

# 중간과제

홍정하

# 과제 작성 및 제출 시 주의사항

- ▶ 제출기한: 11월17일(수) 23시59분
  - 지각제출 불허
  - 11월18일(수) 수업에서 답안 해설
- ▶ 배점: 총 10점(총 2문제) ⇒ 1번 4점, 2번 6점
- ▶ 과제 제출 형식
  - 과제 게시판에 과제를 해결한 코드의 스크립트 파일(확장자 R) 제출
  - 하나의 스크립트 파일에 과제 1~2번 코드 삽입
- ▶ 과제 해결을 위한 강의 범위
  - 02\_R\_기본사용법.pdf~ 09\_R\_제어문과다중파일처리.pdf
  - 수업 중에 다룬 코드/기능만을 사용할 것.
- ▶ 과제 작성 기간 동안 과제 의도 또는 (첨부된 출력결과에 문제의도와 다른 출력결과가 포함될 수도 있으므로) 제시된 출력결과의 적절성에 대한 질문은 가능하나, 과제해결 또는 힌트를 위한 질문 불가. 그러나 과제 해결과 직접 관련없는 질문은 가능

## ▶ 채점기준: 감점제

- 수업에서 다루지 않은 기능/코드를 사용하면 하나당 -1점
  - 부정행위 방지 차원
  - 복습 차원 ← 배운 내용을 충분히 숙지하고 활용하고 있는가?
- 오류/미해결기능/오분석기능 하나당 -1점
- 코드 라인 수 제한 있음(각 문제 참조). 코드 라인 수가 1줄 초과 시 -1점
  - 1 줄에 입력가능하나 가독성을 위해 여러 라인으로 분리한 코드, 즉, R 콘솔 창에서 첫 머리에 +로 표시되는 코드는 1줄 코드로 인정.
  - 그러나 한 줄에 ; 이용하여 여러 명령코드를 입력하는 것은 불인정
- 마지막 라인의 공백라인 여부는 채점 대상 아님
- 코딩수업에서 인정할 만한 코드로 작성할 것.(거의 수작업에 가까운 코드 등)

## ▶ 부정행위 가담자는 모두 F 처리(타인의 도움 금지)

# 첨부파일

- ▶ 데이터 파일: 각 문제의 입력 파일
  - 1번 과제: Q1 폴더(Q1\_o1.txt, Q1\_o2.txt)
  - 2번 과제: Q2 폴더(a01.txt ~ d11.txt, 총 20개)
  
- ▶ 출력 파일: 각 문제의 출력 정답 파일
  - 1번 과제: Out1.txt
  - 2번 과제: Out2 폴더(A01.txt ~ D11.txt, 총 20개)
  
- ▶ 과제 설명 파일
  - 중간과제.pdf

# 과제 풀이 방법

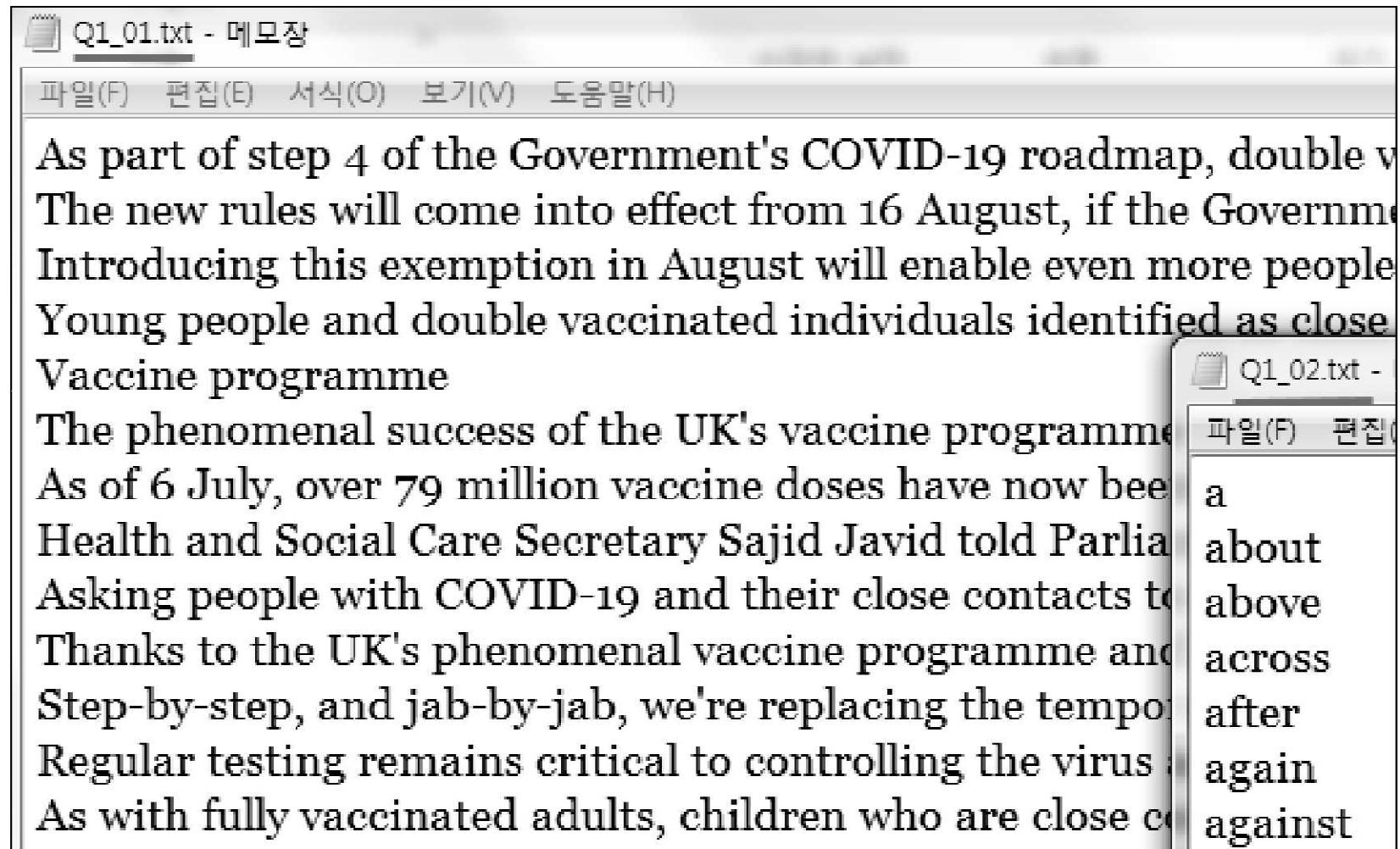
## ▶ 과정

1. 관찰: 각 문제의 입력 파일과 출력 파일 비교
2. 추론: 입력과 출력을 비교 관찰한 내용을 토대로 처리 과정과 필요한 코드 정리
3. 코드 작성: 코드 작성 및 실행 후 출력 파일과 정답 출력 파일을 비교하여 적절성 검토

## ▶ 자기 평가 및 주의사항

- 1차 목표는 과제의 완벽한 해결, 2차 목표는 문제 해결 소요 시간 단축(1번, 2번 문제 해결에 1시간 이상 소요 되었다면 학습 부족 자가 진단. 그러나 프로그래밍 완전 초보자는 그럴 수 있음)
- 가급적 강의자료 또는 웹 검색 도움 없이 문제 풀이 시도, 가급적 웹 검색은 하지 말 것.
- 타인의 도움 또는 의견 절대 금지

# 1번 과제 입력 데이터: Q1 폴더 파일



# 1번 과제: 배점 4점(코드 라인수 $\leq 10$ )

현재 작업폴더 아래에 Q1 폴더를 풀어  
놓고 다음을 수행하는 코드 작성

1. 작업폴더 변경없이 현재 작업폴더 아래 Q1  
폴더의 Q1\_o1.txt, Q2\_o2.txt 파일을 각각 어  
휘 단위 벡터로 불러오기
2. Q1\_o1.txt 파일의 어휘를 소문자 변환 및 어  
휘 양 옆 하나 이상의 문장부호 연쇄 제거
3. Q1\_o1.txt 파일의 어휘 중 Q1\_o2.txt 파일에  
포함된 어휘들의 빈도 추출, 즉, Q1\_o1.txt 파  
일과 Q1\_o2.txt 파일에 공통적으로 분포하는  
어휘의 Q1\_o1.txt 파일에서 빈도
4. 오른쪽 그림과 같은 형식으로 Q1\_out.txt 파  
일 출력(적절한 처리를 했다면 out1.txt과 출  
력결과가 완전히 동일해야 함)



	어휘	빈도
1	a	14
2	across	3
3	after	4
4	against	1
5	all	1
6	almost	1
7	already	1
8	also	1
9	an	1
10	and	19
11	any	1
12	anyone	1

총 97개 어휘

## 2번 과제 입력 데이터: Q2 폴더 파일

### ▶ 마크업언어로 작성된 텍스트 파일 20개

a01.txt  
a05.txt  
a20.txt  
a37.txt  
a41.txt  
a44.txt  
b08.txt  
b16.txt  
b20.txt  
b24.txt

- Q2 폴더 20개 파일 중 일부 파일명
- Q2 폴더 20개 파일 모두 확장자 txt

```
<TEI xmlns="http://www.tei-c.org/ns/1.0"><teiHeader>  
  <fileDesc>  
    <title>"Hartsfield Files"  
    <author>August 17, 1961, "Urged strong."  
    <editor>"Sam Caldwell Joins"  
    <publisher>March 6, 1961, p. 1 "Legislators Are Moving" by  
    <contributor>"Legislator to fight" by Richard Ashworth  
    <date>"House Due Bid..."  
    <page>p. 18 "Harry Miller Wins..."  
  </title></titleStmt><editionStmt><edition>A pa  
  <text xml:id="A01" decls="A">
```

- 입력 파일 본문 예
- TEI 문서



- ▶ <body>...</body> : 본문 영역
- ▶ <p> ... </p>: 하나의 문단 영역
- ▶ <s n="문장번호">... </s>: 하나의 문장 영역
- ▶ < w type="품사" (subtype ="하위분류")>어휘</w>: 하나의 어휘 영역
- ▶ <c type="pct">문장부호</c>: 하나의 문장부호 영역
- ▶ <mw pos="품사">어휘</mw> : 축약형 I'm, doesn't
  
- ▶ 주의:
  - ▶ 반드시 예제 파일을 텍스트파일편집기로 열어보고 파일의 특성을 주의 깊게 관찰할 것.
  - ▶ 어휘 뒤에 스페이스문자가 있는 경우와 없는 경우 모두 혼재

## 2번 과제: 배점 6점(코드 라인수 $\leq 12$ )

- ▶ Q2 폴더를 현재 작업폴더 밑에 압축을 풀어 놓고
  1. 현재 작업폴더 밑에 Q2\_out 폴더를 만들기
  2. 작업폴더 변경없이 현재 작업폴더 밑 Q2 폴더 내 20개 파일에 대해 각 파일을 불러와서 TEI 바디 영역, 즉, `<w ...>`와 `</w>` 사이, `<c ...>`와 `</c>` 사이, `<mw ...>`와 `</mw>` 사이의 어휘부(문장부호, 숫자문자 포함)만 추출하여 하나의 라인에 하나의 어휘를 파일로 출력하되(입력 파일 1당 출력 파일 1개, 출력 파일 개수: 20개, `<w ...>`와 `</w>` 사이, `<c ...>`와 `</c>` 사이, `<mw ...>`와 `</mw>` 사이의 문자열을 하나의 어휘로 처리)
  3. 작업폴더 변경없이 현재 작업폴더 밑 [Q2\_out] 폴더에 파일명 첫 문자를 대문자로 변환하여 출력(예: a01.txt  $\Rightarrow$  A01.txt)
    - 적절한 처리를 했다면 out2 폴더 내 20개 파일과 출력결과 동일(주의: 출력 어휘에 스페이스문자 포함 불가, 어휘 출력으로 공백라인 출력 불가, 단, 마지막 라인의 공백라인 여부는 채점 대상 아님)

```

<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <fileDesc>
    <title>
      "Hartsfield Files"
      August 17, 1961 "Sam Caldwell"
      "Legislators Are Moving to fight" by Richard Ashworth
      "House Due Bid..."
      p.18 "Harry Miller Wins..."
    </title></titleStmt><editionStmt><edition>
      <text xml:id="A01" decls="A">
        <body><p><s n="1"><w type="AT">The</w>
        </p>
        <p><s n="2"><w type="AT">The</w>
        </w> <w type="VBN">conducted</w>
        </p>
        <p><s n="3"><w type="AT">The</w>

```

•입력 파일 예  
•밀줄 친 부분 추출

•출력 파일 예

Jury  
said  
Friday  
an  
investigation  
of  
Atlanta's  
recent  
primary  
election  
produced  
no  
evidence