11강 텍스트 데이터의 시각화 2

숭실대학교 정보통계보험수리학과
이정진 교수

1. 한글 문서의 워드클라우드
2. 트위터 문서의 워드클라우드

1. 한글 문서의 워드 클라우드

1 한글 문서의 워드 클라우드

텍스트 데이터 분석

1) 텍스트 데이터베이스에서 문서들의 집합인 코퍼스 생성

2) 코퍼스의 단어줄기 추출 (word stemming)

3) 언어의 특성에 따른 추가 정제작업 : 대소문자 통일 등

4) 불용어(stop word) 제거 : 전치사, 관사 등

5) 단어의 출현빈도 조사 → 워드 클라우드(word cloud)

6) 데이터베이스 특징을 의사 결정에 이용

1 한글 문서의 워드 클라우드

워드 클라우드 알고리즘

단계 1) 단어들이 그려질 영역을 설정한다.
랜덤하게 할 수도 있고 지정할 수도 있다.

단계 2) 각 영역에서 다음을 반복한다.

단계 3) 각 영역에 그릴 단어의 폰트 크기와 색을
출현 빈도를 이용하여 결정한다.

단계 4) 단어 회전 여부를 결정한다.

단계 5) 다른 인접 영역의 단어와 겹치지 않는지
확인하고 단어를 그린다.

1 한글 문서의 워드 클라우드

[예제 6.2] 한글 문서에 대한 워드 클라우드

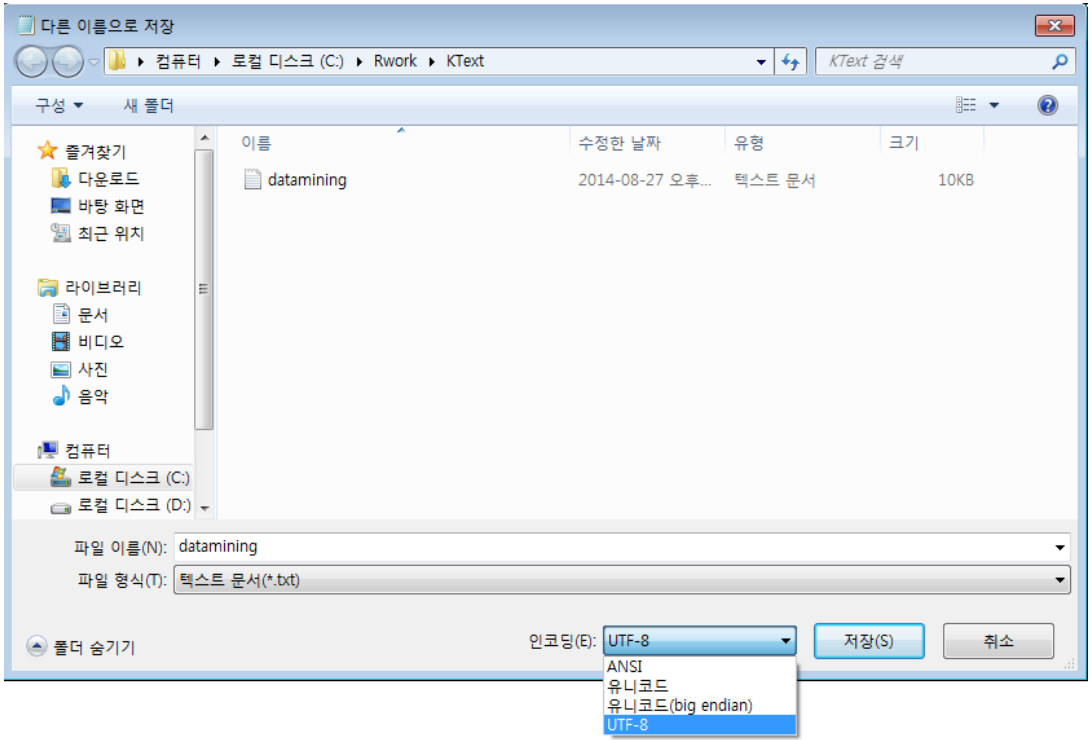
c:Rwork/KText/datamining.txt

데이터마이닝(data mining)이란 대량의 데이터를 분석하여 중요한 정보를 얻어 현실에 응용하는 기법을 의미한다. 예를 들어, 웹로그 데이터의 분석은 젊은 사람들이 제일 많이 이용하는 웹사이트는 무엇이며, 이 웹사이트 중에서 어떠한 웹페이지가 가장 많은 관심의 대상인지 파악하여, 이들을 대상으로 하는 광고 전략을 세울 수 있다.

※ 주의 : 위와 같은 한글 텍스트 파일을 R의 KoNLP 패키지를 이용하여 처리하기 위해서는 저장 인코딩(encoding) 방식을 "UTF-8" 로 지정하여야 한다.

1 한글 문서의 워드 클라우드

[예제 6.2] 한글 문서에 대한 워드 클라우드



※ 주의 : 노트패드에서 한글 인코딩(encoding) 방식 선택 "UTF-8" 로 지정

1 한글 문서의 워드 클라우드

[예제 6.2] 한글 문서에 대한 워드 클라우드

```
# 워드 클라우드를 위한 패키지 설치
install.packages(c("tm", "wordcloud", "KoNLP", "RColorBrewer"))

library(tm)
library(wordcloud)
library(KoNLP)
library(RColorBrewer)
```

1 한글 문서의 워드 클라우드

[예제 6.2] 한글 문서에 대한 워드 클라우드

```
# 파일을 한글 형식으로 불러옴
ktext <- Corpus(DirSource("C:/Rwork/KText", encoding="UTF-8",
recursive=TRUE))
```

1 한글 문서의 워드 클라우드

[예제 6.3] 여러 한글 문서에 대한 워드 클라우드

c:Rwork/Text/14th.txt
친애하는 7천만 내외 동포 여러분,
48년 전 오늘, 우리는 벅찬 감격 속에서 조국광복을 맞았습니다.
우리는 그날의 감격을 되새기면서, 못다 이룬 광복의 완성을 다짐하기 위해 이 자리에 모였습니다.
1997년 8월 15일 대통령 김영삼

c:Rwork/Text/15th.txt
친애하는 7천만 내외 동포 여러분,
48년 전 오늘, 우리는 벅찬 감격 속에서 조국광복을 맞았습니다.
우리는 그날의 감격을 되새기면서, 못다 이룬 광복의 완성을 다짐하기 위해 이 자리에 모였습니다.
2002년 8월 15일 대통령 김대중

※ 주의 : 위와 같은 한글 텍스트 파일을 R의 KoNLP 패키지를 이용하여 처리하기 위해서는 저장 인코딩(encoding) 방식을 "UTF-8" 로 지정하여야 한다.

1 한글 문서의 워드 클라우드

[예제 6.3] 여러 한글 문서에 대한 워드 클라우드

c:Rwork/Text/16th.txt
존경하는 국민 여러분,
그리고 해외동포 여러분,
오늘은 참으로 뜻깊은 날입니다. 58년 전 오늘, 우리의 아버지 어머니들은 일본 제국주의의 압제에서 해방되었습니다.
2007년 8월 15일 대통령 노무현

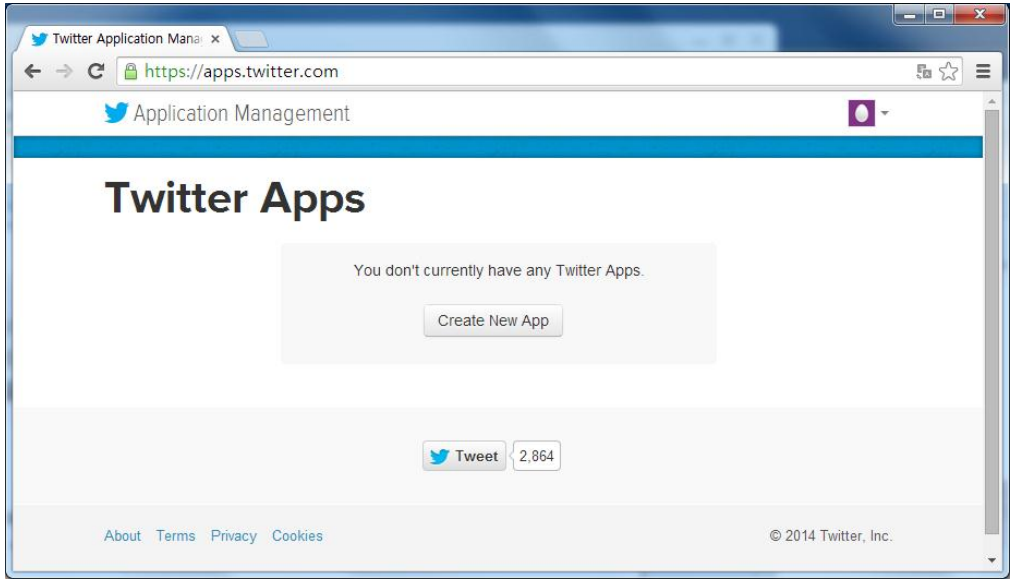
c:Rwork/Text/17th.txt
존경하는 국민 여러분! 재외동포와 국가유공자, 그리고 내외귀빈 여러분!
60년 전 오늘, 바로 이 자리에서 대한민국 정부 수립이 선포되었습니다. 5천년 한민족의 역사가 임시정부와 광복을 거쳐 대한민국으로 계승되는 순간이었습니다.
2012년 8월 15일 대통령 이명박

※ 주의 : 위와 같은 한글 텍스트 파일을 R의 KoNLP 패키지를 이용하여 처리하기 위해서는 저장 인코딩(encoding) 방식을 "UTF-8" 로 지정하여야 한다.

2. 트위터 문서의 워드 클라우드

2 트위터 문서의 워드 클라우드

트위터 API 키 만들기



1) 자신의 트위터 계정을 이용하여 http://apps.twitter.com 사이트에 로그인을 한다. 우선 새로운 앱을 만들어야 한다. "Create New App" 버튼을 클릭한다.

2 트위터 문서의 워드 클라우드

트위터 API 인증하기

1) R에서 사용하기 위해서는 앞서 생성한 API 키를 이용하여 인증을 받아야 한다. 이를 위해서는 우선 다음과 같이 인증 객체를 생성한다. 앞서 받은 API ke와 API secret를 이용한다.

```
> cred <- OAuthFactory$new(consumerKey="******************dIs9dZqT",
consumerSecret="**********************1d7RbgBlyFW2AG8nTBEYUnQaGGXyV2Z",
requestURL="https://api.twitter.com/oauth/request_token",
accessURL="https://api.twitter.com/oauth/access_token",
authURL="https://api.twitter.com/oauth/authorize")
```

2 트위터 문서의 워드 클라우드

[예제 6.4] “빅데이터”라는 말이 포함된 1000개의 트위터 문서에 대한 워드 클라우드

```
install.packages(c("tm", "wordcloud", "twitteR", "KoNLP"))
require(tm)
require(wordcloud)
require(KoNLP)
require(twitteR)
```

정리

- 한글 문서의 워드클라우드는 영어 문서의 경우와 유사하나
  한글 텍스트 파일을 R의 KoNLP 패키지를 이용하여야 한다.
- 한글 문서를 R에서 처리하기 위해서는
  저장 인코딩(encoding) 방식을 "UTF-8" 로 지정하여야 한다.
- 트위터 문서의 워드 클라우드는 한글과 유사하나
  자신의 트위터 계정을 이용하여야 한다.

다음시간안내

다변량 자료의 시각화 이해 1