**<데이터 애널리틱스 중간 실습 정리>**

1. **R 기초**

* R은 데이터 분석을 위한 통계분석 기법과 알고리즘, 시각화 기능을 지원하는 오픈 소프트웨어 도구
* R은 패키지를 이용, R CRAN사이트를 통해서 1만 5천여개의 패키지를 자유롭게 다운로드하고 이용가능 함
* 1:1 상호작용방법으로 진행하거나, 일정 구간을 지정해 일괄 실행 방식으로 처리할 수 있음
* R스튜디오는 R 콘솔에 비해 편리한 IDE라는 통합 분석 개발환경을 제공함
* getwd() : 현재 디렉토리 확인

1. **데이터 구조 & 데이터 분석 기초**

* 거듭제곱: ^또는\*\*, 나머지: %%, 몫: %/%
* 행렬(2차원 배열) -> matrix사용, ‘nrow’ = 행의 수
* 리스트 -> 벡터의 하나 이상의 각 원소
* 데이터 프레임

1. C바인드 -> 열 추가
2. R바인드 -> 행 추가(dataframe함수 사용)

* 벡터

1. R에서 가장 작은 데이터 단위
2. 실수형, 논리형(TF), 문자형,

* 팩터

1. 명목형 변수를 대표하기 위해 사용되는 특별한 용도의 벡터(범주형)
2. “Levels=“구문을 통하여 범주 명시 가능

* 매트릭스 -> 데이터프레임과 다르게 한가지 유형만으로 구성, 수학적 연산에 주로 사용
* save함수 -> 변수 저장, load함수 -> 변수 출력
* write.csv -> csv파일 저장, read.csv -> csv파일 불러오기
* ‘stringsAsFactors = FALSE’ -> 팩터 구조X
* str(변수명) -> 데이터 프레임의 구조나 벡터, 리스트를 포함한 R 데이터 구조를 표시
* summary(변수명) -> 일반적인 요약 통계
* range() -> 최대값과 최소값의 공간
* diff(range()) -> 최소최대 차이
* IQR

1. Q3 – Q1으로 3사분위수에서 1사분위수를 뺀 사분위수 범위를 의미
2. Q3+1.5 IQR 이상의 값을 보통 의심되는 이상값 혹은 극단값이라 함

* boxplot, hist -> main, xlab, ylab을 설정할 때는 +로
* table -> 일원배치
* round -> 반올림
* 함수 -> 하나 이상의 명령어로 반복 사용 가능한 기능을 구현한 것

1. **K평균 군집분석**

* ‘NA’ -> 결측치
* 결측값을 포함할 수 있도록 ifany 작성 -> table(teens$gender, useNA = "ifany")
* 더미코딩 -> 남녀와 같이 우선순위가 없는 명목형 변수일 때 사용
* !is.na(teens$gender) -> teens$gender에서 NA가 아닌 것
* na.rm = TRUE -> 결측값을 포함하지 않고 계산
* aggregate(data = teens, age ~ gradyear, mean, na.rm = TRUE)
* 졸업년도별 나이의 평균
* ave(teens$age, teens$gradyear, FUN = function(x) mean(x, na.rm = TRUE))
* 졸업년도와 나이를 활용해 개인에 대한 예측된 나이 계산
* as.data.frame -> 데이터프레임으로 변환, lapply -> 리스트화, scale -> 정규화

1. **계층적 군집분석**

* row.names -> 행의 이름을 변경
* sapply -> 결과를 벡터로 반환
* dist(utilities.df.norm[,c(6,8)] -> 6열과 8열의 유사도를 분석
* dist(utilities.df.norm, method = "euclidean") -> 8개 변수의 유사도를 모두 분석
* single(단일): 최단거리, complete(완전): 최장거리, average(평균): 각 개체마다 모두 계산 후 평균, median(중심): 그룹별 중앙점 거리
* hang = -1 -> 그림을 보기 좋게
* ann = FALSE -> 주석X로 축에 이름이 없음
* rect.hclust(hc2, k=6) -> 6개의 군집으로 분류하여 표시
* Colv = NA -> row로 분석함
* col = rev -> 어두울수록 높은 값 배정

1. **자기조직화 지도**

* rlen -> 학습횟수
* alpha -> 학습률 ex) c(0.3, 0.01) 0.3부터 진행할 때마다 0.01씩 감소
* type = "changes" => 학습그래프(Training Progress) 파악
* type = "counts” => 학습된 som 모델의 각 뉴런이 몇 개의 학습 데이터와 맵핑이 되는지를 그림으로 확인. 이를 통해 som 모델의 질(quality)을 평가
* type = "dist.neighbours" => 노드 간 유사도 확인
* type = "codes" => 각 뉴런에 대한 학습 데이터 가중치 기여율 확인

1. **웹 크롤링**

* ASIN => 아마존이 만든 10자리 고유 식별 번호
* paste0 => 문자열 결합 함수
* 파이프 연산자 " %>%, %<% "

1. Ex) A %>% B => A ) B , 즉 A가 더 큰 개념
2. 화살표 순서대로 처리하면 됨
3. 함수가 많아질수록 더 유용해짐

* html\_nodes => html문서에서 노드를 찾는 함수
* html\_text() => html에서 해당하는 텍스트
* tibble => 열 이름만 적어서 간편하게 Dataframe 생성 가능
* do.call

1. list의 i번째 큰 방에 벡터를 넣은 후
2. for문이 끝난 후 list안에 있는 큰 방들에 대해 rbind를 시킴
3. 데이터가 커질수록 유용해짐

* gsub => 텍스트를 지우거나 변경하는 함수

gsub("변경하고 싶은 텍스트", "어떻게 변경할지", 변경할 변수명)

* strsplit => 문장을 나누는 함수
* pattern => 문자열을 다룰 때 사용

1. **TF-IDF**

* 벡터크기 => 루트(각 항목 제곱의 합)
* 검색문과 각 문서간 내적 => 각 인수끼리 곱한 후 더함
* 코사인 유사도 => 내적/검색문 벡터크기\*문서 벡터크기

1. **텍스트 마이닝**

* iconv => 문자 벡터 인코딩 변환 함수
* iconv(enc2utf8(amazon$Review) => amazon의 Review를 UTF-8로 변환
* iconv(ecodot, from="latin1", to="ASCII", sub="") => 라틴어를 영어로
* latin1 => 프랑스어, 독일어, 아이슬랜드어, 스페인어 문자
* ASCII => 미국인 대상의 표준 영어 문자
* Vcorpus => text -> corpus(말뭉치)로 변환하는 함수
* Sys.setlocale() => 현재의 언어 상태 확인
* Sys.setlocale(category = "LC\_ALL", locale = "us") => 한국어를 영어로
* TermDocumentMatrix

1. removePunctuation=T => 모든 문장부호 제거
2. stopwords= "SMART" => 불용어 제거  
   ‘텍스트에 빈번하게 나타나 중요하게 인지될 수 있지만, 실질적으로는 분석에 중요한 단어가 아니므로 제거해줌’
3. tolower = T => 모두 소문자로 변환
4. removeNumbers = T => 숫자 제거
5. wordLength = c(5,5) => 단어길이 설정, ex) 3으로만 설정하려면 c(3,3)
6. stemming = F => 어간 추출X, 에러 뜰 가능성 있음
7. stripWhitespace = T => 두개 이상 공백을 하나의 공백으로

* txt\_mat <- as.textmatrix(as.matrix(tdm))   
  # tdm의 list형태 -> matrix -> textmatrix
* lsa(txt\_mat, dim=10)   
  # txt\_mat을 10개의 차원으로 LSA 만들기
* LSA (Latent Semantic Analysis) 잠재 의미 분석
* tk, dk, sk => 잠재 의미 공간들