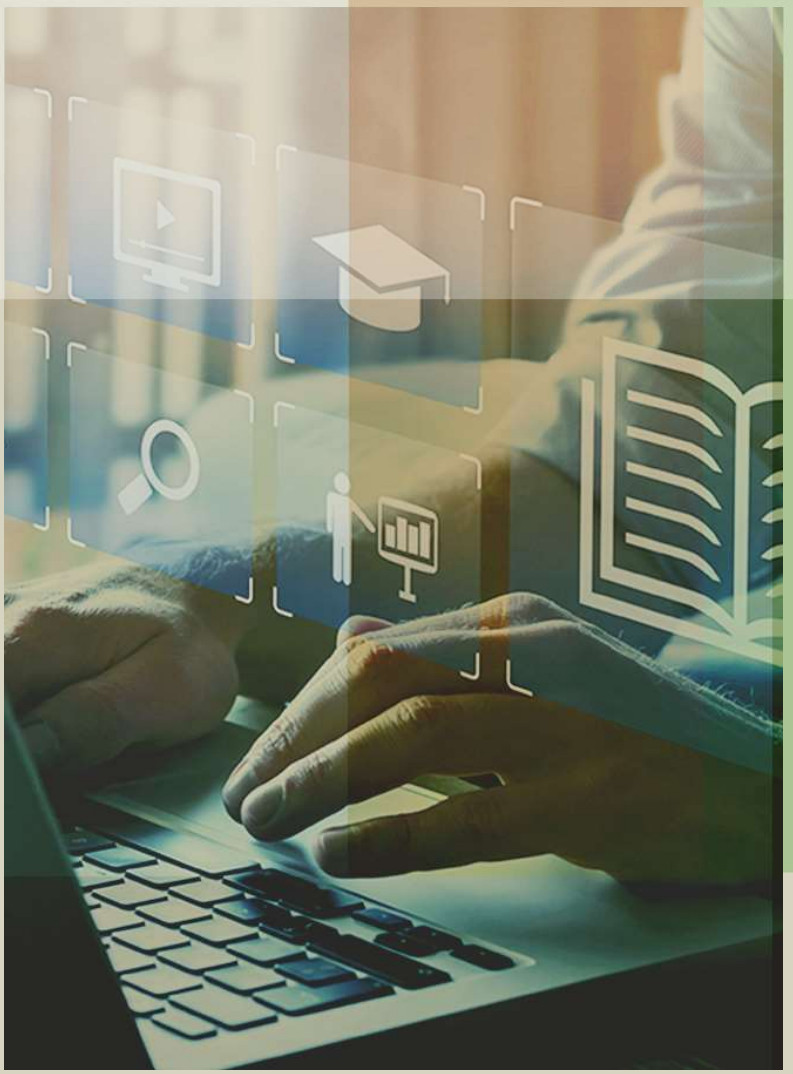
통계·데이터과학과 장영재 교수

학습목차

1. 텍스트데이터와수치형데이터의표현방식의차이
2. 텍스트데이터의통계분석을위한기본가설
3. 토큰화(tokenization)

01

텍스트 데이터와 수치형 데이터의 표현 방식의 차이



1. 텍스트 데이터와 수치형 데이터의 표현 방식의 차이 - ① 수치형 데이터의 표현 방식

- 수치형 데이터는 흔히 벡터, 행렬, 데이터 프레임 등으로 정형화된 구조로 표현
  - 정해진 구조에 따라 표현되고 행과 열에 따라 수치의 의미가 결정되므로 정형적(structure) 데이터라고 부름
    - → 붓꽃(iris) 데이터는 꽃받침(sepal)의 길이와 너비, 꽃잎(petal)의 길이와 너비 등 4개의 설명변수와 붓꽃의 종류에 해당하는 하나의 반응변수로 구성되는 데이터프레임

1. 텍스트 데이터와 수치형 데이터의 표현 방식의 차이 - ② 텍스트 데이터의 표현 방식

텍스트 데이터는 수치형 데이터와 달리 정형화하기 쉽지 않다는 특성

- 텍스트 데이터는 길이에 명시적인 제약이 없는 경우가 많음
- 같은 의미라도 사람에 따라 다양한 단어와 문장 구조를 사용해 표현할 수 있음

1. 텍스트 데이터와 수치형 데이터의 표현 방식의 차이 - ② 텍스트 데이터의 표현 방식

- "비정형적(unstructured)"이라는 단어는 텍스트 데이터가 구조를 갖추지 못했다는 의미가 아니라 수치형 데이터와는 달리 복잡하고 미묘한 문법적인 구조 속에서 표현
  - → "The world's most valuable resource is no longer oil, but data."에서 "no longer … but"이라는 구문의 의미를 정확하게 반영하지 못한다면 엉뚱한 의미로 해석될 수 있음
- 몇 가지 가정들을 기반으로 텍스트 데이터를 수치형 데이터로 표현하는 방법이 필요

02

텍스트 데이터의 통계 분석을 위한 기본 가설

2. 텍스트 데이터의 통계 분석을 위한 기본 가설 - ① 통계적 의미론 가설 (Statistical Semantic Hypothesis)

- 다양한 방식으로 표현되는 복잡한 텍스트 데이터를 통계적으로 분석하기 위해서는 가정을 도입하여 문제를 단순화할 필요
- 통계적 의미론 가설이란 사람들이 단어를 사용하는 통계적 규칙성(pattern)으로부터 말하는 사람 또는 글을 쓴 사람이 뜻하는 바를 알 수 있다는 주장
  - 사람들은 자신의 뜻을 표현하기 위해 필요한 단어들을 선택
    - → 표현하고자 하는 바가 같은 경우 선택된 단어들에 공통적인 특성이 나타남

2. 텍스트 데이터의 통계 분석을 위한 기본 가설 - ② 단어주머니 가설 (Bag-of-Words Hypothesis)

- 단어주머니 가설은 통계적 의미론 가설을 더 구체화한 가설 중 하나
  - 주머니(bag)란 원소들의 중복을 허용하는 집합을 말하며 중복집합(multiset) 또는 다중집합이라고도 함
  - 텍스트 데이터에 사용된 단어들의 빈도(frequency)가 텍스트 데이터의 의미를 결정하는 중요한 단서가 된다는 가설
    - → 서로 다른 두 문서에 포함된 단어들과 그 단어들의 출현 빈도가 비슷할수록(단어주머니가 비슷할수록) 두 문서가 비슷한 의미를 가질 가능성이 높다는 의미

2. 텍스트 데이터의 통계 분석을 위한 기본 가설 - ② 단어주머니 가설 (Bag-of-Words Hypothesis)

- 단어주머니 가설에서는 텍스트 데이터를 단어주머니로 표현할 때 원래의 텍스트 데이터가 가지고 있는 문장 구조와 단어 배열순서 무시
  - 단어의 출현 여부와 빈도만 중요하게 평가
    - → 텍스트 데이터에 포함되어 있던 많은 정보가 사라질 우려
  - 예를 들어 'Kim was loved by everybody.'의 단어주머니 {Kim, was, loved, by, everybody}와 'Everybody loved Kim.'의 단어주머니 {everybody, Kim, loved} 는 유사한 의미
    - → 'Kim loved everybody.'의 단어주머니는 {everybody, Kim, loved}로 표현할 수 있지만 배열순서가 달라져 의미의 차이

2. 텍스트 데이터의 통계 분석을 위한 기본 가설 - ② 단어주머니 가설 (Bag-of-Words Hypothesis)

- 이런 제약에도 불구하고 단순화를 통해 효율적인 데이터 분석을 가능하게 할 뿐만 아니라 많은 실제 분석과제에서 유용성이 입증
  - 지금까지도 텍스트 데이터의 기본적인 수치화 기법으로 많이 활용

3. 단어주머니 생성의 예 (R을 이용한 사례)

- R을 이용하여 텍스트 데이터를 단어 단위로 나누고 단어주머니 생성
  - 텍스트 데이터의 문자열을 분해하여 단어주머니를 만드는 strsplit( ) 함수를 활용
    - → split을 공백(whitespace) " ", 즉 띄어쓰기를 기준으로 지정하면 공백을 기준으로 문자열을 나누고 split을 ""로 지정하면 문자열을 각 문자별로 나눔

```
> x <- c('Kim was loved by everybody.', 'Everybody loved Kim','Kim loved everybody.')
> strsplit(x, split= " ")
> strsplit(x, split = " ")[[2]][3]
[1] "Kim"
```

03

토큰화(tokenization)

3. 토큰화(tokenization) - ① 토큰(token)과 토큰화(tokenization)

- '토큰(token)'이란 고유한 의미를 가지고 있어서 더 이상 나눌 필요가 없는 하나의 단위를 의미(Webster and Kit, 1992)

  - 공백을 기준으로 문자열을 나누어 이를 토큰, 즉 고유한 의미를 가진 최소 단위라고 하였을 때 발생할 수 있는 문제들

    → 축약(contraction)된 표현: 여러 단어가 공백 없이 하나의 묶음으로 묶여 있는 경우 (ex: 'they're'는 they are'의 의미)

    → n-gram: 여러 단어 사이에 공백이 있지만 의미상 하나의 묶음으로 보는 것이 타당한 경우 (ex: 'give up' 등 숙어나 'bed and breakfast', 'grab and go', 'sharing economy', 'hard disk drive', 'the White House' 등과 같은 관용어구)

3. 토큰화(tokenization) - ① 토큰(token)과 토큰화(tokenization)

- 토큰화(tokenization)는 위와 같은 특징을 고려하여 텍스트 데이터를 단일 의미를 가진 토큰 단위로 나누는 것을 의미
  - 축약(contraction)된 표현은 두 개의 토큰으로 나누고 n-gram과 같은 경우는 하나의 토큰으로 인식하는 과정
    - → 토큰화를 통해 식별된 토큰들은 텍스트 데이터의 통계적 분석을 위한 기본 단위

3. 토큰화(tokenization) - ② R을 이용한 토큰화

- R 함수를 이용하여 축약된 표현들을 원래의 단어로 나누고 여러 개의 단어로 구성된 n-gram을 하나의 토큰으로 식별

```
> sub(pattern, replacement, x, ignore.case = FALSE)
> gsub(pattern, replacement, x, ignore.case = FALSE)
```

3. 토큰화(tokenization) - ② R을 이용한 토큰화

- pattern은 텍스트 데이터에서 찾아서 변환할 대상이 되는 문자열을, replacement는 변환 후 입력될 문자열을, x는 패턴을 찾을 텍스트 데이터 벡터를 의미

  → ignore.case는 대소문자와 관련된 인수로 FALSE로 지정되어 있으면 대문자와 소문자를 별개의 문자로 구분

```
> sub(pattern = "ouse", replacement = "ay", x = "mouse in the house")
[1] "may in the house" # 패턴이 일치하는 첫 번째 문자열만 변환

> gsub(pattern = "ouse", replacement = "ay", x = "mouse in the house")
[1] "may in the hay" # 패턴이 일치하는 모든 문자열을 변환
```

3. 토큰화(tokenization) - ② R을 이용한 토큰화

1) 축약된 표현의 토큰화

- 축약된 형태로 자주 사용되는 표현들을 모아 미리 사전을 만들어 두고 텍스트 데이터에서 이 사전에 포함되어 있는 표현들을 찾아 변환

```
> contraction_dict <- list(c("don't", "it's", "you're"), c("do not", "it is", "you are"))
> dictlen <- length(contraction_dict[[1]]) # 축약된 표현의 수

> datstr <- "I don't think you're ready.“
> for (stri in 1:dictlen) {
      datstr <- gsub(pattern = contraction_dict[[1]][stri],
       replacement = contraction_dict[[2]][stri], x = datstr)
 }
> datstr
[1] "I do not think you are ready."
> strsplit(datstr, " ")
[[1]]
[1] "I"     "do"    "not"   "think" "you"   "are"   "ready."
 # 단어들을 원래 형태로 토큰화
```

3. 토큰화(tokenization) - ② R을 이용한 토큰화

2) n-그램(n-gram)의 토큰화

- 미리 사전을 정의해두고 사전에 포함된 표현을 하나의 토큰으로 변환

```
> ngram_dict <- list(c("bed and breakfast", "grab and go", "New York"), c("bed_and_breakfast", "grab_and_go", "New_York"))
> ndictlen <- length(ngram_dict[[1]]) # 관용어구 수
> datstr <- "It's one of the best bed and breakfast in New York."
> for (stri in 1:ndictlen) {
      datstr <- gsub(x=datstr, pattern=ngram_dict[[1]][stri],
          replacement = ngram_dict[[2]][stri])
  }     # 관용어구를 하나의 토큰으로 변환

> datstr
[1] "It's one of the best bed_and_breakfast in New_York."
```

3. 토큰화(tokenization) - ② R을 이용한 토큰화

```
> strsplit(datstr, " ")
[[1]]
[1] "it"              "is"              "one"
[4] "of"              "the"             "best"
[7] "bed_and_breakfast" "in"            "New_York."
# 문장 토큰화의 완성
```

- 우리말에서는 형태소 분석을 통해 어미와 조사를 별도로 분리해내고 토큰화하여 명사와 어간을 중심으로 텍스트 데이터를 분석하기도 함

실습하기

다음시간안내

08

텍스트 데이터의 전처리(2)