

사회학자를 위한 데이터사이언스
INTRODUCTION TO SOCIAL DATA SCIENCE
(3-3-0)

2023년 2학기

Instructor	김현우, PhD (hxk271@cbnu.ac.kr)
Office	N15동 414호
Classroom	N15동 407호(사회조사실습실)
Class hours	01:00 PM – 4:50 PM, 금요일
Course website	https://github.com/hxk271/DataSciSoc

강의 개요

종종 4차 산업혁명과 초연결사회로 은유되는 현대사회에서 통계분석과 데이터 사이언스는 세부적인 전공 영역을 뛰어넘어 창의적 인재의 일반 소양으로 거듭나고 있습니다. 이 분야는 학술 연구 뿐 아니라 정부 정책의 제안 및 비판, 데이터 저널리즘(data journalism), 헬스케어(health care), 비영리단체/기업 운영관리 등 현실적 문제에서도 그 가능성을 폭넓게 증명해 왔습니다. 그러나 막상 현실적인 문제에 적용될 때 인문사회학적 소양을 함께 지닌 전문가의 부족으로 인해 학계와 산업체에서는 최근까지도 많은 아쉬움을 겪어왔습니다. 이에 따라 사회학도에게도 데이터 사이언스와 코딩 교육, 그리고 통계학에 관한 지식과 능력이 요구되고 있습니다.

데이터 사이언티스트에게는 코딩(coding), 수학(mathematics), 그리고 전공 지식(domain knowledge)이라는 세 가지 스킬셋이 요구됩니다. 이 수업은 특히 수학적 기초와 코딩 경험이 튼튼하지 않은 사회학자를 위해 설계되었습니다. 수학은 알고리즘에 관한 직관적 이해를 돋는 정도로만 사용되고, 코딩은 알고리즘을 손쉽게 구현하는 예제를 중심으로 짜여집니다. 그러므로 수학이나 컴퓨터 코딩이 두려워 이 수업을 기피할 필요는 없습니다.

이 수업은 사회학을 전공하는 대학원생을 대상으로 데이터 사이언스의 기초적인 아이디어와 논리, 그리고 알고리즘을 소개합니다. 이 수업에서 우리는 다음의 세 가지 토픽을 차례로 학습합니다.

1. 데이터 사이언스의 소개와 파이썬(Python) 입문
2. 데이터 전처리, API 및 웹 스크래핑
3. 텍스트 분석과 자연어처리(natural language processing)의 기초

이 수업은 일차적으로 데이터 사이언스의 논리와 기법을 다루지만, 수강생 여러분은 자신만의 연구 대상에 이 논리와 기법이 어떻게 적용될 수 있는지 끊임없이 고민해야 합니다. 데이터 사이언스의 논리와 기법을 사회학적 연구에 잘 활용하려면 그 자체로 많은 탐색을 필요로 하기 때문입니다.

선수 과목

학부에서 사회통계학과 사회조사방법론을 반드시 이수해야 합니다. 데이터 사이언스 분야의 타 전공 학생은 최종 학점에서 일정한 패널티를 받게 됩니다.

강의 교재

이 수업에는 세 종류의 교재가 필요합니다. 먼저 파이썬 입문 교재로 박해선(2023)을 사용합니다. 복습을 위해 필수적이므로 반드시 구비하세요. 송석리·이현아 (2019)은 우리 수업 전반부에서 사용하지만 구입이 필수는 아닙니다.

- 박해선. 2023. 『혼자 공부하는 데이터 분석 with 파이썬』. 한빛. (필수)
- 송석리·이현아. 2019. 『모두의 데이터 분석 with 파이썬』. 길벗. (참고)

다음으로 텍스트 분석 교재가 필요합니다. 우리는 박조은·송영숙(2023)을 기본 교재로 삼습니다. 그 밖의 책들도 제법 괜찮으므로, 텍스트 분석을 깊이있게 공부할 생각이라면 모두 손에 넣는 것을 추천합니다.

- 윤태일·이수안. 2018. 『파이썬으로 텍스트 분석하기: 전략커뮤니케이션을 위한 텍스트 마이닝』. 늘봄. (필수)
- 박조은·송영숙. 2023. 『모두의 한국어 텍스트 분석 with 파이썬 - 기초부터 챗GPT까지, 누구나 쉽게 시작하는 자연어 처리』. 길벗. (참고)
- 서대호. 2019. 『잡아라! 텍스트 마이닝 with 파이썬』. 비제이퍼블릭. (참고)

우리 수업에서 별도로 다루지 않지만 텍스트 분석을 사회과학연구에서 제대로 응용하기 위해서는 반드시 내용분석(content analysis)을 공부해야 합니다. 아쉽게도 이 분야의 좋은 한국어 교재를 찾아보기 어렵습니다. 그나마 라이프, 레이시, 피코(2001)가 괜찮은 것 같습니다.

- 라이프, 대니얼·스티븐 레이시·프레드릭 피코. 2001. 『미디어 내용분석 방법론』. 커뮤니케이션북스. (참고)

예비수강생의 요청에 따라 기계학습(machine learning)과 그 알고리즘은 우리 수업에서 다루어지지 않습니다. 이는 치명적인 결함이므로 그나마 최대한 보강하기 위해 다음 교재를 반드시 구비하세요.

- 응, 애널린·케네스 수. 2006. 『수학없이 배우는 데이터 사이언스』. 에이콘. (필수)

학습 보조자료

- Python은 오픈소스이고 공식 튜토리얼은 무료입니다(<https://docs.python.org/3/tutorial>). 설명이 초보자에게는 다소 지나치게 세부적이지만 예제가 매우 파이썬 답습니다(“Pythonic”). 머신러닝 알고리즘을 구현하는데 필요한 라이브러리인 sklearn 역시 오픈소스로 마찬가지로 무료 공식 튜토리얼을 가지고 있습니다(<https://scikit-learn.org>). 설명을 온전히 이해하기 위해서는 약간의 수학 지식을 요구하지만 예제만으로도 충분히 훌륭합니다.
- 유튜브(<https://www.youtube.com>)나 Udacity(<https://www.udacity.com>) 등에서는 수많은 무료 Python 강의가 열려 있습니다. 영어를 겁내지 말고 (필요하다면) 캡션을 사용하세요.
- 구글(<https://www.google.com>)과 스택오버플로(<https://stackoverflow.com>)는 학부 1학년부터 업계 시니어에 이르기까지 모든 데이터 사이언티스트의 가장 든든한 우군입니다. 잘 모를 때는 질문을 영어로 옮겨 검색하세요. 초보자가 가질만한 거의 모든 질문은 이미 누군가가 던졌고 게다가 대답도 있을 가능성이 높습니다. 챗GPT(<https://chat.openai.com>)도 여기에 합류했습니다!

강의 구성

- 강의: 모든 학생은 반드시 수업에 참여해야 합니다. 수업을 통해 기초적인 개념과 알고리즘을 배우고 컴퓨터를 사용하여 실습합니다. 수업 내용은 진행될수록 누적되기 때문에 결석은 향후 이해에 큰 방해가 됩니다.
- 실습: 모든 학생은 반드시 실습에 참여해야 합니다. 전산실에서 Python을 사용하여 실제 코딩을 진행합니다. 배운 내용을 집에서도 복습해야 할 뿐 아니라 교재에서 할당된 부분도 별도로 공부해야 합니다. 중간에 포기하면 아무 의미도 없게 되므로 이해가 갈때까지 모든 수단을 동원하여 탐구를 거듭해야 합니다.
- 과제: 매주 읽을거리를 수업 전에 모두 읽고 한 번 이상 연습해야 합니다. 응·수(2006)를 읽고 사회학 연구에의 적용, 질문 및 토론거리를 1페이지로 준비하여 발표하고 제출합니다.
- 시험: 모두 두 개의 시험이 7주차와 15주차에 실시됩니다(각 30%). 문제가 주어지면 이를 코딩으로 풀어보이고 해설을 제시하는 방식입니다.

학점 안내사항

최종 학점은 다음 기준에 따라 산출됩니다.

- 출석 및 수업 참여 (20%)
- 과제 (20%)
- 시험 2개 (60%)

수강생 유의사항

- 모든 수업은 별도의 안내가 없는 이상 원칙적으로 **대면으로** 진행됩니다.
- 수업 전일에 해당 주차 강의안과 자료가 다음 GitHub 레파지토리에 업로드됩니다:
<https://github.com/hxk271/DataSciSoc>
- 공결은 증빙서류를 모두 갖추어 담당교수에게 **직접** 제출된 경우만 인정하며, 개별적인 공결 행정처리는 일체 **무효임**에 주의하십시오.
- 2회 지각은 1회 결석으로 처리합니다. (첫 수업 및 공결을 포함하여) 3회 이상 결석한 경우 무조건 F이며 예외는 없습니다. 불가피한 사정으로 공결하였을 경우 그 외 일체 지각이나 결석을 하지 않아야 하겠습니다.
- 과제는 선택사항이 아닙니다. 과제를 4회 이상 제출하지 않으면 무조건 F이며 예외는 없습니다.
- 과제를 **표절하거나 표절당하지 않도록** 주의를 기울이십시오. 표절에 조력한 자와 표절한 자 모두 **경고없이** 과제 2개 분량을 무효처리합니다. 그러므로 과제 파일을 전산실에 남기는 등 표절에 가담할 여지를 남기지 마십시오.
- 시험 관련 부정행위자로 판명되었을 때는 학칙 또는 내규에 의거 해당 교과목의 성적을 취소합니다.

장애인 학생 수업안내

장애학생은 본 수업과 관련하여 본인 희망시 다음과 같은 지원이 가능합니다. 담당교수 및 장애학생지원센터와 상담 바랍니다.

- 공통: 도우미 지원(수업, 이동), 대체평가, 별도 발표/시험장소 제공, 선수강 지원, 노트북 사용
- 시각장애: 점자/확대/녹음 교재 및 시험지 제공, 발표/시험시간 연장, 강의자료 텍스트제공
- 청각장애: 지정좌석제, 동영상 자막지원
- 지체장애: 강의실 변경, 지정좌석제, 발표/시험시간 연장

토픽 개요

파이썬 입문	1주차	데이터 사이언스 소개
	2주차	Hello World in Python
	3주차	조건문과 반복문
	4주차	matplotlib를 이용한 시각화
	5주차	공공데이터 분석실습
	6주차	중간시험

목표: 파이썬을 이용하여 간단한 공공데이터를 분석하고 시각화를 실습한다.

데이터 전처리 API 사용법 및 웹 스크래핑	7주차	numpy와 pandas 입문
	8주차	API 호출하기
	9주차	웹 스크래핑
	10주차	자료 전처리

목표: 자료수집과 전처리의 기법을 익힌다

텍스트 분석	11주차	텍스트 전처리
	12주차	키워드 빈도 분석
	13주차	감정분석
	14주차	토픽모델링과 한국어 임베딩
	15주차	기말시험

목표: 자연어처리의 기본 논리와 기법을 습득한다.

세부 일정

1주차	<u>데이터 사이언스의 소개</u>
TOPICS	데이터 사이언스의 구성요소; 산업혁명들과 4차 산업혁명; 데이터 분석가, 데이터 엔지니어, 그리고 데이터 사이언티스트; 구글 콜랩(Colab); 스크립트 언어로서 Python
READINGS	윤태일·이수안 1장, 2장; 박해선 01-1, 01-2; 응·수 1장
2주차	<u>Hello World in Python</u>
TOPICS	문자열(string)과 슬라이싱(slicing); 자료 타입(lists, tuples, sets, dicts); 파일과 자료의 입출력(I/O); 내장함수(built-in functions)
READINGS	응·수 2장
3주차	<u>조건문과 반복문</u>
TOPICS	연산자(Operators); 조건문(if, elif, else); 반복문(for, range, while); 계속과 이탈(continue, pass, break); 단순대입 탐색(brute-force search)
READINGS	윤태일·이수안 3장; 응·수 3장
4주차	<u>matplotlib를 사용한 시각화</u>
TOPICS	시각화의 힘; 히스토그램(histogram); 상자-수염 그림(Box-Whisker Plot); 산점도(scatterplot); matplotlib
READINGS	응·수 4장
5주차	<u>공공데이터 분석실습</u>
TOPICS	반복문과 조건문 활용; 시계열 그래프(time-series line chart)
READINGS	응·수 5장

세부 일정 (계속)

6주차 중간시험

7주차 numpy와 pandas 입문

TOPICS DataFrame; numpy—pandas 변환; 정렬(sort)

READINGS 윤태일·이수안 4장; 박해선 01-3; 응·수 6장

8주차 API 호출하기

TOPICS Application Programming Interface (API); JSON과 XML 포맷

READINGS 윤태일·이수안 5장; 박해선 02-1, 05-1; 응·수 7장

9주차 웹 스크래핑

TOPICS Application Programming Interface (API); JSON과 XML 데이터 다루기

READINGS 박해선 02-2, 05-2; 응·수 8장

10주차 자료 전처리

TOPICS 인덱싱(indexing); 피쳐 엔지니어링(feature engineering); 데이터 결합(merge, concat, join); 집계(aggregation)와 groupby; 기술통계; 재배열(reshaping)

READINGS 박해선 03-1, 03-2; 응·수 9장

세부 일정 (계속)

11주차 키워드 빈도

TOPICS Keywords in Context

READINGS 윤태일·이수안 6장, 7장; 박해선 04-1; 응·수 10장

12주차 감정분석

TOPICS Lexicon-Based Sentiment Analysis

READINGS 윤태일·이수안 11장; 박해선 04-2; 응·수 11장

13주차 의미망 분석

TOPICS Semantic Network; 노드(nodes)와 엣지(edge); 중심성(centrality); 군집(cluster)과 커뮤니티(community); NetworkX

READINGS 윤태일·이수안 8장, 9장; 박해선 05-1; 응·수 12장

14주차 토픽 모델링과 한국어 임베딩

TOPICS embedding, *Word2Vec*

READINGS 윤태일·이수안 10장; 박해선 05-2; 응·수 부록

15주차 기말시험

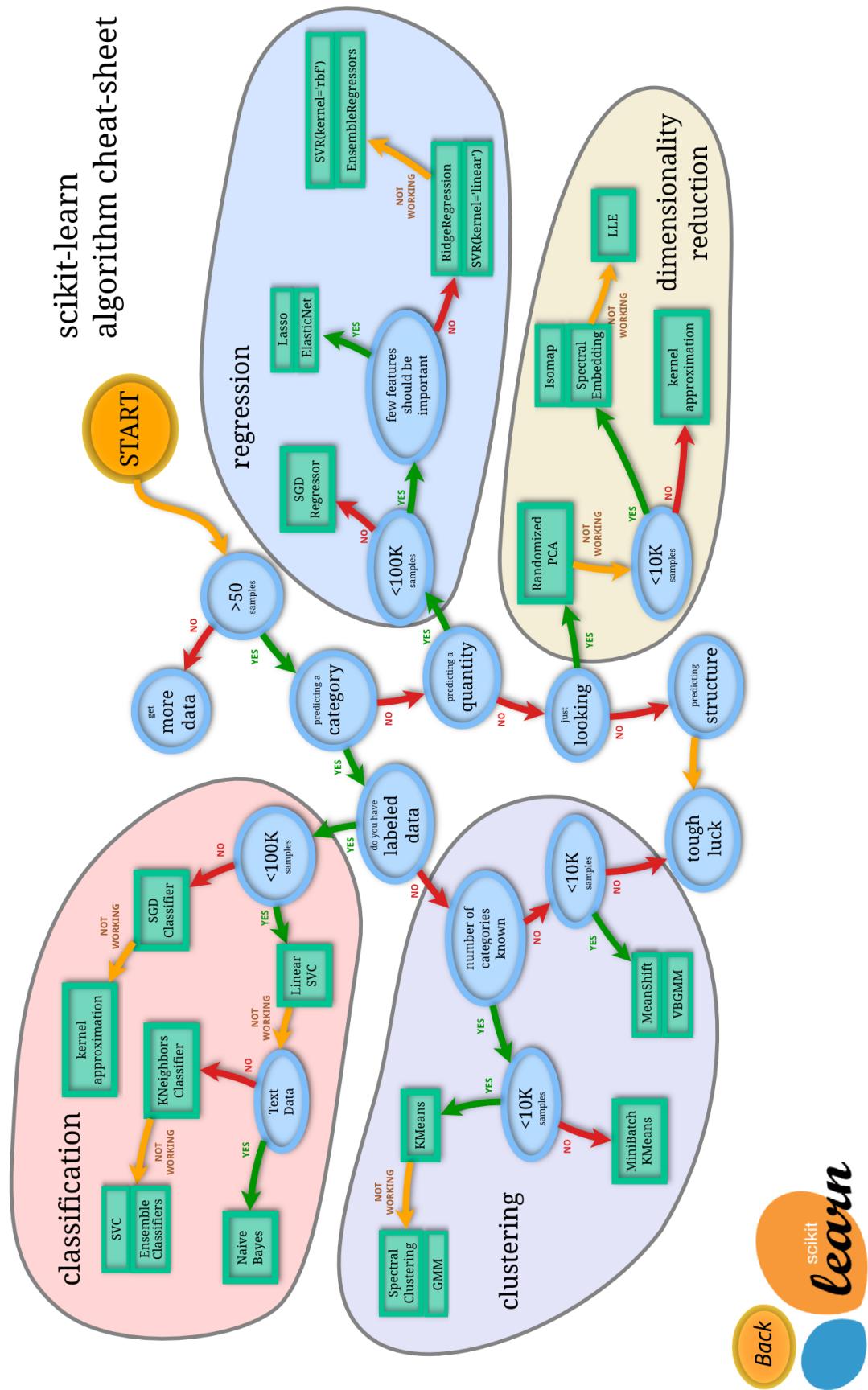


Figure 1: sklearn을 통해 구현할 수 있는 알고리즘