

정보처리및자연어처리

1차 과제

홍정하

과제 작성 및 제출 시 주의사항

- ▶ 제출기한: 4월14일(수) 23시59분(지각제출 불허)
- ▶ 배점: 총 10점(총 6 문제), 1~2번 각 1점, 3~6번 각 2점
- ▶ 과제를 해결한 코드를 포함하는 jupyter notebook 파일(확장자 ipynb) 또는 txt 파일 또는 py 파일을 과제 게시판에 제출
 - 각 문제를 해결한 코드를 notebook 파일 또는 txt/py 파일 하나에 작성
 - notebook 파일 또는 txt/py 파일에 다음과 같이 첫 줄에 “# 문제번호”를 쓰고 둘째 라인에 해당 번호의 답안 코드, 셋째 라인에 해당 변수 값 출력결과 확인을 위한 변수명 기재



```
# 1번
a = 3 + 5
a

8
```

1번 답안 코드

- ipynb는 문제별 출력결과도 제출, txt, py 파일은 문제번호와 코드만 제출

- ▶ 강의에서 다루지 않은 코드/기능 사용 가능
- ▶ 각 문제의 코드 라인 제한수, 주의사항을 유의하여 코드 작성
- ▶ 각 문제마다 활용하는 변수명과 생성하는 변수명이 다르니, 문제별 변수명에 주의할 것.
- ▶ 1, 2, 3번 제한 코드 라인수 준수
 - 한 줄 코드로 입력 가능하나 가독성을 위해 jupyter notebook에서 여러 라인으로 분할하여 제시하는 경우 한 줄 코드로 인정
 - ;을 이용하여 하나의 라인에 두 개 이상의 명령코드 사용 금지
 - 단, import 라인 및 변수 값 출력결과 확인을 위한 라인은 제한 코드 라인수에 포함되지 않음

▶ 채점기준

- 각 문제에서 제한 코드 라인수를 초과하거나 금지하는 기능/코드를 사용하면 해당 문제는 오답 처리
- 오류/미처리 기능 하나당 -1점
- 코딩수업에서 인정할 만한 코드로 작성할 것.(출력결과를 목록으로 만들어 출력하거나 거의 수작업에 가까운 코드 등)

▶ 과제 작성 기간 동안 과제 의도 또는 제시된 출력결과의 적절성에 대한 질문은 가능하나, 과제해결 또는 힌트를 위한 어떤 질문도 불가.

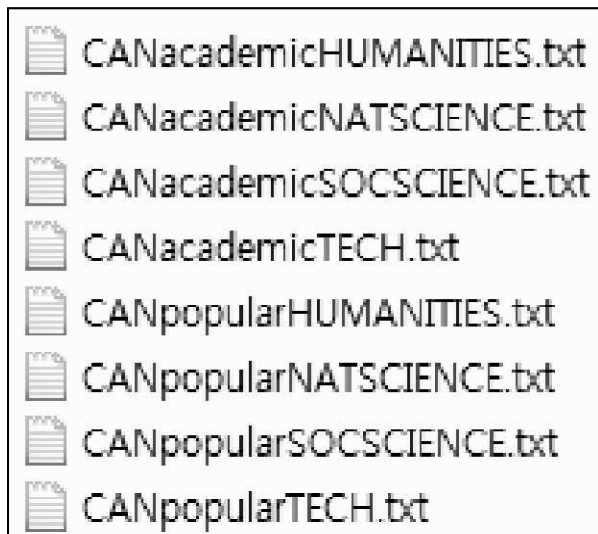
▶ 부정행위 가담자는 모두 F 처리

첨부파일 안내

- ▶ 자연어처리_1차과제.pdf ⇐ 지침 및 문제
- ▶ Q1 폴더 ⇐ 1 번 문제 데이터 파일
- ▶ Q2 폴더 ⇐ 2 번 문제 데이터 파일

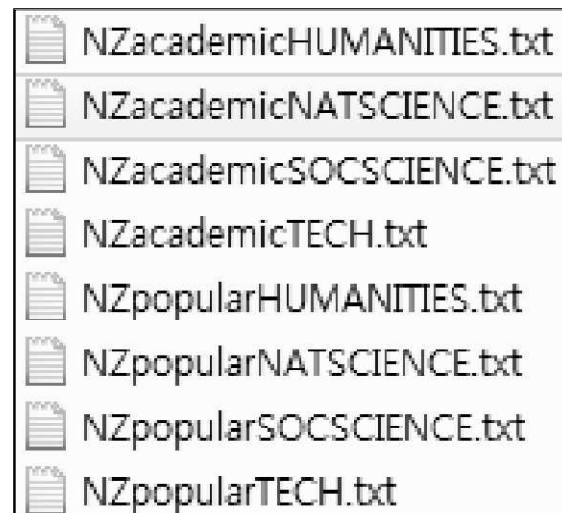
데이터파일: Q1 폴더

현재 작업폴더 밑에 Q1 폴더를 만드시오. Q1 폴더 내의 16개 파일을 이용하여 다음 슬라이드의 1번 문제를 해결하시오.



A list of 8 text files, each preceded by a document icon. The files are:

- CANacademicHUMANITIES.txt
- CANacademicNATSCIENCE.txt
- CANacademicSOCSCIENCE.txt
- CANacademicTECH.txt
- CANpopularHUMANITIES.txt
- CANpopularNATSCIENCE.txt
- CANpopularSOCSCIENCE.txt
- CANpopularTECH.txt



A list of 8 text files, each preceded by a document icon. The files are:

- NZacademicHUMANITIES.txt
- NZacademicNATSCIENCE.txt
- NZacademicSOCSCIENCE.txt
- NZacademicTECH.txt
- NZpopularHUMANITIES.txt
- NZpopularNATSCIENCE.txt
- NZpopularSOCSCIENCE.txt
- NZpopularTECH.txt

1번: 배점 1점

현재 작업 폴더 밑 Q1 폴더의 16개 파일을 불러와서 다음의 결과를 얻을 수 있도록 변수 a를 있도록 한 줄 코드로 만드시오. (주의: categories의 출력값을 Q1 폴더 파일명으로부터 추출해야 함.)

```
a.categories()
```

```
['HUMANITIES', 'NATSCIENCE', 'SOCSCIENCE', 'TECH']
```

```
len(a.fileids())
```

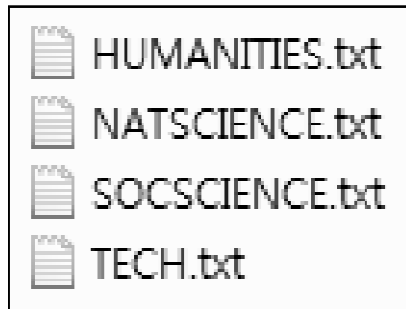
```
16
```

```
print(a.fileids())
```

```
['CANacademicHUMANITIES.txt', 'CANacademicNATSCIENCE.txt', 'CANacademicSOCSCIENCE.txt', 'CANacademicTECH.txt', 'CANpopularHUMANITIES.txt', 'CANpopularNATSCIENCE.txt', 'CANpopularSOCSCIENCE.txt', 'CANpopularTECH.txt', 'NZacademicHUMANITIES.txt', 'NZacademicNATSCIENCE.txt', 'NZacademicSOCSCIENCE.txt', 'NZacademicTECH.txt', 'NZpopularHUMANITIES.txt', 'NZpopularNATSCIENCE.txt', 'NZpopularSOCSCIENCE.txt', 'NZpopularTECH.txt']
```

데이터파일: Q2.zip

현재 작업폴더 밑에 Q2 폴더를 만드시오. Q2 폴더 내의 4개 파일을 이용하여 다음 슬라이드의 2번 문제를 해결하시오.



2번: 배점 1점

현재 작업 폴더 및 Q2 폴더의 4개 파일을 불러와서 다음의 결과를 얻을 수 있도록 변수 b를 구성하도록 한 줄 코드로 만드시오.

```
b.fileids()
```

```
['HUMANITIES.txt', 'NATSCIENCE.txt', 'SOCSCIENCE.txt', 'TECH.txt']
```

3번: 배점 2점

2번의 변수 b와 for 구문을 이용하여 다음을 수행(코드라인 수 제한 없음)

- 파일별 문자수 = 공백문자 및 문장부호 제외 문자수
- 파일별 어휘수 = 알파벳 또는 숫자문자 하나 이상 포함 어휘수
- 아래의 서식을 참고하여 함수 print를 이용하여 출력

파일명: HUMANITIES.txt	문자수: 530,867	어휘수: 110,177
파일명: NATSCIENCE.txt	문자수: 538,039	어휘수: 110,008
파일명: SOCSCIENCE.txt	문자수: 566,873	어휘수: 115,451
파일명: <u>TECH.txt</u>	문자수: <u>530,235</u>	어휘수: <u>106,654</u>

문자 15 칸
왼쪽정렬

문자 10 칸
왼쪽정렬
3자리 쉼마

문자 10 칸
왼쪽정렬
3자리 쉼마

4번: 배점 2점

2번의 변수 `b`를 이용하여 빈도 3,000 이상의 문장부호를 빈도 내림차순으로 정렬하여 다음과 같은 변수 `c`를 출력하는 한 줄 코드를 작성하시오.

<code>len(c)</code>
7
<code>print(c)</code>
<code>[(',', 22634), ('.', 21953), ('-', 6226), ('"', 4888), (')', 3703), ('(', 3694), ('"', 3095)]</code>

5번: 배점 2점

2번의 변수 b, 4번의 변수 c, nltk의 tabulate 함수를 이용하여 다음과 같은 파일별 변수 c의 문장부호 빈도교차표를 출력하기 위한 코드를 제시하시오.(코드 라인 수 제한 없음)

	,	.	-	')	("
HUMANITIES.txt	5770	5087	1222	1640	584	584	877
NATSCIENCE.txt	6016	6091	1567	674	1515	1515	427
SOCSCIENCE.txt	6078	5674	1706	1845	701	701	1042
TECH.txt	4770	5101	1731	729	903	894	749

6번: 배점 2점

2번의 변수 b와 for 구문, nltk similar을 이용하여 다음을 수행(코드라인 수 제한 없음)

- 파일별 'evidence'와 분포 맥락이 어휘 목록을 추출
- 아래와 같이 파일명을 print 함수를 이용하여 출력한 후 nltk similar를 이용하여 출력

```
HUMANITIES.txt
```

```
one those time land life workshops islands flooding way that other  
death nature signs some them works loss most definition
```

```
NATSCIENCE.txt
```

```
way need two implications absence so difficult argue year application  
data same form soil surface zone work association material consisting
```

```
SOCSCIENCE.txt
```

```
one time part members which areas aware much discussion method crucial  
discounting history the of measure and over another not
```

```
TECH.txt
```

```
is most possible selection use all those some but uncertainty into one  
i follows assumed be which mentioned many such
```