R로 배우는 데이터분석 입문 2차 과제

[주의사항]

- 각 문제에 대한 소스코드를 먼저 쓰고, 정답을 기재하는 방식으로 작성할 것
- 소스코드는 평가대상은 아니며, R을 이용해 문제를 풀었는지 확인하기 위한 용도임 (R이 아닌 다른 프로그램으로 푼 경우 0점 처리)
- 부분점수가 없으므로 신중하게 정답을 작성하여 제출할 것
- 유효숫자는 소수 셋째자리(넷째 자리에서 반올림)으로 할 것

숙명여대 전·현직 총장님의 신년사(new year address.csv)를 아래 코드로 불러와서 전처리 한 후 문제에 답하시오. 전처리에서 오류가 발생하면 오답으로 처리되니 신중하게 진행하시오.

```
library(dplyr)
library(stringr)
library(tidytext)
library(tidyr)
library(KoNLP)
raw_NYA <- read.csv("new year address.csv")
NYA <- raw_NYA %>% mutate(value = str_replace_all(value, "[^가-힣]", " "), value = str_squish(value))
NYA_noun <- NYA %>% unnest_tokens(input = value, output = word, token = extractNoun)
frequency_three <- NYA_noun %>% count(president, word) %>% filter(str_count(word) > 1)
```

1. (3점) 세 총장님의 신년사에서 TF-IDF가 가장 높은 단어는 각각 무엇인가?

2. (2점) 장윤금총장님 신년사에서 '디지털'과 '데이터' 두 개 단어가 모두 포함된 문장을 찾아서 각 문장의 첫 번째 단어를 제시하시오.

3. (2점) 황선혜총장님 신년사에서 TF-IDF 상위 10위에 포함되고, 동시에 빈도수 상위 10위에도 포함된 단어는 무엇인가?

4. (2점) 강정애총장님 신년사에서 TF-IDF 상위 10위에 포함되고, 동시에 빈도수 상위 10위에도 포함된 단어는 무엇인가?

5. (3점) '숙명'이나 '교육'과 같이 빈도수가 높은 단어가 TF-IDF 상위에 포함되지 못하는 이유가 무엇인지 한 줄로 설명하시오.

6. (3점) frequency_three를 기준으로 1)president가 황선혜총장님인 경우를 제외하여 frequency_two를 만든 후, 2)tf, idf, tf_idf 변수를 제거하고, 3)names_from으로 president를, values_from으로 n을, NA는 0으로 하는 wide form을 만드시오. 강정애총장님과 장윤금총장님 신년사에 공통적으로 출현하는 단어 중 빈도합계(= 두 총장님 신년사에서 빈도수의 합계)가 두 번째로 높은 단어는 무엇인가?
(힌트: 변수 제거 코드 df$var <- NULL)

7. (2점) 문제6에서 만든 wide form을 기준으로, odds-ratio를 강정애총장님 신년사에서의 단어 비중 대비 장윤금총장님 신년사에서의 단어 비중이라고 정의하자. 상대적으로 강정애총장님 신년사에서 비중이 높은 상위 두 개 단어는 무엇이며, 이들의 odds-ratio는 각각 얼마인가?

8. (3점) (문제7에서 연결됨) 강정애총장님 신년사에서 차지하는 비중과 장윤금총장님 신년사에서 차지하는 비중이 가장 유사한 단어 중에서 빈도수가 가장 높은 단어와 그때 odds-ratio 및 빈도수는 각각 얼마인가?

9. (2점) raw_NYA를 단어 기준으로 토큰화하고, 군산대에서 만든 감정사전을 활용하여 단어별 감정점수를 부여한 후, 감정점수가 2점인 단어가 몇 개나 되는지 확인하시오. (주의! 중복된 단어는 1개로 처리하시오)

10. (3점) 세 총장님 신년사 전체 문장에 대해서 감정점수를 각각 구하시오.