20장 텍스트마이닝

1. 서론

이때까지 다뤄온 수치형, 이진형, 다중범주형이 아닌 텍스트 데이터.

수많은 문서의 검토와 같은 대규모 검색과정등에서 전처리 등.

2. 텍스트의 표로 나타내는 표현 : 용어 - 문서 행렬과 "단어 주머니"

다음과 같은 3개의 문장

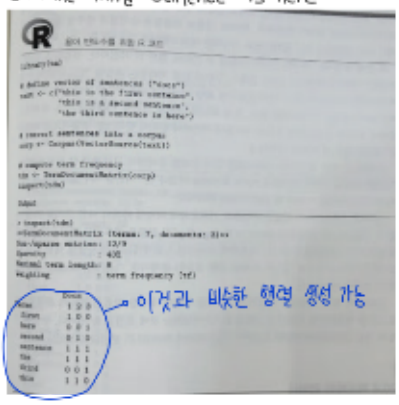
S1 : this is the first sentence

S2 : this is a second sentence

S3 : the third sentence is here

→ 이 세 개의 문장 (문서라고 불리는)에서 단어('용어')들을 용어-문서 행렬 (term-document matrix)로 나타낼 수 있다.

각 행은 각 단어, 각 열은 문장



각 셀 안의 숫자는 문장에서 해당 단어의 빈도 수 → '단어 주머니' (bag-of-words) 접근법

↳ 문서는 순서와 문법, 신택스가 상관없는 단어들의 집합으로 단순하게 취급

3. 단어주머니 vs 문서단위의 의미추출

텍스트마이닝에서의 작업
- 문서가 어떤 클래스에 속하는지 표시하거나 유사한 문서들의 클러스터링
- 문서에서 좀 더 구체적인 의미의 추출

첫 번째 작업에서는 '말 뭉치' (Corpus) 라고 하는 상당한 양의 문서집합과 문서로부터 예측변수들을 추출할 수 있는 능력,

분류과제에 대해 모델을 학습하기 위해 사전에 레이블된 대량의 문서가 필요.

↳ 사용된 모델들은 수치형과 범주형 데이터를 위해서 이미 취급했던 표준 통계모델, 기계학습 예측모델

두 번째 작업에는 하나의 문서만 사용, 훨씬 더 포괄적 → 컴퓨터가 인간언어를 이해하기 위해선 문법, 신택스, 구두점 등을 처리하는

복잡한 알고리즘의 이런 버전 모두를 학습필요. → 자연언어 처리 (NLP)

↳ 단어의 순서나 동의적인 표현 처리가 핵심

⇒ 이 책에서는 문서에 확률적으로 클래스를 배정하거나 유사한 문서들을 클러스터링 하는 것에 초점.

4. 텍스트의 전처리

별도 정리 X.