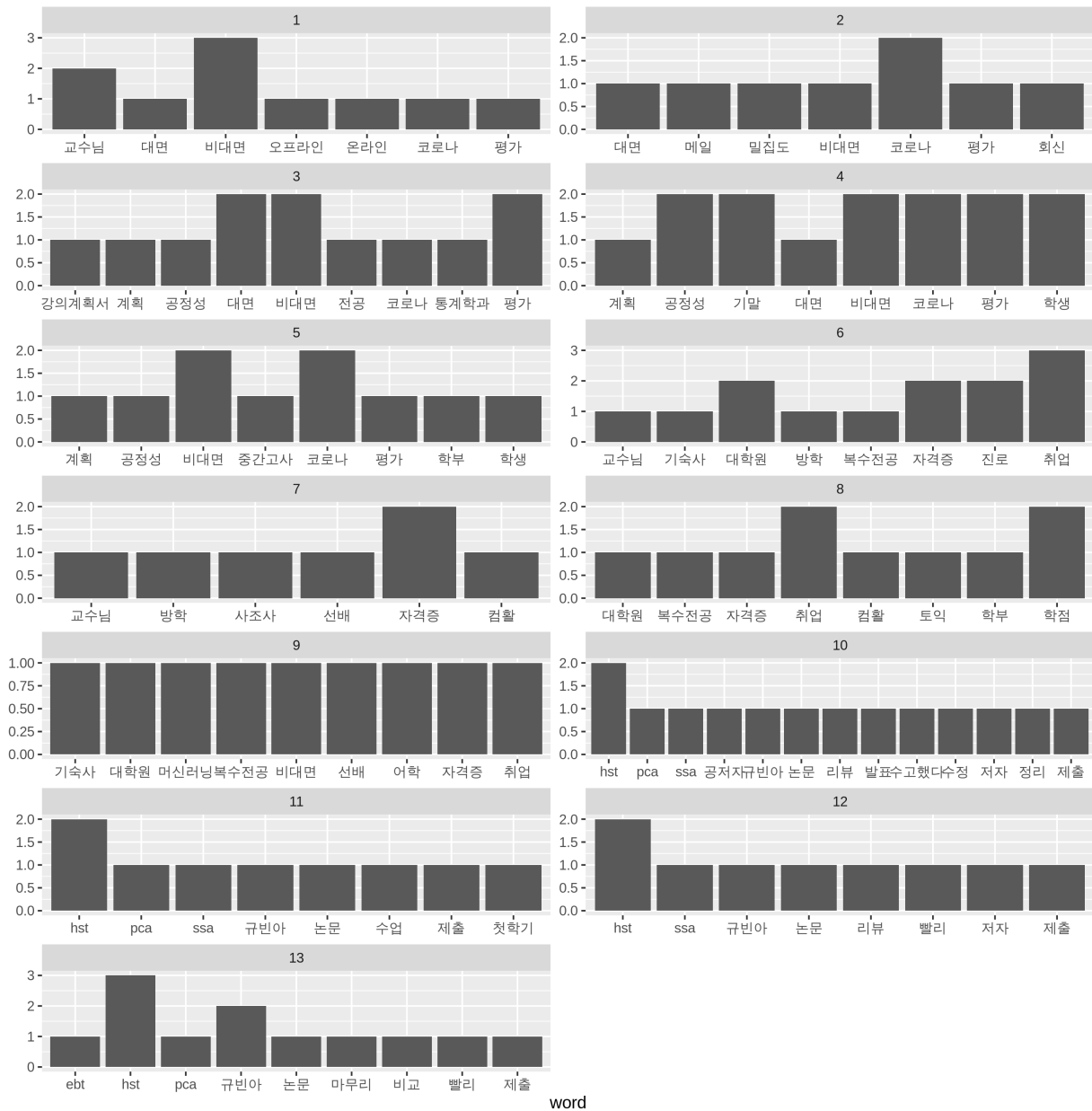


통계계산 기말고사

문제상황

아래는 A씨가 수신한 13통의 이메일에 포함된 단어를 시각화 한 것이다.



문제

토픽모형을 활용하여 A씨가 수신한 13통에 이메일의 단어들이 어떠한 토픽으로부터 생성되었는지 추론하라.

Note

- A씨가 수신한 이메일의 단어들은 LMS에 업로드 된 “data.csv”에 저장되어 있다.
- data.csv의 1열은 단어, 2열은 이메일의 인덱스 3열은 (임의로 부여된) 토픽을 의미한다.
- 토픽모델링을 통하여 적절한 토픽을 추론하고, 분석결과를 “학번.csv”의 형태로 저장하여 분석에 사용한 코드와 함께 제출한다.

힌트

A씨는 최근 크게 3가지 종류의 메일을 수신하였다.

- 토픽1: 학부 중간/기말평가 시행 안내 및 평가 계획 제출 요청
- 토픽2: 평생지도교수 및 전공진로설계 학생상담메일
- 토픽3: 박사과정 지도교수님으로부터 수신한 메일

메일에 포함된 단어를 파악하니 메일1-5는 토픽1에, 메일6-9는 토픽2에 메일10-13은 토픽3에 해당하는 것 처럼 보인다. 시뮬레이션을 통하여 해당 메일들의 단어가 적절한 토픽으로 올바르게 묶였는지 스스로 체크해보라.