1. “1.txt” 파일을 읽어서 각 글자(알파벳)별로 나타나는 횟수를 계산하여, 가장 많이 나오는 글자 10개를 순서대로 histogram을 생성하세요(x 축 : 글자, y 축 : 출현횟수).

* 제출 : 소스 코드
* 배점 : 10점

|  |
| --- |
| “2.txt” 파일은 타이타닉호의 승객 정보와 생존여부에 관한 데이터이다. 이 데이터를 이용하여 승객의 생존여부(**survived)**를 예측 하는 예측모델을 생성하려고 한다. ( 2번 ~ 4번) |

1. 데이터의 결측치(Missing Value)를 찾아내어 적절한 조치를 취하세요.

- 제출 : 소스 코드 및 해설(해설은 소스코드 맨 아래 주석으로 작성)

- 배점 : 15점

1. 예측 모델을 생성하기 위한 변수를 선택하고 그 이유를 설명하세요.

- 제출 : 소스 코드 및 해설(해설은 소스코드 맨 아래 주석으로 작성)

- 배점 : 15점

1. 예측 모델을 생성하기 위하여 가장 알맞게 데이터를 7:3(training data : test data) 비율로 나누세요.

- 제출 : 소스 코드

- 배점 : 30점

|  |
| --- |
| “cardusages.csv” , “usagestype.csv” 카드 사용에 관한 데이터이다. (5번 ~ 6번)  cardusages.csv 컬럼 설명   * date : 사용일자 * card\_number : 카드번호 * type : 사용처 분류 Index * amount : 사용량   usagestype.csv 컬럼 설명   * Index : 사용처 분류 Index * type : 사용처 분류 이름 |

1. Entertainment 로 가장 많이 사용한 월(1월 ~ 12월)과 Transportation 으로 가장 많이 사용한 일(1일 ~ 31일)을 구하세요.

- 제출 : 소스 코드

- 배점 : 15점

1. 각 사용처별 사용량을 꺽은선 그래프(line plot)로 그리고, 각 사용처 분류 이름을 범례로 표시하세요. ( x 축 : 사용일, y 축 : 사용량)

- 제출 : 소스 코드

- 배점 : 15점