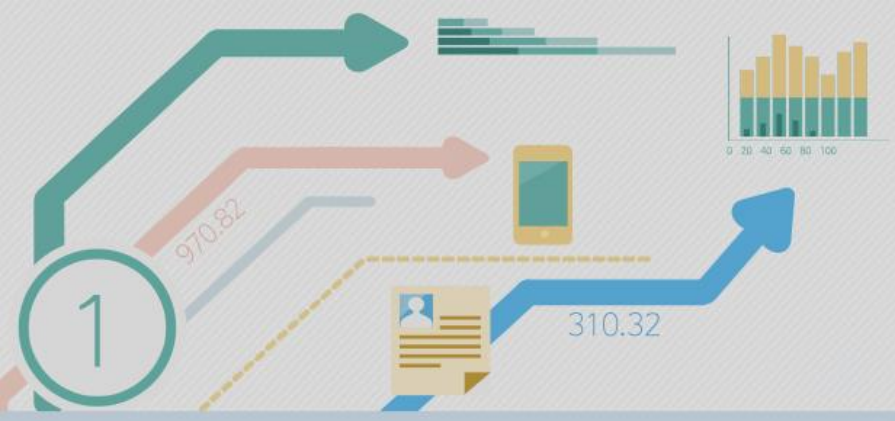


빅데이터 분석 결과 시각화

영화 리뷰 분석 시각화

한국기술교육대학교
온라인평생교육원

학습내용

- 비정형 데이터
- ‘N사’ 영화 리뷰 분석
- ‘D사’ 영화 리뷰 분석 실습

학습목표

- 비정형 데이터를 분석할 수 있는 코딩을 할 수 있다.
- 영화 리뷰 분석하는 방법을 설명할 수 있다.
- RStudio를 통해 직접 영화 리뷰 분석을 할 수 있다.

비정형 데이터

1. 비정형 데이터 정의

◆ 비정형 데이터란?

- 비정형 데이터
  - 문서, 그림, 음성, 영상처럼 구조화되지 않은 데이터
  - 형태와 구조가 다양한 정보

◆ 비정형 데이터의 탄생

① 문서나 음성 정보 등을 숫자로 변환하는 방법으로 설계 및 분석됨

② 점점 방대해지고 의미의 연관성이 높아지는 정보

③ 컴퓨터 처리 방식으로 해석하기 어려워짐

◆ 비정형 데이터의 분석

- 텍스트 마이닝
  - 자연어 처리 방식을 이용하여 정보를 추출하는 기법
  - 정보 검색 - 추출 - 체계화 - 분석의 과정
  - 텍스트 마이닝 관련 사이트

- 오피니언 마이닝
  - 특정 주제나 대상에 대한 사람들의 주관적이고 감정적인 의견을 분석하여 선호도 판별

2. R 스크립트 용어 정리

◆ 작업용 기본 디렉터리(Working Directory)

● 실제로 C드라이브에 R폴더를 만들고 그 디렉터리를 작업용 디렉터리로 만들어 파일을 불러오거나 저장할 수 있음

```
> setwd("c:₩₩R")          C 드라이브에 R폴더를 작업 디렉터리로 설정함
> getwd( )                현재 설정된 작업 디렉터리를 확인함
[1] "c:/R"
```

```
> setwd("c:/R")           슬래시 기호인 것을 주의함 (₩₩와 /는 같음)
> getwd( )
[1] "c:/R“
>
```

◆ print( ) 사용하기

● 화면에 결과를 보여주는 스크립트 작성법

```
> print(1+2)              print 안에 출력하고 싶은 내용 작성함
[1] 3
> 1+2                     print 명령이 생략된 것임
[1] 3
> print('a')              문자를 출력할 때는 작은 따옴표를 붙여야 함
[1] "a"
> 'a‘                     문자만 입력해도 정상적으로 출력됨
[1] "a"
```

```
> print(pi)               소수점일 경우 총 7자리로 출력함
[1] 3.141593
> print(pi,digits=3)      digits로 자릿수 지정할 수 있음
                          → 3을 기입하면 두 자릿수 표현
[1] 3.14
```

2. R 스크립트 용어 정리

◆ R에게 줄 수 있는 자료들

● 숫자형과 주요 산술연산자

```
> 1+2
[1] 3
```

기호	의미	사용 예(결과)
+	더하기	5+6 → 11
-	빼기	5-4 → 1
*	곱하기	5*6 → 30
/	나누기(실수 가능)	4/2 → 2
%/%	나눈 몫 구하기	4.5%/%2 → 2
%%%	나머지 구하기	5%%4 → 1
^ , **	승수 구하기	3^2→ 9, 3^3 → 27

2. R 스크립트 용어 정리

◆R 주요 패키지 설명

● 문법 : extractNoun(문장 또는 변수에서 한글 명사만 추출하는 함수)

```
> v1 <- ("대한민국은 행복한 나라입니다.")        v1 변수에 좌측 문장을 삽입
> v1
 [1] "대한민국은 행복한 나라입니다."
> extractNoun(v1)
[1] "대한" "민국" "행복" "한" "나라"
```

● 코드 명과 코드 해설

- words : 출력할 단어들이나 단어들이 들어가 있는 변수 이름
- freq : 언급된 횟수
- scale : 가장 많이 언급된 글자와 적게 언급된 글자의 크기 비율
- min.freq : 최소언급 횟수 지정(이 값 이상 언급된 단어만 출력)
- max.words : 표시할 최대 단어 개수 지정 (출력된 단어 개수가 설정된 값 이상이라면, 최소 빈도수를 갖는 데이터부터 제거)
- random.order : 출력되는 순서를 임의로 지정
- random.color : 글자 색상을 임의로 지정
- rot.per : 단어 배치를 90° 각도로 출력
- colors : 출력될 단어들의 색상을 지정
- ordered.colors : 이 값을 True로 지정할 경우 각 글자별로 색상을 순서대로 지정 가능
- use.r.layout : 이 값을 False로 할 경우 R에서 c++ 코드 사용 가능

‘N사’ 영화 리뷰 분석

1. 영화 리뷰 가져오기

① 영화포털 사이트 접속

- 영화 리뷰가 있는 포털사이트에 접속하여 자신이 원하는 영화를 검색

② 평점 / 리뷰 탭

- 평점 / 리뷰 탭을 선택하여 네티즌들이 남긴 평점과 리뷰 내용을 확인함
- 포털사이트에서는 데이터 시각화가 잘 이루어지고 있는 것을 확인할 수 있음

‘N사’ 영화 리뷰 분석

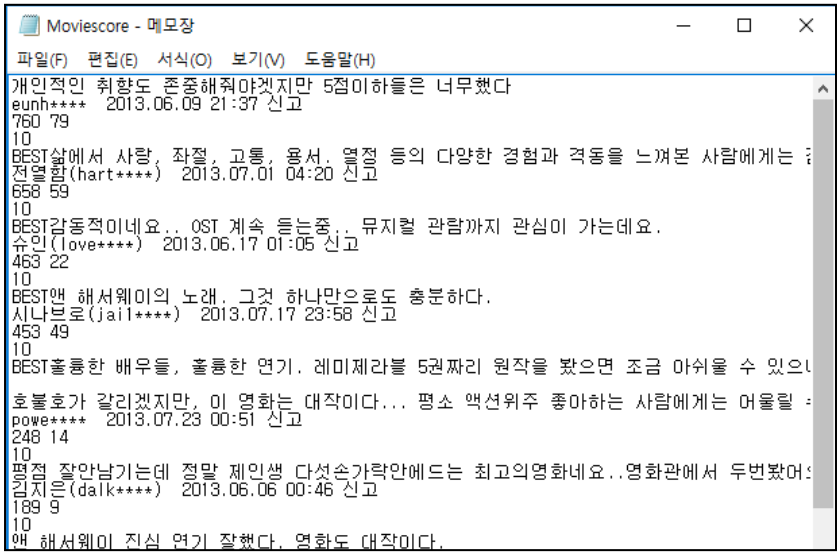
1. 영화 리뷰 가져오기

③ 리뷰 데이터 저장

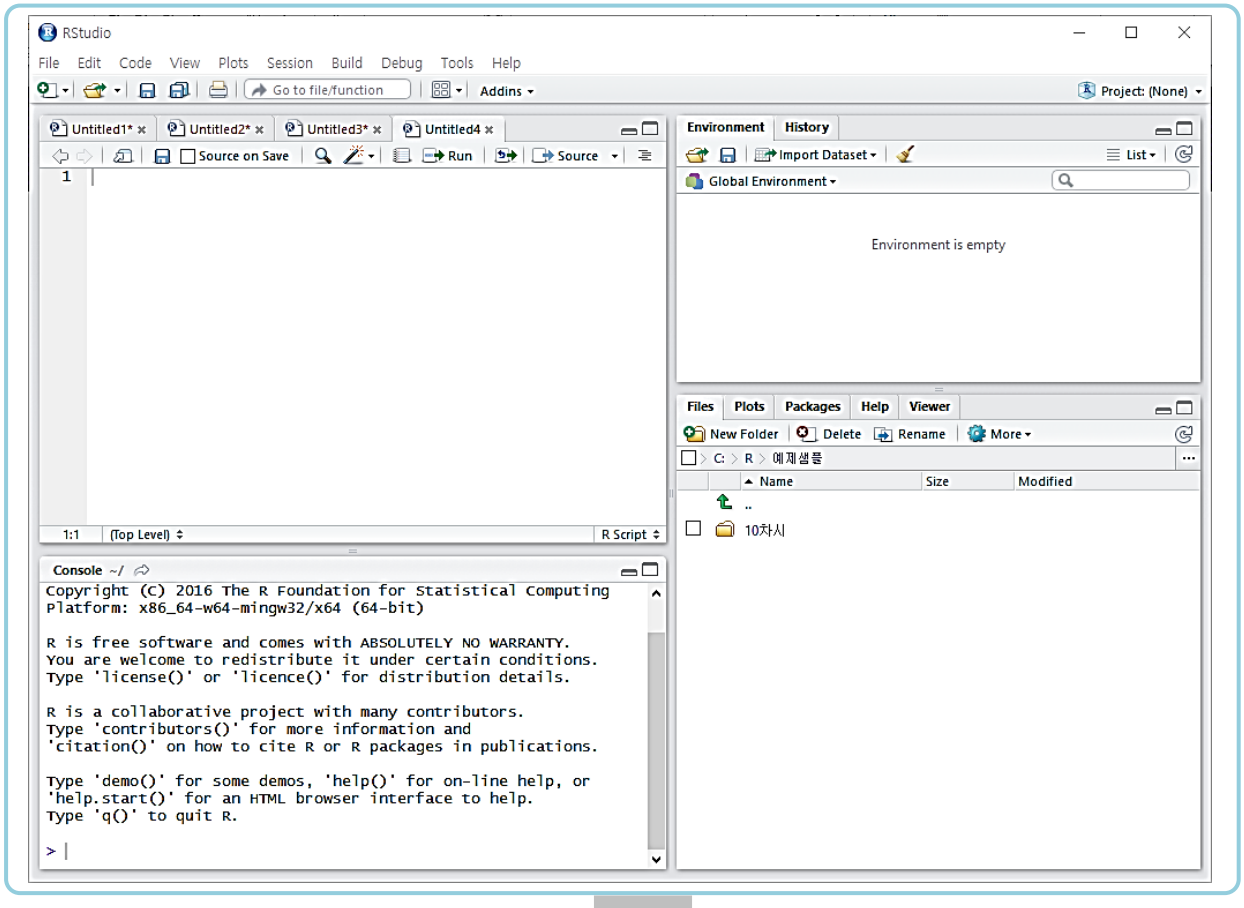
- 평점/리뷰 탭안에 있는 리뷰를 저장하여 Moviescore.txt 파일로 저장함

④ 리뷰 텍스트 데이터

- 리뷰 2만여 건 전체가 아닌 5페이지만 저장하여 Moviescore.txt 파일로 저장함



2. R 스크립트 분석



새로운 R스크립트 화면을 만든 후 Rscript 입력

2. R 스크립트 분석

① 작업용 디렉터리 지정

```
> setwd("c:₩₩R")                    작업 디렉토리는 임의로 지정
```

② 필요한 패키지를 설치 한 후 R에 Loading

```
> install.packages("KoNLP")          한글화 지원 패키지 설치
>install.packages("wordcloud")       워드클라우드 패키지 설치
>install.packages(“RcolorBrewer")    워드클라우드 색상 패키지 설치
> library(KoNLP)                     한글화 지원 패키지 로딩
> library(wordcloud)                 워드클라우드 패키지 로딩
> library(RcolorBrewer)              워드클라우드 색상 패키지 로딩
```

③ 분석할 원본 데이터를 변수로 읽어들임

```
> data1 <- readLines("Moviescore.txt")
> data1
1] “개인적인 취향도 존중해줘야겠지만 5점이하들은 너무했다”
[2] “eunh**** 2013.06.09 21:37 신고”.
.
.
[87] “30 4”
[88] “10”
[89] “보는내내 소름이 돋았습니다.”
```

2. R 스크립트 분석

④ 데이터 중에서 명사만 골라낸 후 data2 변수에 할당함

```
> data2 <- sapply(data1,extractNoun,USE.NAMES=F)
> data2
[1] “BEST” “삶” “사랑” “좌절” “고통” “용서”
[7] “열정” “등” “다양” “한” “경험” “격동”
[13] “사람” “감정” “공유” “속” “듯” “생동감”
[19] “최고” “명작” “온실” “무미건조” “한” “일상”
[25] “속” “영화” “속” “대리만족” “자극” “사람”
[31] “영화”
```

⑤ 두 글자 이상 되는 것만 필터링 함

```
> data3 <- unlist(data2)                                    # 비순차적으로 정렬
> data3 <- Filter(function(x) {nchar(x) >= 2} ,data3)
> data3
[1] “취향” “존중해줘야겠지만”
[3] “5점이하들은” “eunh”
[5] “2013.” “06.”
[7] “09” “21”
[9] “37” “신고”
[11] “760” “79”
```

⑥ 파일로 저장 한 후 테이블 형태로 변환하여 불러들임

```
> write(unlist(data3),"Moviescore2.txt")
```

‘N사’ 영화 리뷰 분석

2. R 스크립트 분석

⑦ 수정 완료된 파일을 read.table 명령으로 다시 변수에 불러들임

```
> data4 <- read.table("Moviescore2.txt")
> wordcount <- table(data4)        # 각 단어별 빈도수를 계산하여 변수에 할당
```

⑧ 가장 많이 언급된 데이터를 내림차순으로 20개 정렬

```
> head(sort(wordcount, decreasing=T),20)
```

⑨ Moviescore2.txt 파일을 불러내어 불필요한 단어를 지워버린 후 Moviescore3.txt 파일로 저장

```
> write(data4, “Moviescore3.txt”)
```

⑩ 저장된 데이터 파일은 read.table 명령어로 다시 불러내어 실행

```
> data4 <- read.table("Moviescore3.txt")
> wordcount <- table(data4)                    # 테이블 데이터의 개수를 변수에 할당
> head(sort(wordcount, decreasing=T),20)
```

⑪ Word Cloud 형태로 그래픽으로 출력

```
> palete <- brewer.pal(9,"Set3")
>ordcloud(names(wordcount),freq=wordcount,scale=c(5,1),rot.per=0.25,
min.freq=1,random.order=F,random.color=T,colors=palete)
```

2. R 스크립트 분석

◆ RStudio실습 순서

- 작업용 디렉터리 지정
- 필요한 패키지를 설치 후 R에 Loading
- 분석할 원본 데이터를 변수로 읽어 들임
- 데이터 중에서 명사만 골라낸 후 data2 변수에 할당
- 두 글자 이상 되는 것만 필터링
- 파일로 저장 한 후 테이블 형태로 변환하여 불러들임
- 수정 완료된 파일을 read.table 명령으로 다시 변수에 불러들임
- 가장 많이 언급된 데이터를 내림차순으로 20개 정렬
- Moviescore2.txt파일을 불러내어 불필요한 단어 지워버린 후 Moviescore3.txt 파일로 저장
- 한번 저장된 데이터 파일은 read.table 명령어로 다시 불러내어 실행
- Word Cloud 형태로 그래픽으로 출력

‘D사’ 영화 리뷰 분석

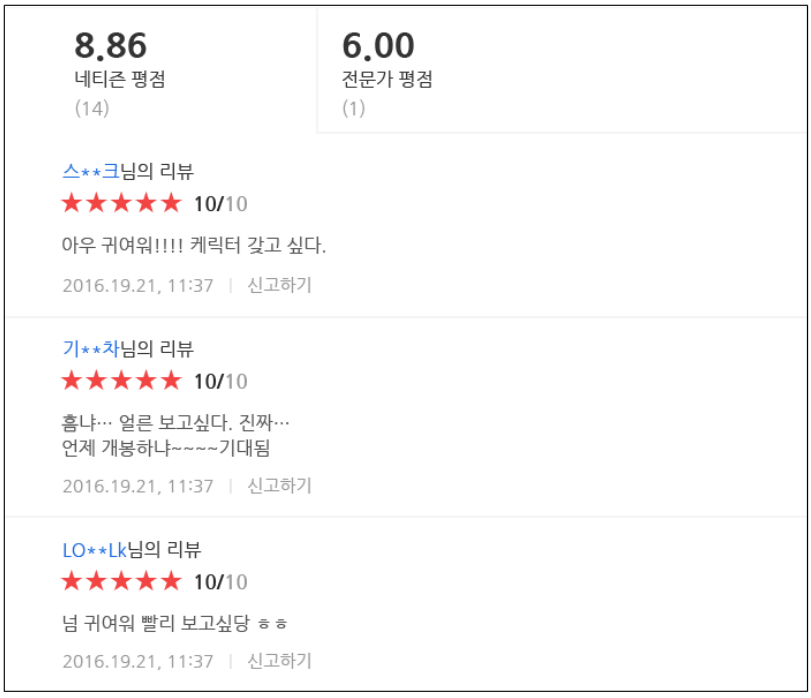
1. 영화 리뷰 가져오기

① 영화포털 사이트 접속

- 영화 리뷰가 있는 포털사이트에 접속하여 자신이 원하는 영화를 검색함

② 평점 / 리뷰 탭

- 개봉되기 전 영화 리뷰를 통해 관객 수를 예측하기 위해, 개봉 전 영화를 검색한 후 리뷰를 가져옴



‘D사’ 영화 리뷰 분석

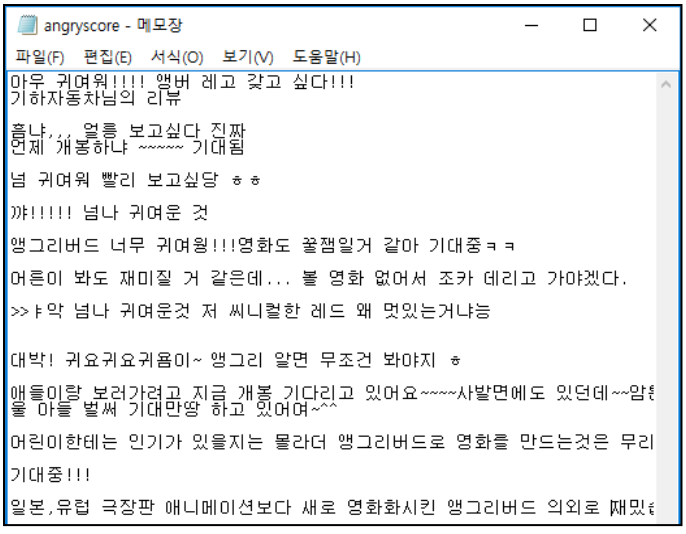
1. 영화 리뷰 가져오기

③ 평점 / 리뷰 데이터 저장

- 평점/리뷰 탭안에 있는 리뷰를 저장하여 movie.txt 파일로 저장함

④ 평점 / 리뷰 텍스트 데이터

- 개봉 전 영화의 평점 리뷰는 빅데이터라기보다는 스몰데이터에 가까움
- 미래 예측을 위해서는 스몰데이터에 대한 분석도 중요함



2. R 스크립트 분석

◆ RStudio실습 순서

- 작업용 디렉터리 지정
- 필요한 패키지를 설치 한 후 R에 Loading
- 분석할 원본 데이터를 변수로 읽어 들임
- 데이터 중에서 명사만 골라낸 후 data2 변수에 할당
- 두 글자 이상 되는 것만 필터링
- 파일로 저장 한 후 테이블 형태로 변환하여 불러들임
- 수정 완료된 파일을 read.table 명령으로 다시 변수에 불러들임
- 가장 많이 언급된 데이터를 내림차순으로 20개 정렬
- Word Cloud 형태로 그래픽으로 출력

핵심요약

1. 비정형데이터

- **비정형 데이터란?**
  - 문서, 그림, 음성, 영상처럼 구조화되지 않은 데이터, 형태와 구조가 다양한 정보
- **작업용 기본 디렉터리**
  - > setwd("c:₩₩R") : C 드라이브에 R폴더를 작업 디렉터리로 설정함
  - > getwd( ) : 현재 설정된 작업 디렉터리를 확인
- **코드 명과 해설**
  - words : 출력할 단어들이나 단어들이 들어가 있는 변수 이름
  - freq : 언급된 횟수
  - scale : 가장 많이 언급된 글자와 적게 언급된 글자의 크기 비율
  - min.freq : 최소언급 횟수 지정(이 값 이상 언급된 단어만 출력)
  - max.words : 표시할 최대 단어 개수 지정 (출력된 단어 개수가 설정된 값 이상이라면, 최소 빈도수를 갖는 데이터부터 제거)
  - random.order : 출력되는 순서를 임의로 지정
  - random.color : 글자 색상을 임의로 지정
  - rot.per : 단어 배치를 90° 각도로 출력
  - colors : 출력될 단어들의 색상을 지정
  - ordered.colors : 이 값을 True로 지정할 경우 각 글자별로 색상을 순서대로 지정 가능
  - use.r.layout : 이 값을 False로 할 경우 R에서 c++ 코드 사용 가능

핵심요약

2. ‘N사’ 영화 리뷰 분석

- **R 스크립트 분석**
  - 작업용 디렉터리 지정
  - 필요한 패키지를 설치 후 R에 Loading
  - 분석할 원본 데이터를 변수로 읽어 들임
  - 데이터 중에서 명사만 골라낸 후 data2 변수에 할당
  - 두 글자 이상 되는 것만 필터링
  - 파일로 저장 한 후 테이블 형태로 변환하여 불러들임
  - 수정 완료된 파일을 read.table 명령으로 다시 변수에 불러들임
  - 가장 많이 언급된 데이터를 내림차순으로 20개 정렬
  - Moviescore2.txt파일을 불러내어 불필요한 단어 지워버린 후 Moviescore3.txt 파일로 저장
  - 한번 저장된 데이터 파일은 read.table 명령어로 다시 불러내어 실행
  - Word Cloud 형태로 그래픽으로 출력

핵심요약

3. ‘D사’ 영화 리뷰 분석 실습

- **R 스크립트 분석**
  - 작업용 디렉터리 지정
  - 필요한 패키지를 설치 한 후 R에 Loading
  - 분석할 원본 데이터를 변수로 읽어 들임
  - 데이터 중에서 명사만 골라낸 후 data2 변수에 할당
  - 두 글자 이상 되는 것만 필터링
  - 파일로 저장 한 후 테이블 형태로 변환하여 불러들임
  - 수정 완료된 파일을 read.table 명령으로 다시 변수에 불러들임
  - 가장 많이 언급된 데이터를 내림차순으로 20개 정렬
  - Word Cloud 형태로 그래픽으로 출력