**4조 발표 구성**

**개 요**

**학 습 목 표 : 각 장의 프로그래밍 활용**

**데이터 출처 : 네이버 뉴스 중 각 분야 별 일일 조회수 / 댓글 랭킹 30, 각 뉴스들의 댓글**

**데이터 컬럼 : 뉴스(랭크, 제목, 부제, 회사), 뉴스 댓글(내용, 갯수, 글자수, 작성자 삭제/규정 미준수 댓글 갯수, 성비, 각 나이대 별 비율)**

활용 목록

|  |  |  |  |
| --- | --- | --- | --- |
| 순 차 | 내 용 | 순 차 | 내 용 |
| 1 | Database | 7 | 연관 분석 |
| 2 | 워드 클라우드 | 8 | 군집 분석 |
| 3 | 교차 분석과 카이 제곱 검정 | 9 | KNN |
| 4 | 집단 간 차이 검정 | 10 | Naïve Bayes |
| 5 | 의사 결정 트리 | 11 | SVM |
| 6 | 인공 신경망 |  |  |

**[1] Database**

**목표 : Oracle 연결 및 크롤링 데이터 저장**

활용 함수

|  |  |  |
| --- | --- | --- |
| 함 수 | 사용 컬럼 | 예상 결과 |
| JDBC |  |  |
| dbConnect |  |  |
| dbGetQuery |  |  |

**[2] 워드 클라우드**

**목표 : 정치 분야 뉴스 제목들의 단어로 워드 클라우드 만들기**

활용 함수

|  |  |
| --- | --- |
| 항 목 | 내 용 |
| library | KoNLP, ggplot2, stringr, RColorBrewer, wordcloud, wordcloud2 |
| 함수 | readLine, sapply, str\_replace, gsub, Filter, sort, ggplot, wordcloud |

**[3] 교차 분석과 카이 제곱 검정**

**목표 : ‘나이대’, ‘선호 분야’ 간의 관련성 여부 확인**

활용 함수

|  |  |
| --- | --- |
| 항 목 | 내 용 |
| 함수 | table(교차분석표), chisq.test |
| 기타 | 리코딩(색인 활용) |

**[4] 집단 간 차이 검정**

**목표 : 분야1과 분야2 뉴스 간 댓글 평균 글자수 차이 확인**

활용 함수

|  |  |
| --- | --- |
| 항 목 | 내 용 |
| library | stats |
| 함수 | shapiro.text, qqnorm, qqline, t.test, qt, var.test, Bartlett.test, TukeyHSD |

**[5] 의사 결정 트리**

**목표 : 어느 나이대이면 어느 분야의 뉴스를 보게 되는지 확인**

**(특정 단어로 검색하면 어떤 분야의 뉴스가 검색될지 확인)**

활용 함수

|  |  |
| --- | --- |
| 항 목 | 내 용 |
| library | rpart, party |
| 함수 | function활용, rpart, ctree, predict |

**[6] 인공 신경망**

**목표 : 각 분야 별 뉴스는 주로 어떤 단어들이 포함되어 있는지 확인**

활용 함수

|  |  |
| --- | --- |
| 항 목 | 내 용 |
| library | nnet, devtools, neuralnet |
| 함수 | nnet, neuralnet, compute |

**[7] 연관 분석**

**목표 : 특정 분야 뉴스 안에 특정 단어1이 있으면 특정 단어2가 같이 있을 확률 구하기**

활용 함수

|  |  |
| --- | --- |
| 항 목 | 내 용 |
| library | arules, arulesViz |
| 함수 | apriori, inpect, sinkm, plot활용 |

**[8] 군집 분석**

**목표 : 특정 분야의 뉴스에서, 세부 주제에 따라 나이대 추측**

활용 함수

|  |  |
| --- | --- |
| 항 목 | 내 용 |
| library | cluster, corrgram |
| 함수 | dist, hclust, cmdscale, t, kmeans, plot활용, rect.hclust, corrgram |

**[9] KNN**

**목표 : 기사 제목을 놓고, 어느 분야에 가까운지 KNN 그래프로 보기**

활용 함수

|  |  |
| --- | --- |
| 항 목 | 내 용 |
| library | class |
| 함수 | ceiling, sqrt, knn |

**[10] Naïve Bayes**

**목표 : 뉴스 댓글들 중 네이버 클린봇이 필터 못한 광고/욕설 등 필터하기**

활용 함수

|  |  |
| --- | --- |
| 항 목 | 내 용 |
| library | tm |
| 함수 | Corpus, tm\_map, findFreqTerms, DocumentTerMatrix, naiveBayes |

**[11] SVM**

**목표 : 나이대 / 시간대 실시간 데이터 저장 및 정치 분야 산점도 그래프 그리기**

활용 함수

|  |  |
| --- | --- |
| 항 목 | 내 용 |
| library | kernlab |
| 함수 | ksvm, prop.table, ggplot, predict |