



[2020 혁신성장 청년인재 집중양성 사업]

프로젝트 기반 데이터 과학자 양성과정

빅데이터 분석

- 3주차 -

#정형 데이터 처리 #문자 처리 #시공간 처리



A table of Contents

- 1 데이터 처리 패키지
- 2 데이터 조회
- 3 시간 데이터 처리
- 4 문자 데이터 처리
- 5 공간 데이터 처리
- 6 Summary

Part 1

데이터 처리 패키지



apply 계열 함수

apply 계열 함수

- 벡터, 행렬 또는 데이터 프레임에 임의의 함수를 적용한 결과를 얻기 위한 함수
- 데이터 전체에 함수를 한 번에 적용하는 벡터 연산을 수행함으로 속도가 빠름

함수	설명	다른 함수와 비교했을 때의 특징
apply()	배열 또는 행렬에 주어진 함수를 적용한 뒤 그 결과를 벡터, 배열 또는 리스트로 반환	배열 또는 행렬에 적용
lapply()	벡터, 리스트 또는 표현식에 함수를 적용하여 그 결과를 리스트로 반환	결과가 리스트
sapply()	lapply와 유사하지만 결과를 벡터, 행렬 또는 배열로 반환	결과가 벡터, 행렬 또는 배열
tapply()	벡터에 있는 데이터를 특정 기준에 따라 그룹으로 묶은 뒤 각 그룹마다 주어진 함수를 적용하고 그 결과를 반환	데이터를 그룹으로 묶은 뒤 함수를 적용
mapply()	sapply의 확장된 버전으로, 여러 개의 벡터 또는 리스트를 인자로 받아 함수에 각 데이터의 첫째 요소들을 적용한 결과, 둘째 요소들을 적용한 결과, 셋째 요소들을 적용한 결과 등을 반환	여러 데이터를 함수의 인자로 적용



유용한 패키지

Plyr 패키지

- 데이터 분할, 특정 함수 적용, 재조합하는 함수 제공
- `adply(배열 > 데이터 프레임)`, `ddply(데이터 프레임 > 데이터 프레임)`
- `mutate(새로운 컬럼을 추가하거나 기존 컬럼 수정)`, `summarise(데이터 요약)`

reshape2 패키지

- 데이터 모양을 바꾸거나 그룹별 요약값을 계산하는 함수가 포함

함수	의미
<code>melt()</code>	여러 컬럼으로 구성된 데이터를 데이터 식별자(id), 측정 변수(variable), 측정값(value)이라는 3개 컬럼으로 변환한다. 만약 한 데이터에 대해 다수의 측정 변수와 측정값이 있다면 이들은 여러 행으로 표현된다. 이렇게 변환된 결과는 variable 컬럼에 측정 대상이 기록되어 있으므로 각 variable마다 value의 통계 값을 계산하는 것이 편리하다.
<code>cast()</code>	<code>melt()</code> 된 데이터를 다시 여러 컬럼으로 변환한다. 데이터에 여러 측정 변수와 측정값이 존재한다면 이들은 모두 새로운 컬럼으로 변환된다. <code>cast()</code> 로 변환된 결과는 마치 스프레드시트에 입력한 데이터 모양과 유사하므로 분석자가 읽기 쉽다. 또한, <code>cast()</code> 시 <code>melt()</code> 된 데이터의 여러 행이 한 셀에 대응하는 경우 데이터의 요약 값을 자동으로 계산해준다.



유용한 패키지

dplyr 패키지

- 데이터 분할, 특정 함수 적용, 재조합하는 함수 제공
- Group by를 추가로 이용하면 그룹별로 다양한 집계 가능

함수명	내용	비고
Filter()	조건에 맞는 데이터 추출	Subset()
select()	열의 추출	Data["Year","Month"]
mutate()	열 추가	Transform()
Arrange()	정렬	order(),sort()
Summarise()	집계	aggregete()

A grayscale background image of a workspace. In the upper right, a portion of a laptop keyboard is visible. In the center, an open notebook with horizontal lines lies flat, with a silver ballpoint pen resting diagonally across its pages. In the lower right, a smartphone is partially visible, its screen dark. A large, dark gray diagonal shape cuts across the right side of the image. On the far left, a solid teal vertical bar runs the full height of the frame.

Part 2

데이터 조회



Part 3

시간 데이터 처리



시간 데이터 처리

시간 데이터 구조

- 날짜 및 시간 데이터
- 수치와 범주 다음으로 자주 다루는 데이터

구분	타입	비고
날짜형	Date	
일시형	POSIXct	• 일시형끼리 비교하거나 시간 차를 계산할 때 유용
	POSIXlt	• 연, 월, 시, 분, 초를 각각 가진다 • 여러 값을 내부 리스트로 가진 형태로 dplyr 처리가 불가

Part 4

문자 데이터 처리

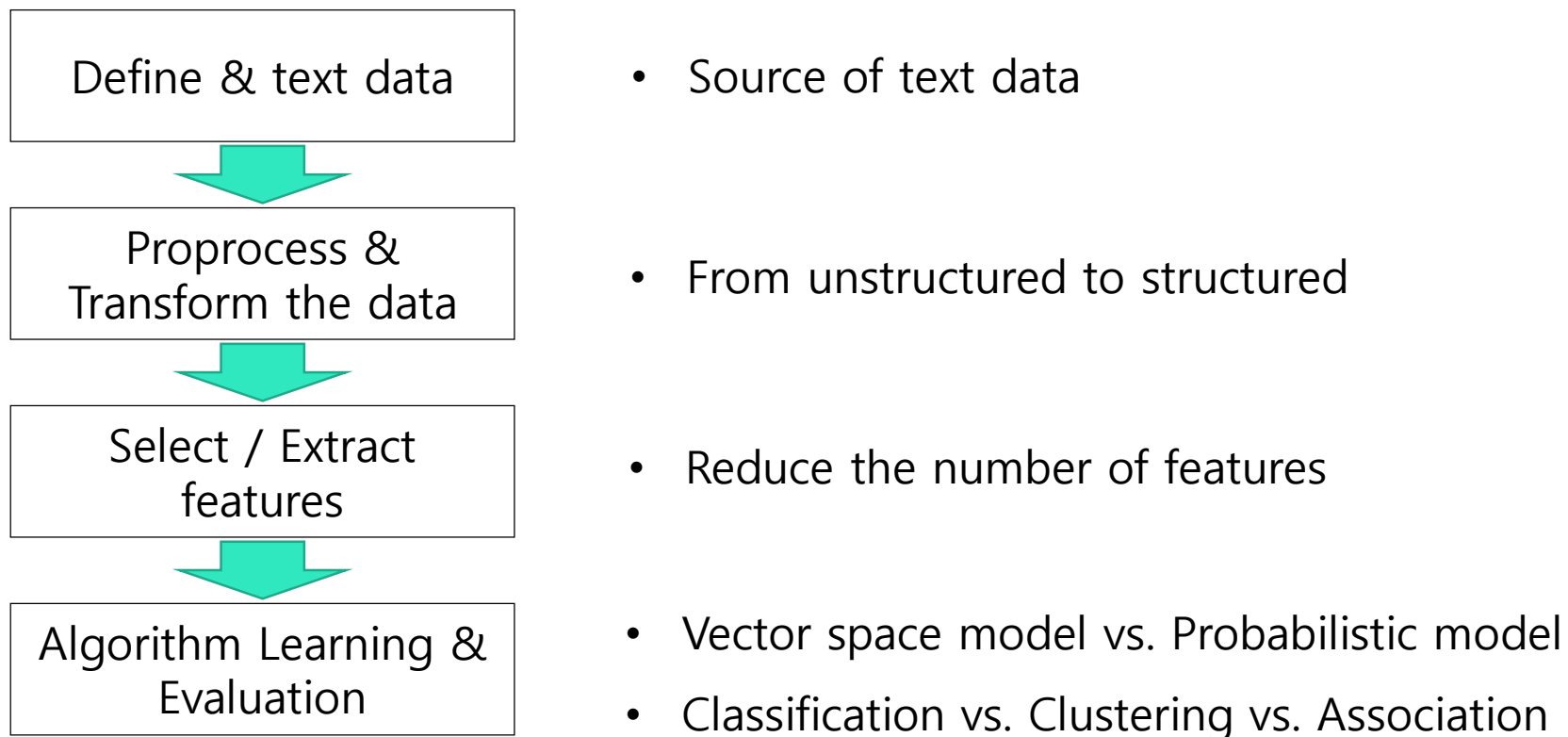




텍스트 마이닝

텍스트 마이닝이란

- 텍스트로부터 유의미한 정보를 추출
- 단어의 출현 빈도와 단어 간의 관계성 등을 파악





Part 5

공간 데이터 처리



공간 자료 개요

공간 자료란?

- 과거에는 종이 지도를 이용하여 출발지, 도착지, 통행 경로 등을 2차원으로 파악
- 현재는 디지털 맵을 검색하여 최적 경로, 교통상황 등 3차원으로 파악 가능
- 평면에 투영하는 방법에 따라 다양한 좌표계가 적용

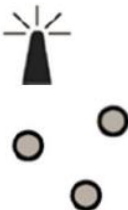


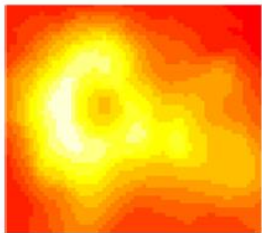
준거 타원체	사용국가
WGS84(1984)	세계적 적용
Everest(1830)	인도
Bessel(1841)	한국, 일본, 독일
Airy(1844)	영국
Clarke(1866)	북아메리카
GRS80(1980)	국제적 채택
Krasovsky(1938)	러시아



공간 자료 개요

공간 자료의 구조

- 벡터(Vector) 형식 : 공간에서의 기하 구조를 점(Point), 선(Line), 면(Polygon)으로 표현
- 래스터(Raster) 형식 : 공간상에서 기하 구조를 Grid Data(or Pixel Data)로 불리는 격자형 Cell로 표현
- 도면이나 지도와 같은 도형적 요소의 유형, 위치, 크기, 다른 지형요소와의 공간적 위상 관계를 나타내는 도형 정보와 지형 요소에 관련된 서술적 특성을 나타내는 속성 정보로 구분

벡터(Vector) 형식			래스터(Raster)
점(point)	선(line)	면(polygon)	격자(grid)
			

A grayscale photograph of a workspace. In the background, a laptop is partially visible. In the foreground, a silver pen rests on an open, lined notebook. To the right of the notebook, a smartphone lies flat. A solid teal vertical bar is positioned on the far left side of the image.

Part 6

Summary