텍스트마이닝

Preview

- 텍스트 데이터도 모델링이 가능한가?
  - 신동엽 시인과 안도현 시인의 아래 시를 가지고 모델링하면 새로운 시 "껍데기는 가라 사월도 알맹이만 남고 껍데기는 가라 " 가 어느 시인의 것인지 예측할 수 있나?

샘플1: "우리들의 어렸을 적 황토 벗은 고갯마을 할머니 등에 업혀 누님과 난, 곧잘 파랑새 노랠 배웠다." _신동엽

샘플2: "누가 하늘을 보았다 하는가 누가 구름 한 자락 없이 맑은 하늘을 보았다 하는가" _신동엽

샘플3: "너에게 묻는다 연탄재 함부로 발로 차지 마라 너는 누구에게 한번이라도 뜨거운 사람이었느냐" _안도현

샘플4: "눈 내리는 만경들 건너가네 해진 짚신에 상투 하나 떠가네 가는 길 그리운 이 아무도 없네" _안도현

- 텍스트 모델링이 가능하다면 유용한 응용이 많음
  - 영화 관람평을 모델링하면 흥행 예측 가능
  - 상품에 대한 댓글을 분석하여 마케팅 전략 세움
  - 트윗을 분석하여 대선이나 총선 결과 예측
  - 주식 관련 댓글을 보고 주가 예측

01.1 텍스트 데이터의 성질

■ 텍스트 데이터는 다음과 같은 독특한 성질을 가짐

- 비정형 데이터다. 문서마다 길이가 천차만별이며, 문서에 나타난 단어의 종류도 제각각이다. 문장 중간에 나타나는 숫자와 특수 기호, 외국어의 종류와 위치가 다양하다.
- 잡음이 많은 데이터다. '하다', '그리고', '위해' 등과 같은 불용어가 많으며, 구두점이 자주 나타난다. 그리고 어미가 심하게 변형되어 나타난다. 예를 들어, '뛰다'의 어미는 '뛰니, 뛴, 뜀' 등으로 변형되어 문서에 나타난다.
- 애매성이 많다. 예를 들어, 'time flies like an arrow'를 '시간은 화살처럼 빠르다'로 해석할 수도 있고 '시간 파리는 화살을 좋아한다'로 해석할 수도 있다.
- 텍스트 분석에는 구문론syntax과 의미론semantic이 있다. 문법만 따지는 구문론 수준에서는 위의 영어 문장을 파악하는 데 혼란이 있지만, 의미론에서는 '시간은 화살처럼 빠르다'로 해석할 수 있다. 의미론은 단어의 의미를 파악하여 문서를 분석해야 하므로 훨씬 어렵다.
- 언어가 다양하다. 영어에서는 'I'가 목적어가 되면 'me'이지만, 한국어에서는 '나는'이 '나를'이 되는 것처럼 조사를 이용해 목적어가 된다.

01.2 텍스트 데이터의 처리 파이프라인

텍스트 마이닝

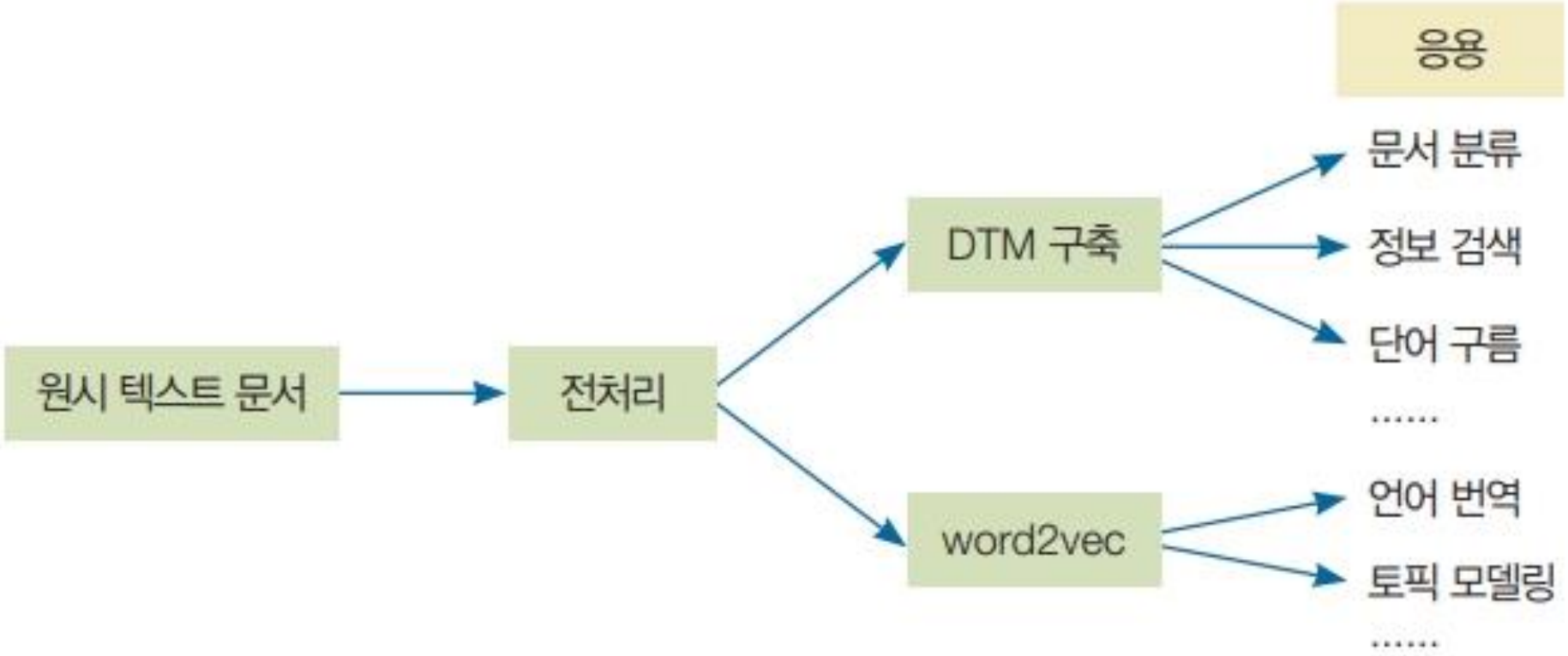
- 텍스트 데이터에서 유용한 정보 또는 지식을 찾아내는 일

용어

- 문서$_{document}$
  - 예) [그림 11-1]의 위키 설명문, 뉴스에서 뉴스 꼭지 하나하나, 트윗에서 트윗 하나, 댓글에서 댓글 하나, 신동엽 시인의 시 하나하나
- 말뭉치$_{corpus}$
  - 특정 분야에서 발생하는 문서의 집합
  - 예) 특정 연도에 치러지는 대선 관련 기사, 사회학자가 모은 한 달간 트윗 문서 전체, 국문학자가 모은 신동엽 시인의 시 전체

01.2 텍스트 데이터의 처리 파이프라인

■ 텍스트 처리 파이프라인



01.2 텍스트 데이터의 처리 파이프라인

■ 전처리 과정

- 전처리를 마치면 어느 정도 정보 손실이 있으나 다음 단계 처리에 적합한 형태가 됨

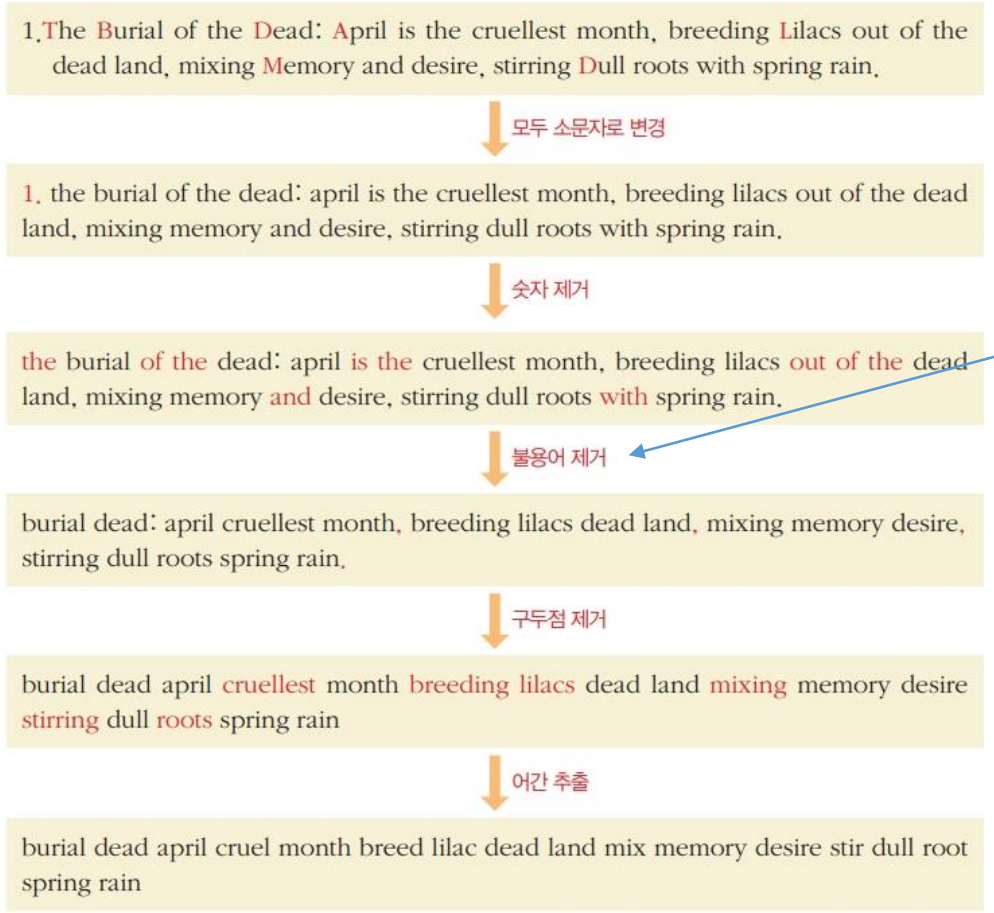


그림 11-3 텍스트 데이터의 전처리 예

불용어$_{\text{stop word}}$란 검색 색인 단어로 의미가 없는 단어
예) a, the, and, 그리고, 또는, 및

02 DTM 구축

- 텍스트 데이터는,
  - 비정형 데이터라 그 상태로는 시각화 함수를 적용할 수 없고 모델링할 수도 없음
  - 문서를 이들 함수에 적용하려면 일정한 크기의 벡터로 변환해야 함
  - DTM은 문서를 벡터로 변환하는 기술

NOTE 문서 단어 행렬, 즉 DWM Document Word Matrix이라 부르는 대신 DTM Document Term Matrix이라 부르는 이유는 사전을 만들 때 단어만 대상으로 하지 않고 일반적으로 n-그램을 대상으로 하기 때문이다. n-그램이란 연속으로 나타나는 n개 단어를 말한다. 예를 들어 "Data science is exciting and motivating."의 2-그램은 data-science, science-is, is-exciting, exciting-and, and-motivating이다. n-그램을 사용하면 단어가 나타나는 순서 정보를 어느 정도 보완할 수 있다는 장점이 있다.

02.1 DTM이란?

DTM

- 문서에 나타난 단어의 빈도를 표현하는 행렬
- 예제) 말뭉치에 다음과 같은 세 개의 문서가 있다고 가정

D1 : "Data science is exciting and motivating."
D2 : "I like literature class and science class."
D3 : "What is data science?"

- 사전$_{dictionary}$ 만들기 (문서에 나타난 단어를 모으면 사전이 됨)
- 예제에서는 9개의 단어가 사전을 구성
- [표 11-1]은 사전에 있는 단어별로 발생 빈도를 나타냄

표 11-1 문서별 단어 발생 빈도

	data	science	exciting	motivating	I	like	literature	class	what
D1	1	1	1	1	0	0	0	0	0
D2	0	1	0	0	1	1	1	2	0
D3	1	1	0	0	0	0	0	0	1

02.1 DTM이란?

DTM 예제

- 3개 문서를 다음과 같이 9차원 벡터로 표현

$$D1=(1, 1, 1, 1, 0, 0, 0, 0, 0)$$
$$D2=(0, 1, 0, 0, 1, 1, 1, 2, 0)$$
$$D3=(1, 1, 0, 0, 0, 0, 0, 0, 1)$$

- DTM 형태로 쓰면,

$$DTM=\begin{bmatrix}1&1&1&1&0&0&0&0&0\\0&1&0&0&1&1&1&2&0\\1&1&0&0&0&0&0&0&1\end{bmatrix}$$