



13주차: 텍스트 마이닝

ChulSoo Park

School of Computer Engineering & Information Technology
Korea National University of Transportation



학습목표 (13주차)

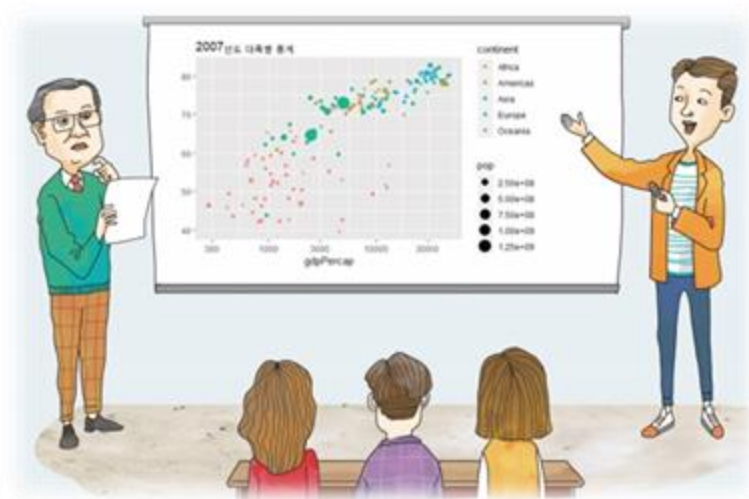
- ❖ 텍스트 마이닝의 개념 이해
- ❖ 텍스트 문서 벡터 변환(DTM) 구축
- ❖ 단어 구름(word cloud) 개념 이해
- ❖ 문서 분류
- ❖ 한국어 처리와 KoNLP를 이용한 텍스트 마이닝



11

CHAPTER

텍스트 마이닝



11.1 텍스트 마이닝 기초

11.2 DTM 구축

11.3 단어 구름

11.4 문서 분류

11.5 영어 텍스트 마이닝을 통한 한국어 처리

11.6 KoNLP를 이용한 한국어 텍스트 마이닝

요약



Review(예측 오류 발생)

- 첫째, 세상은 불확실성 투성이(코로나바이러스감염증-19(COVID-19))
 - 목소리를 보고 성별을 구분하는 경우, 목소리가 가는 남성은 여성의 음성과 흡사. 같은 남성이라도 평소 굵은 목소리를 냈는데 무척 피곤한 경우 가는 목소리
 - 기상청의 날씨 예측 오류, 프로모션은 매출 변화에 중요한 영향 요인
- 둘째, 데이터의 불완전성
 - 데이터를 측정할 때 기구의 불완전성이나 사람의 불완전성
 - 데이터 양이 작아 현장을 완전히 대표하지 못함. 예) colon 데이터의 경우 아무리 많은 대장암 환자 데이터를 모으더라도 데이터에 없는 특수 체질이 새로 발생함
- 데이터 과학이 할 수 있는 일
 - 엄정한 성능 평가 기준을 세우고, 여러 모델을 성능 평가하여 가장 뛰어난 모델을 선택하고, 성능이 일정 수준 이상이 되면 현장 설치



Review(정확률)

■ 가장 널리 활용되는 정확률

- 전체 Sample 수 : n , 정답 수 : n_1 , 오답 수 : n_2
- 정확률 : $\frac{n_1}{n}$, 오류율 : $\frac{n_2}{n}$
- 예) iris에 있는 150개 샘플
 - ✓ 정확률 : 150개 중에 102개를 맞추었다면 $102/150=68\%$
 - ✓ 오류율은 $48/150=32\%$

■ 기각(rejection) 기능이 있는 경우

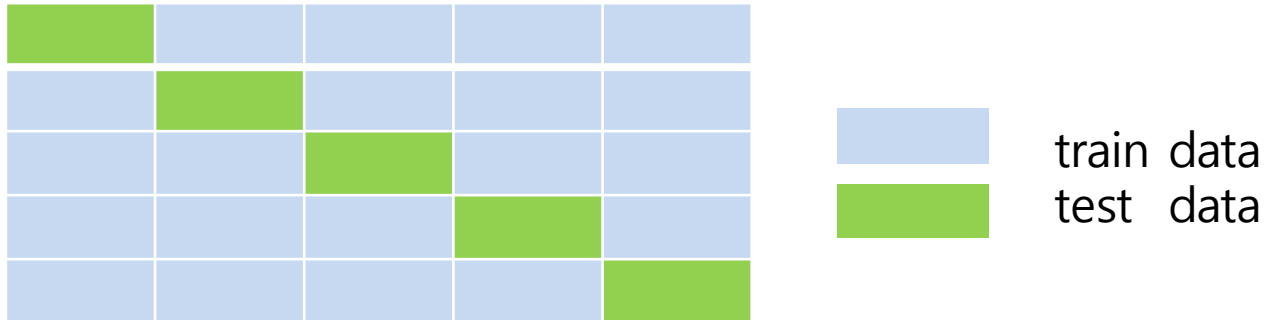
- 전체 Sample 수 : n , 정답 수 : n_1 , 오답 수 : n_2 , 기각한 샘플 수 : n_3
($n = n_1 + n_2 + n_3$)
- 정확률 : $\frac{n_1}{n}$, 오류율 : $\frac{n_2}{n}$, 기각률 : $\frac{n_3}{n}$



Review(교차 검증과 모델 선택 등)

■ k -겹 교차 검증

- 데이터를 k 개 부분 집합으로 분할 ([그림]은 5-겹 교차 검증)
- 학습과 평가를 k 번 반복하고, 평균을 취함



■ 모델의 선택

- 첫째, 하이퍼 매개변수는 기본값으로 두고 모델 선택
- 둘째, 하이퍼 매개변수 최적화와 모델 선택을 동시에 수행



■ 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?

샘플 1: "우리들의 어렸을 적 황토 벗은 고갯마을 할머니 등에 업혀 누님과 난, 곧잘 파랑새 노렐 배웠다." _ 신동엽

샘플 2: "누가 하늘을 보았다 하는가 누가 구름 한 자락 없이 맑은 하늘을 보았다 하는가" _ 신동엽

샘플 3: "너에게 묻는다 연탄재 함부로 차지 마라 너는 누구에게 한번이라도 뜨거운 사람이었느냐" _ 안도현

샘플 4: "눈 내리는 만경들 건너가네 해진 짚신에 상투 하나 떴가네 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



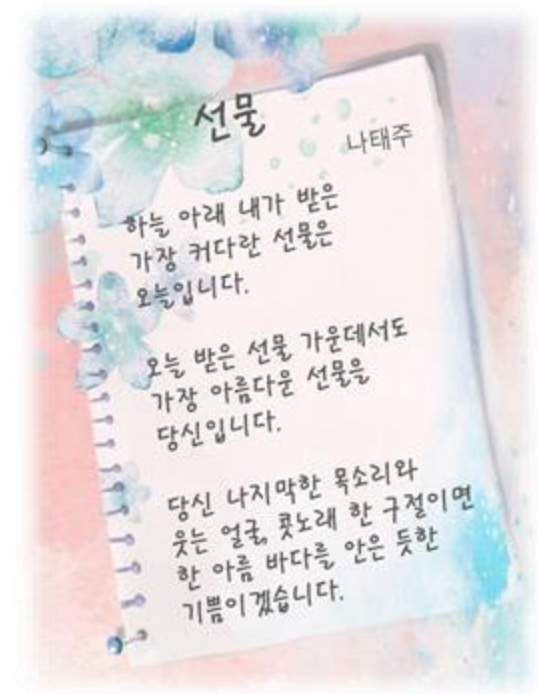
- [illegible]



11.1 텍스트 마이닝 기초

■ 텍스트 데이터는 다음과 같은 독특한 성질을 가짐

- 1) 비정형의 data이다. 길이, 숫자, 특수 기호, 표현 방법 등등
- 2) 잡음이 많은 data이다. "하다", "위해" 등과 같이 불용어가 많고 구두점도 자주 나타난다.
- 3) 애매성이 많다. 사슴 같은 목, 화살 같은 세월, 거북 같은 행동 등등
- 4) 텍스트 분석에는 구문론(syntax)과 의미론(semantic)이 있다. 의미론은 단어의 의미(문맥)를 파악해서 문서를 해석해야 하므로 훨씬 어렵다. (HOT?)
- 5) 언어가 다양하다.



Hot

위키백과, 우리 모두의 백과사전.

Hot 또는 **HOT**은 다음과 같은 뜻이 있다.

- Hot는 뜨거운, 뜨겁다라는 뜻을 가진 영어 단어다.
- H.O.T.: 대한민국의 5인조 보이 그룹
- Hot (태양의 음반)
- Hot (에이브릴 라빈의 노래)



■ 이러한 텍스트 data 처리방법은?

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute

Help
Learn to edit
Community portal
Recent changes
Upload file

Tools

What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export

Download as PDF
Printable version

Data science

From Wikipedia, the free encyclopedia

Not to be confused with [information science](#).

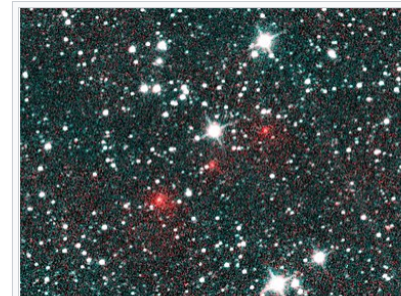
Data science is an [interdisciplinary](#) field that uses scientific methods, processes, algorithms and systems to extract [knowledge](#) and insights from structured and [unstructured data](#),^{[1][2]} and apply knowledge and actionable insights from data across a broad range of application domains. Data science is related to [data mining](#), [machine learning](#) and [big data](#).

Data science is a "concept to unify [statistics](#), [data analysis](#), [informatics](#), and their related methods" in order to "understand and analyze actual phenomena" with data.^[3] It uses techniques and theories drawn from many fields within the context of [mathematics](#), [statistics](#), [computer science](#), [information science](#), and [domain knowledge](#). 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]}

Contents [hide]

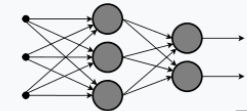
- 1 [Foundations](#)
 - 1.1 [Relationship to statistics](#)
- 2 [Etymology](#)
 - 2.1 [Early usage](#)
 - 2.2 [Modern usage](#)
- 3 [Impact](#)
- 4 [Technologies and techniques](#)
 - 4.1 [Techniques](#)
- 5 [References](#)

- 하이퍼텍스트
- 참고 문헌 인용 부호
- 벡터, 행렬 데이터 프레임과는 확연히 다름



The existence of [Comet NEOWISE](#) (here depicted as a series of red dots) was discovered by analyzing [astronomical survey data](#) acquired by a [space telescope](#), the [Wide-field Infrared Survey Explorer](#).

Part of a series on
**Machine learning
and
data mining**



■ 텍스트 마이닝

- 텍스트 데이터에서 유용한 정보 또는 지식을 찾아내는 일(흥행, 악성 댓글, 징조 등)

■ 용어

- 문서(document)

- 예) [그림 11-1]의 위키 설명문, 뉴스에서 뉴스 꼭지 하나하나, 트윗에서 트윗 하나, 댓글에서 댓글 하나, 신동엽 시인의 시 하나하나

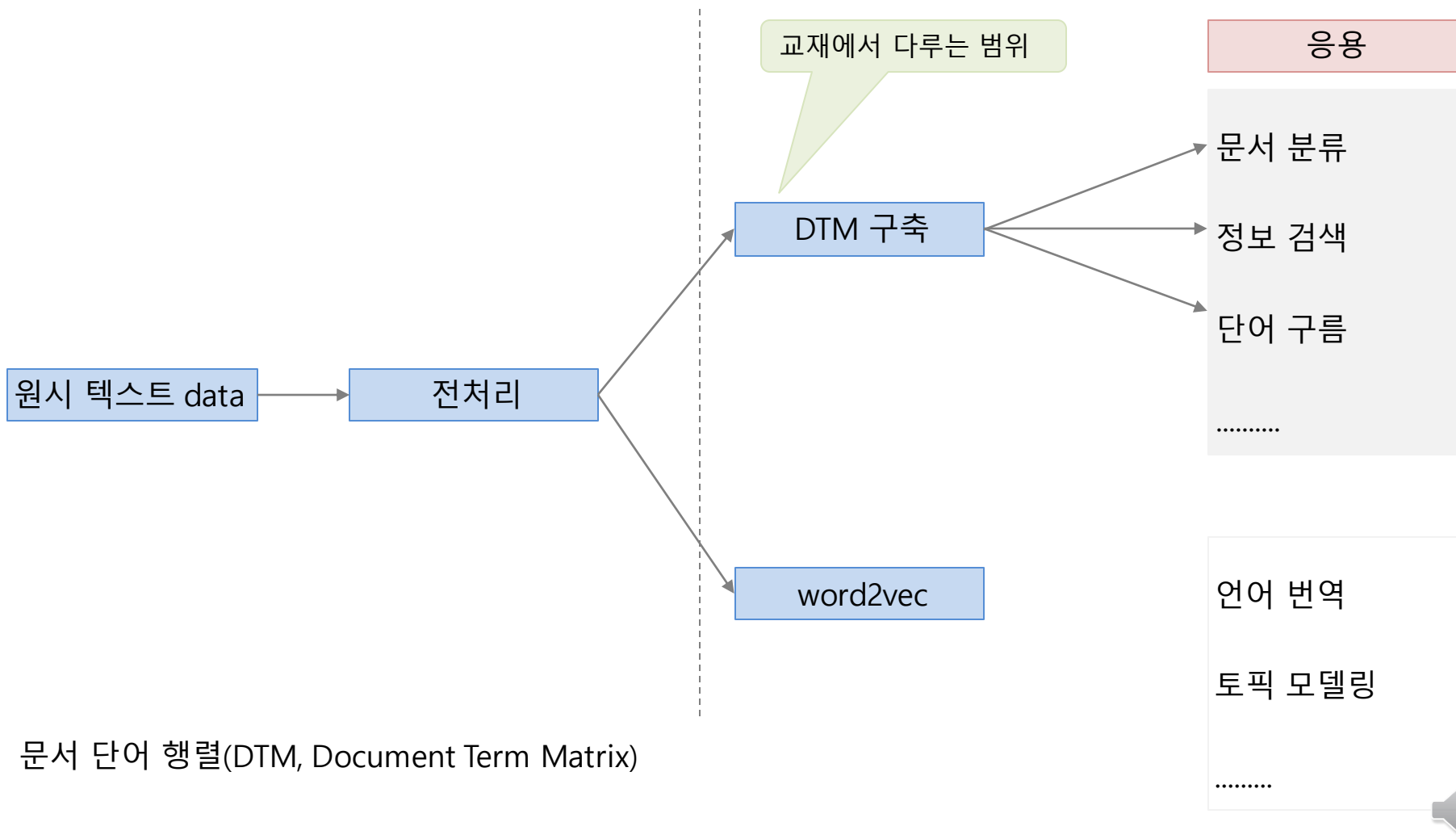
- 말뭉치(corpus)

- 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, 한국교통대학교 관련 6개월간 4대 일간지 기사, ICT 관련 정보통신 전문지 기사(2020년)



11.1 텍스트 마이닝 기초

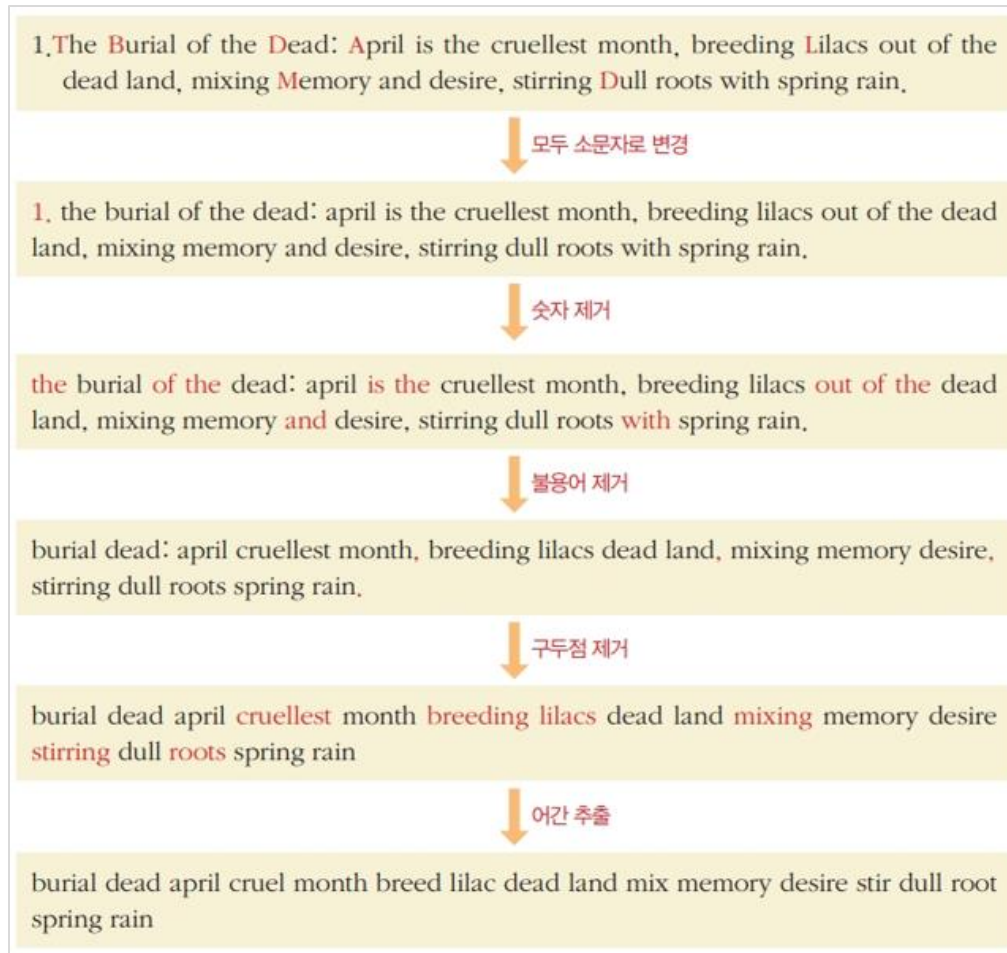
■ 텍스트 처리 파이프라인



11.1 텍스트 마이닝 기초

■ 전처리 과정

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



엘리엇의 황무지(The Waste Land)

1. 죽은 자의 매장

사월은 가장 잔인한 달
죽은 땅에서 라일락을 키워 내고
추억과 욕정이 뒤섞고
잠든 뿌리를 봄비로 깨운다.

불용어(stop word)란 검색 색인 단어로 의미가 없는 단어
예) a, the, and, or, 그리고, 또는, 및



■ 텍스트 데이터

- 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

※ 문서 단어 행렬

DWM(Document Word Matrix)라 하지 않고 DTM(Document Term Matrix)이라고 부르는 이유는 사전을 만들 때 단어만 대상으로 하지 않고 일반적으로 n-그램을 대상으로 하기 때문이다.

- n-그램 : 연속으로 나타나는 n개의 단어
- 예) "Data science is exciting and motivating" 의 2-그램 Data-science, science-is, is-exciting, exciting-and, and-motivating
- n-그램을 사용하면 단어가 나타나는 순서 파악에 장점이 있음



■ 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?”
- 먼저 전처리작업 : 소문자 변환, 구두점과 불용어 제거, 숫자 제거 등
- 사전(dictionary) 만들기 (문서에 나타난 단어를 모으면 사전이 됨)
- 예제에서는 9개의 단어가 사전을 구성
- 다음 표는 각각의 문장에 대해 발생 빈도를 채운 DTM이다


단어

↓

문서

→

	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



■ 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

$$DTM = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 2 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \leftarrow \text{행렬}$$



■ DTM 추가 설명

① 사전 구축

- 실제에서는 사전의 크기는 보통 수만~수십만
- **말뭉치**에서 수집할 수도 있고, 국어사전에 실려있는 단어 집합을 사용할 수도 있음

② 문서가 벡터로 표현되므로 거리 또는 유사성 측정 가능

- 앞의 예제에서 D1 벡터는 D2 벡터보다 D3 벡터와 가까움
- 랜덤 포리스트나 SVM 등의 적용이 가능해짐

③ 정규화 필요성

- 문서가 길면 벡터의 길이가 커져서 유사한 문서와 거리가 멀어짐
- 벡터의 길이를 1로 만드는 정규화를 적용하여 해결 가능(벡터의 크기로 나누어 줌)

④ DTM은 희소 행렬

- 한번도 발생하지 않아 0인 칸이 아주 많음

	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


⑤ 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



■ 실제 DTM 구축 : 예제, 위키피디아의 “data science” 문서

-
- https://en.wikipedia.org/wiki/Data_science



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute

Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

Article **Talk** Read Edit View history Search Wikipedia

Data science

From Wikipedia, the free encyclopedia

Not to be confused with [information science](#).

Data science is an [interdisciplinary](#) field that uses scientific methods, processes, algorithms and systems to extract [knowledge](#) and insights from structured and [unstructured data](#),^{[1][2]} and apply knowledge and actionable insights from data across a broad range of application domains. Data science is related to [data mining](#), [machine learning](#) and [big data](#).

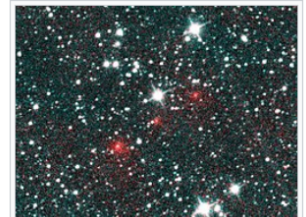
Data science is a "concept to unify [statistics](#), [data analysis](#), [informatics](#), and their related methods" in order to "understand and analyze actual phenomena" with data.^[3] It uses techniques and theories drawn from many fields within the context of [mathematics](#), [statistics](#), [computer science](#), [information science](#), and [domain knowledge](#). 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]}

Contents [hide]

- [Foundations](#)
 - [Relationship to statistics](#)
- [Etymology](#)
 - [Early usage](#)
 - [Modern usage](#)
- [Impact](#)
- [Technologies and techniques](#)
 - [Techniques](#)
- [References](#)

Foundations [edit]

Data science is an interdisciplinary field focused on extracting knowledge from data sets, which are typically large (see [big data](#)), and applying the knowledge and actionable insights from data to solve problems in a wide range of application domains.^[6] The field encompasses preparing data for analysis, formulating data science problems, analyzing data, developing data-driven solutions, and presenting findings to inform high-level decisions in a broad range of application domains. As such, it incorporates skills from computer science,



The existence of [Comet NEOWISE](#) (here depicted as a series of red dots) was discovered by analyzing [astronomical survey](#) data acquired by a [space telescope](#), the [Wide-field Infrared Survey Explorer](#).

Part of a series on Machine learning and data mining



Problems [show]



■ 실제 DTM 구축 : 예제, 위키피디아의 “data science” 문서

- RCurl, rvest 라이브러리로 웹 서버에 접속
- XML 라이브러리로 웹 문서 처리

Console C:/RSources/ ↗

```
> library(RCurl)
> library(XML)
> library(rvest)
>
> # url='https://en.wikipedia.org/wiki/Data_science'
> # t=read_html(url)
> t = readLines('https://en.wikipedia.org/wiki/Data_science',n=577)
> 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 변환해 줌

Console C:/RSources/ ↗

```
> # url='https://en.wikipedia.org/wiki/Data_science'
> # t=read_html(url)
> t = readLines('https://en.wikipedia.org/wiki/Data_science')
경고메시지(들):
In readLines("https://en.wikipedia.org/wiki/Data_science") :
  'https://en.wikipedia.org/wiki/Data_science'에서 불완전한 마지막 행이 발견되었습니다
```



■ 전처리 ([교재 그림 11-3] 참조)

- tm 라이브러리는 데이터 마이닝 함수 제공
- SnowballC 라이브러리는 어간을 추출하는 함수 제공

```
library(tm) 텍스트 마이닝의 여러가지 함수 제공
library(SnowballC) 어간을 추출하는 함수 제공

doc = Corpus(VectorSource(clean_doc)) 텍스트 문서 -> corpus 변환
inspect(doc) 말뭉치, 용어 문서 매트릭스 또는 텍스트 문서에 대한 자세한 정보를 표시합니다

# 문장을 매개변수 값에 따른 전처리 작업
doc = tm_map(doc, content_transformer(tolower)) # 소문자로 변화
doc = tm_map(doc, removeNumbers) # 숫자 제거
doc = tm_map(doc, removeWords, stopwords('english')) # 영어 불용어 제거
doc = tm_map(doc, removePunctuation) # 구두점 제거
doc = tm_map(doc, stripwhitespace) # 공백 제거
```

- tm_map 함수는 지정된 매개변수 값에 따라 전처리 수행



■ 전처리 전 말뭉치(corpus) data[7]

```
Console C:/RSources/
> doc = Corpus(VectorSource(clean_doc))
> doc
<<SimpleCorpus>>
Metadata: corpus specific: 1, document level (indexed): 0
Content: documents: 13
> class(doc)
[1] "SimpleCorpus" "Corpus"
> doc[[7]]$meta
author      : character(0)
datetimestamp: 2021-04-17 06:05:27
description  : character(0)
heading
id
language
origin
> doc[[7]]$con
[1] "The term
s an alternati
Classificatio
ce as a topic.
the Chinese A
tatistics shou
ics shed inacc
describing da
ciplinary conc
```

```
Console Jobs x
C:/RSources/
author      : character(0)
datetimestamp: 2021-05-18 01:18:51
description  : character(0)
heading      : character(0)
id           : 2
language     : en
origin       : character(0)
> doc[[2]]$content
[1] "data science interdisciplinary field uses scientific methods processes algorithms
systems extract knowledge insights structured unstructured data apply knowledge actio
nable insights data across broad range application domains data science related data m
ining machine learning big data "
```

■ 전처리 전 말뭉치(corpus) data[7]

> doc = tm_map(doc, content_transformer(tolower)) 대문자 -> 소문자

> doc[[7]]\$content

[1] "The term "data science" has been traced back to 1974, when Peter Naur proposed it as an alternative name for computer science.[21] In 1996, the International Federation of Classification Societies became the first conference to specifically feature data science as a topic.[21] However, the definition was still in flux. After the 1985 lecture in the Chinese Academy of Sciences in Beijing, in 1997 C.F. Jeff Wu again suggested that statistics should be renamed data science. He reasoned that a new name would help statistics shed inaccurate stereotypes, such as being synonymous with accounting, or limited to describing data.[22] In 1998, Hayashi Chikio argued for data science as a new, interdisciplinary concept, with three aspects: data design, collection, and analysis.[20]"

Console C:/RSources/ ↗

> doc[[7]]\$content

[1] "the term "data science" has been traced back to 1974, when peter naur proposed it as an alternative name for computer science.[21] in 1996, the international federation of classification societies became the first conference to specifically feature data science as a topic.[21] however, the definition was still in flux. after the 1985 lecture in the chinese academy of sciences in beijing, in 1997 c.f. jeff wu again suggested that statistics should be renamed data science. he reasoned that a new name would help statistics shed inaccurate stereotypes, such as being synonymous with accounting, or limited to describing data.[22] in 1998, hayashi chikio argued for data science as a new, interdisciplinary concept, with three aspects: data design, collection, and analysis.[20]"



■ 전처리 전 말뭉치(corpus) data[7]

```
> doc = tm_map(doc, removeNumbers)
```

 숫자 제거

```
> doc[[7]]$content
```

```
[1] "The term "data science" has been traced back to 1974, when Peter Naur proposed it  
t as an alternative name for computer science.[21] In 1996, the International Federat  
ion of Classification Societies became the first conference to specifically feature d  
ata science as a topic.[21] However, the definition was still in flux. After the 1985  
lecture in the Chinese Academy of Sciences in Beijing, in 1997 C.F. Jeff Wu again su  
ggested that statistics should be renamed data science. He reasoned that a new name w  
ould help statistics shed inaccurate stereotypes, such as being synonymous with accou  
nting, or limited to describing data.[22] In 1998, Hayashi Chikio argued for data sci  
ence as a new, interdisciplinary concept, with three aspects: data design, collectio  
n, and analysis.[20]"
```

```
> doc[[7]]$content
```

```
[1] "the term "data science" has been traced back to , when peter naur proposed it as  
an alternative name for computer science.[21] in , the international federation of cla  
ssification societies became the first conference to specifically feature data scienc  
e as a topic.[21] however, the definition was still in flux. after the lecture in the  
chinese academy of sciences in beijing, in c.f. jeff wu again suggested that statis  
tics should be renamed data science. he reasoned that a new name would help statistic  
s shed inaccurate stereotypes, such as being synonymous with accounting, or limited t  
o describing data.[22] in , hayashi chikio argued for data science as a new, interdisci  
plinary concept, with three aspects: data design, collection, and analysis.[20]"
```



11.2 DTM 구축

■ 전처리 전 말뭉치(corpus) data[7]

> doc = tm_map(doc, removeWords, stopwords('english')) 영어 불용어 제거

Console C:/RSources/ ↗

> doc[[7]]\$content

[1] "the term "data science" has been traced back to , when peter naur proposed it as an alternative name for computer science.[] in , the international federation of classification societies became the first conference to specifically feature data science as a topic.[] however, the definition was still in flux. after the lecture in the chinese academy of sciences in beijing, in c.f. jeff wu again suggested that statistics should be renamed data science. he reasoned that a new name would help statistics shed inaccurate stereotypes, such as being synonymous with accounting, or limited to describing data.[] in , hayashi chikio argued for data science as a new, interdisciplinary concept, with three aspects: data design, collection, and analysis.[]"

[1] " term "data science" traced back , peter naur proposed alternative name computer science.[] , international federation classification societies became first conference specifically feature data science topic.[] however, definition still flux. lecture chinese academy sciences beijing, c.f. jeff wu suggested statistics renamed data science. reasoned new name help statistics shed inaccurate stereotypes, synonymous accounting, limited describing data.[] , hayashi chikio argued data science new, interdisciplinary concept, three aspects: data design, collection, analysis.[]"



11.2 DTM 구축

■ 전처리 전 말뭉치(corpus) data[7]

> doc = tm_map(doc, removePunctuation) 구두점, 특수문자 등 제거

```
> doc[[7]]$content
[1] " term "data science" traced back , peter naur proposed alternative name
computer science [], international federation classification societies became f
irst conference specifically feature data science topic.[] however, definition s
till flux. lecture chinese academy sciences beijing, c.f. jeff wu suggeste
d statistics renamed data science. reasoned new name help statistics shed inac
curate stereotypes, synonymous accounting, limited describing data.[] , hayash
i chikio argued data science new, interdisciplinary concept, three aspects: data
design, collection, analysis.[]"
```

```
> doc[[7]]$content
[1] " term "data science" traced back peter naur proposed alternative name c
omputer science international federation classification societies became first c
onference specifically feature data science topic however definition still flux
lecture chinese academy sciences beijing f jeff wu suggested statistics
renamed data science reasoned new name help statistics shed inaccurate stereotyp
es synonymous accounting limited describing data hayashi chikio argued data
science new interdisciplinary concept three aspects data design collection analy
sis"
```



11.2 DTM 구축

■ 전처리 전 말뭉치(corpus) data[7]

> doc = tm_map(doc, stripWhitespace) 공백문자 제거

```
> doc[[7]]$content
```


```
[1] " term "data science"   traced back   peter naur proposed   alternative name c  
omputer science   international federation   classification societies became first c  
onference   specifically feature data science   topic however definition still flux  
   lecture   chinese academy sciences beijing   cf jeff wu suggested statistics  
   renamed data science reasoned   new name help statistics shed inaccurate stereoty  
pes   synonymous accounting limited describing data   hayashi chikio argued data  
science   new interdisciplinary concept   three aspects data design collection analy  
sis"
```

```
> doc[[7]]$content
```

```
[1] " term "data science" traced back peter naur proposed alternative name computer s  
cience international federation classification societies became first conference spec  
ifically feature data science topic however definition still flux lecture chinese aca  
demy sciences beijing cf jeff wu suggested statistics renamed data science reasoned n  
ew name help statistics shed inaccurate stereotypes synonymous accounting limited des  
cribing data hayashi chikio argued data science new interdisciplinary concept three a  
spects data design collection analysis"
```



■ DocumentTermMatrix 함수로 DTM 구축 (전처리 후)

```
Console C:/RSources/ 
> dtm = DocumentTermMatrix(doc)
> dim(dtm)
[1] 13 363
> inspect(dtm)
```

dim 함수는 dtm의 행과 열의 개수를 알려줌

inspect 함수는 중요 문서 10개와 중요 단어 10개를 보여줌 (중요성은 발생 빈도로 결정)

<<DocumentTermMatrix (documents: 13, terms: 363)>> 위키에 있는 문장 13개 각각을 문서로 간주함
363개의 단어를 추출하여 사전 구축

Non-/sparse entries: 480/4239

Sparsity : 90%

Maximal term length: 17

weighting : term frequency (tf)

sample :

Terms

Docs	big	data	field	information	knowledge	learning	name	new	science	statistics
10	0	2	0	0	0	0	0	0	1	0
12	4	6	0	1	0	0	0	1	1	0
13	0	2	0	0	0	1	0	0	2	0
2	1	6	1	0	2	1	0	0	2	0
3	0	5	0	2	1	0	0	0	6	2
4	1	10	2	2	2	1	0	0	5	2
5	0	11	2	0	0	0	1	1	8	8
6	0	4	1	0	0	0	1	1	2	3
7	0	5	0	0	0	0	2	2	4	2
9	0	8	1	0	0	2	2	1	7	1



Thank you

