10강 텍스트 데이터의 시각화 1

숭실대학교 정보통계보험수리학과
이정진 교수

1. 텍스트 데이터란?
2. 텍스트 마이닝이란?
3. 워드클라우드란?
4. 영어 문서의 워드클라우드

1. 텍스트 데이터란?

1 텍스트 데이터란?

텍스트 데이터란?

- 컴퓨터, 정보통신 기술의 발전으로 축적된 문자 정보
- 예 1. 도서관의 책, 논문 정보
- 예 2. 언론기관의 뉴스 정보
- 예 3. 포탈사이트의 게시물, 트윗 등

2. 텍스트 마이닝이란?

2 텍스트 마이닝이란?

텍스트 마이닝이란?

- 텍스트 데이터베이스 : 문자 정보 데이터 집합
- 정보의 단위 : 문서
- 문서들의 집합 : 코퍼스 (corpus)
- 데이터베이스에서 유용한 정보 탐색 : 텍스트 마이닝

3. 워드 클라우드(word cloud)란?

3 워드 클라우드(word cloud)란?

워드 클라우드(word cloud)란?

1) 텍스트 데이터베이스에서 문서들의 집합인 코퍼스 생성

2) 코퍼스의 단어줄기 추출 (word stemming)

3) 언어의 특성에 따른 추가 정제 작업 : 대소문자 통일 등

4) 불용어(stop word) 제거 : 전치사, 관사 등

5) 단어의 출현빈도 조사
   → 중요 단어를 구름처럼 배치
   → 데이터 베이스 특징을 의사 결정에 이용

3 워드 클라우드(word cloud)란?

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.

3 워드 클라우드(word cloud)란?

워드 클라우드에 이용되는 R의 패키지

- tm
- wordcloud
- RColorBrewer
- SnowballC

4. 영어 문서의 워드 클라우드

4 영어 문서의 워드 클라우드

[예제 6.1] 두 문서에 대한 워드 클라우드

c:Rwork/Text/sampletext.txt

Text mining, also known as text data mining or knowledge discovery from textual databases, refers to the process of extracting interesting and non-trivial patterns or knowledge from text documents. Regarded by many as the next wave of knowledge discovery, text mining has very high commercial values. Last count reveals that there are

c:Rwork/Text/wordcloud.txt

How I used R to create a word cloud, step by step
R is less scary than you thought! R, the open source package, has become the de facto standard for statistical computing and anything seriously data-related (note I am avoiding the term 'big data' here ? oops, too late!). From data mining to predictive analytics to data visualisation, it seems like any

4 영어 문서의 워드 클라우드

[예제 6.1] 두 문서에 대한 워드 클라우드

```
# 워드 클라우드를 위한 패키지 설치
install.packages(‘tm’)
install.packages('SnowballC')
Install.packages(‘wordcloud’)

library(tm)
library(SnowballC)
library(wordcloud)
```

4 영어 문서의 워드 클라우드

Tm_map()의 일반적인 형태

tm_map(x, FUN, ...)

〈주요 인수 설명〉

x	코퍼스
FUN	변환함수(transformation function), 소문자화, 불용어 제거, 숫자 제거 등

4 영어 문서의 워드 클라우드

wordcloud()의 일반적인 형태

```
wordcloud(words,freq,scale=c(4,.5),min.freq=3,max.words=Inf,random.order
=TRUE,
random.color=FALSE, rot.per=.1,      colors="black",ordered.colors=FALSE,
use.r.layout=FALSE, fixed.asp=TRUE, ...)
```

4 영어 문서의 워드 클라우드

wordcloud()의 일반적인 형태

주요 인수 설명

- words 단어들의 코퍼스
- freq 단어들의 빈도수
- scale 단어크기의 범위를 지정하는 크기 2인 벡터
- min.freq 그림에서 이 최소빈도수보다 작은 단어들은 생략
- max.words 그림에 나타내는 최대 단어 수, 빈도수가 작은 단어들 제거됨
- random.order 단어를 랜덤하게 배치. FALSE면 단어 빈도수의 내림차순으로 그림
- random.color 단어의 색을 랜덤하게 선택,
  FALSE면 단어 빈도수에 근거하여 색을 선택
- rot.per 90도 회전시키는 단어들의 비율
- colors 단어의 색상, 빈도수 적은 것에서 큰 것으로
- ordered.colors 단어의 색상을 순서대로 정해줌

정리

- 텍스트 데이터 : 문자로 이루어진 데이터
- 텍스트 마이닝 : 데이터베이스에서 유용한 정보 탐색
- 텍스트 데이터 시각화 : 워드 클라우드
  - 텍스트 데이터 전체에 나타나는 중요 단어를 구름처럼 표시
  - 중요 단어를 색, 크기를 변화시키면서 보여줌
  - 단어의 특징을 의사결정에 이용

다음시간안내

텍스트 데이터의 시각화 2