머신러닝 과제

마감일: 6월 14일 수업시간 시작전
제출방법: 이메일로 제출. hoyoung23@uos.ac.kr
제출내용: 코드와 과제문서를 이메일에 첨부하여 제출.

비고사항: 선제출에 대한 가산점은 없으나 제시된 예측율 이상 달성하는 경우
1번에 대해 1.5점 2번에 대해 1.5점 등 최대 3점 총점에 가산점 부여.

제출이 늦은 경우, 최초 2시간까지 감점 3점,
이후 6시간 늦을때마다 5점씩 감점하여
(즉, 마감시간보다 8시간 늦으면 8점감점)
30시간이 경과한 이후 제출은 0점 처리.

데이터는 수업 웹페이지에서 다운받아,
> load("~~ 본인의 저장 디렉토리 \\homeworkdata.RData")로 읽어들일 것.
> 1번과 2번을 풀기 위한 데이터는 wordmatrix 라는 object로 저장되어 있으며
3번을 풀기 위한 데이터는 we_data 라는 object로 저장되어 있음.

해당 데이터는 출판진행중으로 현재시점에서는
외부공개불가로 과제에만 사용할 것.

주어진 데이터는 자연어 사전처리를 대략적으로 끝낸 DTM (Document-Term Matrix)를 벡터화한 자료로, 국정원의 인터넷 게시판 활동 518개의 댓글에 대해 7589개의 단어 범주의 존재 여부를 표기한 매트릭스 자료이다. 다시 말해 해당 자료의 형태를 개념적으로 그려보면, 아래와 같다.

	분류	단어1	단어2	…
댓글1	1	1	0	
댓글2	0	1	1	
….				
댓글3	1	0	0	

원래 데이터에는 댓글1, 댓글2로 되어 있지 않은 단순한 행의 구분만 있을 뿐이다. 분류는 2가지로 댓글을 분류한 표기인데, 1은 경우는 이명박 대통령 및 당시 여당지지 이고 0은 전직 대통령 및 야당 비난임을 검찰이 범죄일람표에 표기한 내용이다. 분류에 해당하는 변수명은 "cat"으로 되어 있으며 나머지는 실제 한글 단어이다. 본 과제는 해당 매트릭스를 통해 댓글을 분류하는 작업에 대한 과제이다.

1. 데이터를 70%는 학습자료로 만들고 30%는 테스트 자료로 구성하여

1) 중간 고사 이전에 배운 방법론 중 한 가지를 활용하여,
댓글 분류에 대한 예측율을 적어도 약 89%이상 달성하라. (5점)

2) 1)번에서 선택한 방법에 대한 K-Fold 교차 검증법을 실시하여
예측율을 구할 것. (예측율이 상승해야). (5점)

3) 해당 방법의 문제점은 없는지 서술하시오 (3점)

2. 다층 퍼셉트론 딥러닝(a.k.a. 기본 딥러닝 모형)으로 댓글 문서 분류를 실시하여,

1) 예측율을 적어도 약 97% 달성할 것 (5점).

2) 과대적합이 나타나서 훈련데이터에 대한 학습을 테스트 데이터에 적용하면 약 80% 정도로 나타난다. 해당 내용을 보여줄 것 (3점),

3. GloVe 단어 임베딩 방법을 활용하여,

1) 사업+대통령을 조합하였을 때 코사인 유사도가 가장 높은 단어들 10개를
제시하시오(2점)