중간고사

박성은

2022-10-29

## 중간고사 개요

본 중간고사는 **텍스트 정보처리와 NLP**의 학업 성취도를 평가하기 위한 시험입니다.

응시자는 다음 문제들이 제시하는 목적을 이해하고, 문제에서 제시하는 방법에 따라 R 스크립트를 knitr 청크에 채워넣어 답하여야 합니다.

### 평가 기준

학업 성취도를 평가하는 방법은 다음과 같습니다.

* 주어진 문제에 대한 해결 여부 확인하여, 문제 해결 완료 시 부여한 배점 획득
* 본 평가에서는 심화문제를 추가 출제하여, 배점 30점을 추가 부여합니다.
  + 문제 해결 완료: 문제에 부여한 배점 획득
  + 문제 해결 실패: 문제에 부여한 배점 일부 획득
    - 문제를 수행하는 과정의 R 스크립트를 평가하여,
    - **문제를 해결한 정도, 방법의 독창성, 효율성을 고려하여 문제에 부여한 배점 \* 0.8 내에서 점수 부여 가능**

상기 평가 기준처럼 주어진 문제를 최종적으로 수행하지 못하더라도 부분점수가 주어지므로, 최선을 다해서 수행 과정으로서의 R 스크립트를 기술해 주기 바랍니다.

### 배점

본 평가는 100점 만점을 기준으로 출제됩니다. 그리고 개별 문항마다 난이도나 수행 소요 시간을 고려하여 배점을 부여하였습니다. 배점은 문제와 함께 제시되며, **심화문제를 추가 출제하여, 배점 30점을 추가 부여합니다.**

### 주의사항

* 본 평가는 텍스트 정보처리와 NLP를 수행하는 기초 스킬을 테스트합니다. 그리고 R 코드로 문제를 해결하므로 재현성에 입각해서 평가합니다.
  + 그러므로, **제출한 R 마크다운 문서의 R 스크립트가 문제 해결을 재현하지 못하면 점수를 부여하지 않습니다.**
* 본 평가는 주어진 시간 내에 응시하는 것을 전제로 출제됩니다. 그래서 난이도와 수행 소요 시간을 고려하여 배점하였습니다.
  + 그러므로, **감독관이 공지한 시험 종료 시간 내에 제출하지 않을 경우에는 점수를 부여하지 않습니다.**
  + 응시자는 문제를 해결하기 위해서 작성한 **R 마크다운 파일(exam\_question\_mid.Rmd)**과
  + R 마크다운 파일의 **실행 결과인 MS-Word 파일(exam\_question\_mid.docx)을 함께 제출**해야 합니다.

### 제출방법

* 파일이름 변경
  + exam\_question\_mid\_성명.docx
  + exam\_question\_mid\_성명.Rmd
* 파일제출 email
  + [choonghyun.ryu@gmail.com](mailto:choonghyun.ryu@gmail.com)

## 응시 환경 테스트 및 준비사항

### 응시자 성명 기입

본 파일의 헤더 **author: 홍길동**에서 홍길동을 응시자의 이름으로 변경하세요.

### 환경 테스트

다음 청크에서 에러가 발생하지 않아야 시험을 정상적으로 수행할 수 있습니다. 문제가 발생할 경우에는 감독관에게 연락하십시요.

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.4.1   
## ✔ readr 2.1.2 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(tidytext)  
library(bitTA)  
library(koscrap)  
library(wordcloud2)  
  
## bitTA 패키지 버전 확인 : 1.3.1.9000인가?  
packageVersion("bitTA")

## [1] '1.3.1.9000'

## Mecab 형태소 분석  
morpho\_mecab("아버지가 방에 들어가신다.")

## NNG NNG   
## "아버지" "방"

### 준비사항

**응시자는 네이버 뉴스를 수집하기 위한 오픈 API 키를 미리 준비해야 합니다.**

## 재현가능 연구와 R 마크다운

### 청크 구조의 이해

#### 문제 1

* 배점: 5점

다음은 cars라는 데이터 프레임을 집계하는 knitr 청크입니다. 청크의 옵션을 추가해서 R 명령어가 출력되지 않도록 수정하십시요.

## speed dist   
## Min. : 4.0 Min. : 2.00   
## 1st Qu.:12.0 1st Qu.: 26.00   
## Median :15.0 Median : 36.00   
## Mean :15.4 Mean : 42.98   
## 3rd Qu.:19.0 3rd Qu.: 56.00   
## Max. :25.0 Max. :120.00

#### 문제 2

* 배점: 5점

다음은 로그를 구하는 함수의 log()의 사용 예제입니다. 그런데 이 knitr 청크를 경고 메시지가 발생합니다. 그런데 이 경고 메세지가 보기 좋지 않아서 문서에서 제거하고 싶습니다. 청크의 옵션을 추가해서 경고 메시지를 제거하십시요.

log(-1:3)

## [1] NaN -Inf 0.0000000 0.6931472 1.0986123

### 마크다운의 이해

#### 문제 3

* 배점: 5점

마크다운 태그를 이용해서 다음 문장을 볼드체로 출력하세요. (2주차. “마크다운 태그 이해하기 > 글자 형식 태그의 이해” 참고)

**이 문장은 중요한 내용입니다. 그러니 볼드체로 출력합니다.**

#### 문제 4

* 배점: 5점

다음 “타이틀 5”는 제목을 나타내는 타이틀입니다. header 5를 적용하세요. (2주차. “마크다운 태그 이해하기 > 문장 서식 태그의 이해” 참고)

##### 타이틀 5

#### 문제 5

* 배점: 10점

아래에 “텍스트 정보처리와 NLP”라는 텍스트에 강의노트가 있는 URL인 “[https://choonghyunryu.github.io/NLP\_TA/"를](https://choonghyunryu.github.io/NLP_TA/%22를) 링크를 겁니다. (2주차.”마크다운 태그 이해하기 > 이미지와 링크 삽입 태그의 이해” 참고)

[텍스트 정보처리와 NLP](https://choonghyunryu.github.io/NLP_TA/)

## 텍스트 데이터 수집하기 및 형태소분석

### bitTA와 tidyverse의 이해

#### 문제 6

* 배점: 5점

bitTA의 buzz 데이터에는 맘스홀릭과 레몬테라스 게시물의 일부가 포함되어 있습니다.

**맘스홀릭과 레몬테라스의 게시물이 각각 몇건이지 알 수 있도록 다음 청크를 완성하세요.**

1. 주석(#)을 제거한 후, 괄호안을 채워서 문제를 해결하세요.
2. 만약 문제를 해결하지 못했을 경우에는 주석(#)을 제거하지 마십시요.

library(tidyverse)  
library(bitTA)  
  
buzz %>%   
 group\_by(SECTION) %>%   
 tally()

## # A tibble: 4 × 2  
## SECTION n  
## <chr> <int>  
## 1 레몬테라스 211  
## 2 맘스홀릭 632  
## 3 월급쟁이재테크연구카페 135  
## 4 흥부야재테크하자 22

### 텍스트 데이터 수집과 형태소 분석

#### 문제 7

* 배점: 45점

다음 지시사항을 수행하기 위한 R 코드를 완성하세요. (4주차. “마크다운 태그 이해하기 > 데이터 수집 시나리오” 참고) (6주차. “형태소분석과 감성분석 > 형태소분석과 감성분석 실습” 참고)

* 만약 문제를 해결하지 못했을 경우에는, 에러가 발생하는 곳에 주석(#)을 삽입하여 다음문제를 진행하세.

1. koscrap, bitTA, tidytext, tidyverse 패키지를 사용합니다. (5점)
2. “예금 금리” 키워드로 100건의 뉴스를 검색해옵니다.
   * 유사도 정렬 기준으로 수집합니다. (5점)
   * 결과는 news\_deposit\_rate에 저장합니다. (5점)
3. 수집한 데이터의 description\_text 변수를 일반명사 형태소분석 토큰화합니다. (15점)
   * 입력변수 : description\_text
   * 출력변수 : terms
   * 이 토큰화한 결과를 docs\_term에 할당하세요.
4. docs\_term의 변수 terms의 빈도를 구하되 상위 10개 명사를 추출하세요. (15점)

# 1.  
library(koscrap)  
library(bitTA)  
library(tidytext)  
library(tidyverse)  
  
# Naver 뉴스 API 인증키  
client\_id <- "d20X1HTRRYSrYFd3PPkM"  
client\_secret <- "eAnNTHdYhA"  
  
# 검색 키워드  
keyword <- "예금금리"  
  
# 2.  
news\_deposit\_rate <- search\_naver(keyword, client\_id = client\_id, client\_secret = client\_secret, sort = "sim",  
 do\_done = TRUE, max\_record = 100)

## \* 검색된 총 기사 건수는 161877건입니다.  
## - (100/100)건 호출을 진행합니다.

# 3.  
docs\_term <- news\_deposit\_rate %>%   
 tidytext::unnest\_tokens(  
 output = "terms",  
 input = "description\_text",  
 token = morpho\_mecab  
 )   
  
# 4.  
docs\_term %>%   
 group\_by(terms) %>%   
 tally() %>%   
 arrange(desc(n)) %>%   
 filter(row\_number() <= 10)

## # A tibble: 10 × 2  
## terms n  
## <chr> <int>  
## 1 금리 251  
## 2 은행 189  
## 3 저축 162  
## 4 인상 115  
## 5 상품 91  
## 6 고객 77  
## 7 제공 52  
## 8 혜택 49  
## 9 예금 48  
## 10 관계자 43

## 정규표현식

### 정규표현식의 이해

#### 문제 8

* 배점: 20점

다음 지시사항을 수행하기 위한 R 코드를 완성하세요. (5주차. “정규표현식의 이해 > stringr 패키지를 이용한 문자열 조작” 참고) (5주차. “정규표현식의 이해 > 정규표현식 기반 텍스트 정제 실습” 참고)

1. str\_01을 정제하세요. (10점)
2. str\_02로 다음을 수행하세요.
   1. 텍스트에 공백 포함 여부를 판별하세요. (5점)
      * 공백을 들어가 있는 텍스트는 TRUE, 공백이 없으면 FALSE로 결과값 출력
   2. 부로 시작하는 텍스트를 추출하세요. (5점)
      * grep() 함수를 사용합니다.

library(stringr)  
  
# 1.  
str\_01 <- "분석 방법을\n실습 중심으로 습득함으로써 과학적 텍스트 분석 역량을\n배양합니다."  
str\_01 <- gsub("\n", " ", str\_01)  
str\_01

## [1] "분석 방법을 실습 중심으로 습득함으로써 과학적 텍스트 분석 역량을 배양합니다."

# 2.  
str\_02 <- c(  
 "비비디바비디부 BiBiDi", "케세라세라", "15881234", "똑딱 똑딱 똑똑 딱딱",   
 "묻고\n답하기", "Hello, 뿜빠라뿜빰뿌뿌뿌", "빠라라라라람바 빠라라라라람바",   
 "지국총 지국총 어사와", "아흐다롱디리, 으허허", "63빌딩",   
 "파크원빌딩 현대백화점 6층...", "롯데백화점빌딩 건너편 starbuks",  
 "브루펜", "부루펜", "브루팬", "부르팬")  
  
# 2.1.  
str\_detect(str\_02, "^\\S+$", negate = TRUE)

## [1] TRUE FALSE FALSE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE  
## [13] FALSE FALSE FALSE FALSE

# 2.2.  
grep("^부", str\_02, value=TRUE)

## [1] "부루펜" "부르팬"

#### 추가 문제

* 배점: 추가점수 30점
* 수행한 범위에 따라 부분점수 인정합니다.

강의노트 웹 사이트의 Tidyyverse Hands-on중 “Operate string with stringr”의 URL은 다음과 같습니다.

* <https://choonghyunryu.github.io/NLP_TA/tidyverse/handson-stringr.html>

1. “대표함수 기능” 테이블의 내용을 웹 크롤링으로 수집하세요.
2. 수집한 테이블에서 “함수이름” 변수중에서 앞의 “str\_”을 제거하세요.
3. “반환 데이터 유형”이 “character”인 것만 추출하여,
   * “함수 기능”을 명사로 토큰화한 다음,
   * 가장 많이 빈발하는 단어 3개를 구하세요.

(힌트) - R에서 변수 이름에 공백이 들어 있을 때에는 단어들을 를 이용해서 묶어줍니다. - 예 - "반환 데이터 유형"이라는 변수는 -반환 데이터 유형`으로 표현합니다.

#