10강. 텍스트 데이터의 시각화 1

◈ 담당교수 : 이정진

들어가기

■ 주요용어

용어	해설
텍스트마이닝 (text minig)	텍스트 데이터베이스에서 좀 더 효율적인 방법으로 유용한 정보를 탐색하는 방법.
줄기단어 (stem word)	동사 등에서 변하는 어미를 제거한 단어
불용어 (stop word)	문서 중에 나타나는 빈도는 높으나 의미가 없는 단어. 예를 들면 관사, 또는 '것' 등 다른 색을 칠해 시각화
코퍼스 (corpus)	문서들의 집합
워드 클라우드 (word cloud)	텍스트 데이터베이스의 한 문서에서 단어의 출현 빈도를 이용하여 많이 나타난 단어일수록 더 큰 글자로 화면 가운데 배치하여 단어 구름을 만들어 시각화하는 그림

연습문제

1. 문서들의 집합을 ( )라 부른다.

정답 : 코퍼스 (corpus)

2. 동사 등의 단어에서 변화하는 어미를 제거하는 작업을 ( ) 이라 한다.

정답 : 단어줄기 추출 (stemming)

3. 텍스트마이닝에 사용되는 R패키지가 아닌 것은?

① tm
② word cloud
③ sp
④ KoNLP

정답 : ③
해설 : sp는 지리적 데이터 시각화에 사용되는 패키지이다.

4. 다음 중 불용어에 해당하는 단어는?
① 국민
② 민족
③ 것
④ 나라

정답 : ③
해설 : '것'은 대개 분석의 대상이 되지 않는 불용어이다.

5. 단어의 출현 빈도를 이용하여 많이 나타난 단어일수록 더 큰 글자로 화면 가운데 배치하여 단어 구름을 만드는 것을 무엇이라 하나?
① 단어 분석
② 워드 클라우드
③ 빈도 분석
④ 13

정답 : ②
해설 : 단어 구름을 워드 클라우드(word cloud)라 한다

정리하기

1. 텍스트 데이터는 형태나 크기가 일정치 않은 비정형 데이터이다. 이러한 비정형 데이터의 모임인 텍스트 데이터베이스에서 유용한 정보를 탐색하는 것을 텍스트마이닝(text mining)이라 한다.

2. 텍스트 데이터베이스에 나타나는 단어의 출현 빈도를 이용하여 텍스트 데이터베이스 내의 중요 단어를 구름처럼 시각화 하는 것을 워드 클라우드(word cloud)라고

한다.

3. 텍스트 데이터베이스에 크기가 서로 다른 문서들이 있을 때 이 문서들의 집합인 코퍼스(corpus)를 생성하고 전치사나 관사, 접속사, 그리고 관심의 대상이 되지 않는 단어들, 즉 불용어(stop word)를 제거한다. 여기서 어미 등을 제거한 단어줄기(stem)를 추출한다.

4. 정제된 코퍼스 안에 있는 모든 단어들의 출현 빈도를 조사하여 중요 단어를 추출한다. 텍스트마이닝은 각 문서가 어떠한 중요 단어를 보유하고 있는지 조사하여 이를 이용한 분류분석, 군집분석 등 다양한 모형을 적용하여 분석하는 것이다.

5. 워드 클라우드(word cloud)는 중요 단어의 출현 빈도를 이용하여 가장 많이 나타난 단어일수록 더 큰 글자나 색 또는 회전으로 구분하여 단어 구름을 만들어 시각화하는 것이다.

6. 워드 클라우드는 텍스트 데이터베이스 전체의 주요 내용이 어떠한 것인지 살펴볼 수 있어서 최고 경영자의 의사결정에 도움을 줄 수 있다.

참고자료

1. R 프로그램 사이트 : http://tm.r-forge.r-project.org/
2. http://cran.r-project.org/
3. http://www.ddokbaro.com/
4. http://statistical-research.com/
5. http://davetang.org/