제10강
텍스트 마이닝

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 텍스트 마이닝 기초

- 텍스트는 소통의 원천적 수단
  - 스마트폰으로 문자 전송
  - 트위터, 페이스북에 텍스트 전송

- 텍스트는 지금까지 다룬 데이터와 확연히 다른 성질을 가짐

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.

01.1 텍스트 데이터의 성질

- 위키피디아의 "data science" 문서 예제

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured,[1][2] similar to data mining.

Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.[3] It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.

Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.[4][5]

In 2012, when Harvard Business Review called it "The Sexiest Job of the 21st Century",[6] the term "data science" became a buzzword. It is now often used interchangeably with earlier concepts like business analytics,[7] business intelligence, predictive modeling, and statistics. Even the suggestion that data science is sexy was paraphrasing Hans Rosling, featured in a 2011 BBC documentary with the quote, "Statistics is now the sexiest subject around." [6] Nate Silver referred to data science as a sexed up term for statistics.[8] In many cases, earlier approaches and solutions are now simply rebranded as "data science" to be more attractive, which can cause the term to become "dilute[d] beyond usefulness."[9] While many university programs now offer a data science degree, there exists no consensus on a definition or suitable curriculum contents.[7] To its discredit, however, many data-science and big-data projects fail to deliver useful results, often as a result of poor management and utilization of resources.[11][12][13][14]

그림 11-1 텍스트 문서의 예(위키피디아에서 'data science' 검색)

이런 텍스트 데이터를
어떻게 처리할까?

01.2 텍스트 데이터의 처리 파이프라인

텍스트 마이닝

- 텍스트 데이터에서 유용한 정보 또는 지식을 찾아내는 일

용어

- 문서$_{\text{document}}$
  - 예) [그림 11-1]의 위키 설명문, 뉴스에서 뉴스 꼭지 하나하나, 트윗에서 트윗 하나, 댓글에서 댓글 하나, 신동엽 시인의 시 하나하나
- 말뭉치$_{\text{corpus}}$
  - 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

그림 11-2 텍스트 처리 파이프라인

DTM : Document Term Matrix

01.2 텍스트 데이터의 처리 파이프라인

전처리 과정

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

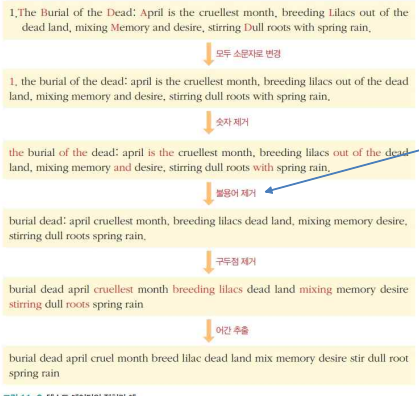


그림 11-3 텍스트 데이터의 전처리 예

불용어$_{stop\ word}$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 Matrix)이라 부르는 대신 DTM(Document Term Matrix)이라 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,

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

02.1 DTM이란?

- 부연 설명
  - 사전 구축
    - 실제에서는 사전의 크기는 보통 수만~수십만
    - 말뭉치에서 수집할 수도 있고, 국어사전에 실려있는 단어 집합을 사용할 수도 있음
  - 문서가 벡터로 표현되므로 거리 또는 유사성 측정 가능
    - 앞의 예제에서 D1 벡터는 D2 벡터보다 D3 벡터와 가까움
    - 랜덤 포리스트나 SVM 등의 적용이 가능해짐

    정규화 필요성
    - 문서가 길면 벡터의 길이가 커져서 유사한 문서와 거리가 멀어짐
    - 벡터의 길이를 1로 만드는 정규화를 적용하여 해결 가능

    DTM은 희소 행렬
    - 한번도 발생하지 않아 0인 칸이 아주 많음

    DTM은 단어 사이의 상호작용을 표현 못함
    - “Data science is exciting and motivating”과 “Data is exciting and science is motivating”은 같은 벡터로 변환됨
    - 2-그램이나 3-그램으로 해결 가능하나 열의 개수가 기하급수적으로 커짐

02.2 R을 이용한 전처리와 DTM 구축

- 위키피디아의 “data science” 문서로 실험
  - RCurl 라이브러리로 웹 서버에 접속
  - XML 라이브러리로 웹 문서 처리

```
> library(RCurl)
> library(XML)

> t = readLines('https://en.wikipedia.org/wiki/Data_science')
> d = htmlParse(t, asText = TRUE)
> clean_doc = xpathSApply(d,"//p", xmlValue)
```

readLines 함수는 지정된 URL에서 html 파일을 읽어옴

  - htmlParse와 xpathSApply 함수는 웹 문서를 R의 데이터 형으로 변환해줌

02.2 R을 이용한 전처리와 DTM 구축

- 전처리 ([그림 11-3] 참조)
  - tm 라이브러리는 데이터 마이닝 함수 제공
  - SnowballC 라이브러리는 어간을 추출하는 함수 제공

```
> library(tm)
> library(SnowballC)

> doc = Corpus(VectorSource(clean_doc))
> inspect(doc)

> doc = tm_map(doc, content_transformer(tolower))
> doc = tm_map(doc, removeNumbers)
> doc = tm_map(doc, removeWords, stopwords('english'))
> doc = tm_map(doc, removePunctuation)
> doc = tm_map(doc, stripWhitespace)
```

tm_map 함수는 지정된 매개변수 값에 따라 전처리 수행

02.2 R을 이용한 전처리와 DTM 구축

- DocumentTermMatrix 함수로 DTM 구축

```
> dtm = DocumentTermMatrix(doc)
> dim(dtm)
[1]  17 576

> inspect(dtm)
<<DocumentTermMatrix (documents: 17, terms: 576)>>
Non-/sparse entries: 788/9004
Sparsity           : 92%
Maximal term length: 22
Weighting          : term frequency (tf)
Sample             :
    Terms
Docs data field many methods now science scientists statistical statistics term
  10    8     0    0       1   0       5          2           1          0    0
  13   14     0    0       0   0       9          0           4          0    0
  14   11     0    1       0   0       4          3           2          1    3
  15    9     1    1       0   2      10          0           0          0    0
  16   13     3    2       0   1       7          1           0          3    0
  17    8     0    0       0   0       6          0           1          0    0
  5     5     0    3       0   4       5          0           0          3    3
  6     3     0    0       2   0       3          0           0          0    4
  8     7     0    0       0   0       5          1           2          3    1
  9     6     3    0       1   0       3          0           1          2    0
```

dim 함수는 DTM의 행과 열의 개수를 알려줌

inspect 함수는 상세 내용을 요약하여 보여줌

위키에 있는 문장 17개 각각을 문서로 간주함
576개의 단어를 추출하여 사전 구축

17*576=9792개의 칸 중에서
9004개는 0이라는 사실을 알려줌

03 단어 구름word cloud

- DTM을 구성하는 단어를 가시화
  - 단어마다 중요성이 다르고, 단어 사이에 연관관계 정보가 있음. 예) data는 field보다 science와 연관성이 더 높음

    단어 구름은 이런 정보를 2차원 공간에 표시하는 가시화 기법

    중요도가 높은 단어는 큰 폰트를 써서 중앙에 배치. 연관성이 높은 단어는 가까이 배치

03.1 wordcloud 라이브러리

- 단어 구름 예제
  - 앞 절에서 구한 DTM을 활용

```
> library(wordcloud)

> m = as.matrix(dtm)
> v = sort(colSums(m), decreasing = TRUE)
> d = data.frame(word = names(v), freq = v)
> wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.words = 100, random.
order = FALSE, rot.per = 0.35)
```

as.matrix 함수는 DTM을 행렬 표현으로 변환

sort 함수는 빈도(중요도)가 높은 순서로 단어를 정렬

중요도가 높은 상위 100개만 그리라는 옵션

세로로 배치할 단어의 비율을 35%로 하라는 옵션

그림 11-4 단어 구름

03.1 wordcloud 라이브러리

- 단어 구름의 여러 가지 옵션
  - RColorBrewer 라이브러리 (6.3.2절 참조)를 이용해 색상 입히기

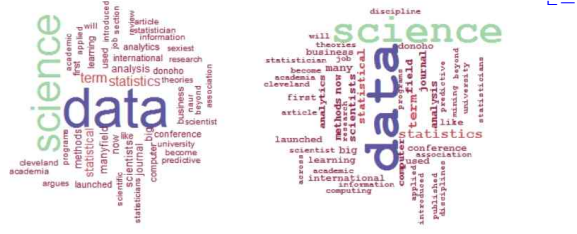
```
> library(RColorBrewer)
> pal = brewer.pal(11,"Spectral")
> wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.words = 50, random.
order = FALSE, rot.per = 0.50, colors = pal)
> wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.words = 50, random.
order = FALSE, rot.per = 0.50, colors = pal, family = "mono", font = 2)
```

가로 세로 비율

색상

단어 개수

폰트



(a) 단어 개수 조절 (b) 폰트 조절

그림 11-5 wordcloud의 옵션

03.2 wordcloud2 라이브러리

- 보다 뛰어난 wordcloud2 라이브러리
  - ◆□■●◆는 wordcloud 이후에 나온 새로운 버전
  - 자동으로 색상 입혀주고 단어를 다양한 방향으로 배치해 줌

```
> library(wordcloud2)
> wordcloud2(d)
```

그림 11-6 wordcloud2로 만든 단어 구름

03.2 wordcloud2 라이브러리

- wordcloud2의 여러 가지 옵션들

wordcloud2에는 단어 개수 지정하는 옵션이 없어 미리 추출

배경 모양

```
> d1 = d[1:200, ]          # 200개 단어만 표시
> wordcloud2(d1, shape = 'star')
> wordcloud2(d1, minRotation = pi/4, maxRotation = pi/4, rotateRatio = 1.0)
```

단어 방향 범위 지정

그림 11-7 wordcloud2의 변형

03.3 빈도 표시하기

- [illegible]와 findAssocs 함수

```
> findFreqTerms(dtm, lowfreq = 12)
[1] "data"       "science"    "statistics" "term"

> findAssocs(dtm, terms = 'harvard', corlimit = 0.7)
$harvard
sexiest     job  review
   0.94    0.79    0.79

> barplot(d[1:10, ]$freq, las = 2, names.arg = d[1:10,]$word, col = 'lightblue',
main = '발생 빈도 상위 단어', ylab = '단어 빈도')
```

상위 12개 단어만 표시하라는 옵션

'harvard'와 상관관계가 0.7이상인 단어를 표시하라는 옵션

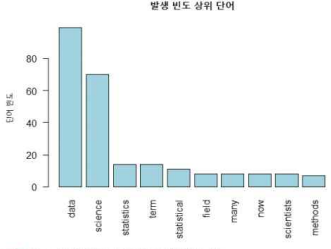


그림 11-8 중요도에 따라 단어를 나열한 막대그래프

03.4 텍스트 이외의 응용: gapminder 예제

- wordcloud는 데이터 프레임을 사용
  - 첫번째 열은 단어, 두번째 열은 해당 단어의 빈도수인 데이터 프레임
  - 텍스트 이외에도 이런 형식을 갖추면 단어 구름 가능함

  예) gapminder에서 첫번째 열은 대륙, 두번째 열은 해당 대륙의 인구가 되도록 데이터 추출

```
> library(gapminder)
> library(dplyr)

> pop_siz = gapminder%>%filter(year==2007)%>%group_by(continent)%>%summarize(sum
(as.numeric(pop)))
> d = data.frame(word = pop_siz[, 1], freq = pop_siz[, 2])
> wordcloud(words = d[, 1], freq = d[, 2], min.freq = 1, max.words = 100, random.
order = FALSE, rot.per = 0.35)
> wordcloud2(d)
```

(a) wordcloud 사용 (b) wordcloud2 사용

그림 11-9 gapminder를 단어 구름으로 그린 모양

감사합니다.