O'REILLY®

밑바닥부터 시작하는

데이터 과학 2판

데이터 분석을 위한 파이썬 프로그래밍과 수학·통계 기초

프로그래밍 인사이트

조엘 그루스 지음
김한결, 하성주, 박은정 옮김

지은이 **조엘 그루스** Joel Grus
Allen Institute for Artificial Intelligence의 연구원이다. 이전에는 구글에서 소프트웨어 엔지니어로, 여러 스타트업에서는 데이터 과학자로 근무했다. 현재 시애틀에 거주하면서 여러 개의 데이터 과학 밋업에 참여하고 있다. joelgrus.com에 종종 글을 쓰며, @joelgrus라는 계정으로 종일 트위터를 한다.

옮긴이 **김한결** https://hank110.github.io
KAIST에서 산업 및 시스템공학을 전공하고 서울대학교 산업공학과에서 석사 학위를 받았다. 에너지, 자동차, 금융, 제조 등 다양한 분야에서 데이터 분석 프로젝트를 진행했다. 《D3를 이용한 시각적 스토리텔링》(인사이트)을 공역했다.

옮긴이 **하성주** http://shurain.net
서울대학교 컴퓨터공학부 최적화 연구실에서 박사 학위를 받았다. 데이터를 바탕으로 지속적인 개선을 달성하는 방법에 관심이 있다.

옮긴이 **박은정** https://lucypark.kr
서울대학교에서 다양한 도메인의 텍스트를 다룬 후 2016년, 데이터마이닝으로 박사 학위를 받았다. 졸업 직후 네이버에 입사해서, 현재는 기계번역 모델을 개발하며 언어 장벽을 낮추기 위해 노력하고 있다.

밑바닥부터 시작하는 **데이터 과학** 2판

데이터 분석을 위한 파이썬 프로그래밍과 수학·통계 기초

Data Science from Scratch 2/E
by Joel Grus

밑바닥부터 시작하는 데이터 과학 2판

초판 1쇄 발행 2016년 5월 31일 **초판 5쇄 발행** 2018년 10월 22일 **2판 1쇄 발행** 2020년 3월 12일 **지은이** 조엘 그루스 **옮긴이** 김한결·하성주·박은정 **펴낸이** 한기성 **펴낸곳** 인사이트 **편집** 이지연 **본문 조판** 최우정 **제작·관리** 신승준·박미경 **용지** 월드페이퍼 **출력·인쇄** 현문인쇄 **제본** 자현제책 **등록번호** 제2002-000049호 **등록일자** 2002년 2월 19일 **주소** 서울시 마포구 연남로5길 19-5 **전화** 02-322-5143 **팩스** 02-3143-5579 **블로그** http://blog.insightbook.co.kr **이메일** insight@insightbook.co.kr
ISBN 978-89-6626-257-1 93000 책값은 뒤표지에 있습니다. 이 책의 정오표는 https://github.com/insightbook/data-science-from-scratch/wiki/Errata-(2nd-Edition)에서 확인하실 수 있습니다. 이 도서의 국립중앙도서관 출판예정도서목록(CIP)은 서지정보유통지원시스템 홈페이지(http://seoji.nl.go.kr)와 국가자료공동목록시스템(http://www.nl.go.kr/kolisnet)에서 이용하실 수 있습니다. (CIP제어번호: CIP2020002009)

프로그래밍 인사이트

《밑바닥부터 시작하는 데이터 과학 2판》 미리보기

차례

추천사

'세계적인'이라고 표현하면 별로 좋아하지 않을 것 같은, 세계적인 데이터 과학자이자 파워 트위터리안 조엘 그루스의《밑바닥부터 시작하는 데이터 과학 2판》의 추천사를 쓰는 내가 자랑스럽다. 게다가 이 책의 번역자들은 하나같이 우리나라 인공지능 영역의 젊은 거인들이 아닌가? 뭔가 퍼거슨 전 맨유 감독이 지휘하고 박지성과 손흥민이 뛰는 경기에 귀빈으로 초대받은 기분이다. 내가 이렇게 대단한 사람인가...

감상은 뒤로하고 지금 우리의 필드를 돌아보자. 빅데이터, 인공지능, 머신러닝, 데이터 과학... 이제 더는 일반인에게도 생소하지 않은 이 단어들은 우리 생활 속에 자리 잡고 있다. 많은 학생과 개발자 들이 데이터 과학자가 되기 위해 공부하고 이를 통해 새로운 기회를 찾고 있다. 텐서플로, 케라스, 사이킷런 등 다양한 도구를 사용하여 캐글이나 국내외 데이터 분석 사이트에서 자신의 실력을 뽐낼 수도 있다. 요즘엔 초등학생도 "딥러닝을 한다. 데이터 과학을 한다"고 할 정도로 대중화되고 있다.

하지만 좀 더 자세히 들여다보자. 우리는 도대체 얼마나 알고 있는가? 얼마나 설명할 수 있을까? 예를 들어 텍스트 간의 유사성을 비교할 때 왜 유클리드 거리(Euclidean distance)를 쓰지 않고, 코사인 거리를 쓰는 걸까? 왜 데이터가 적을 때는 경사 하강법(Gradient Descent)보다 정규 방정식(Normal Equation)을 쓰라고 할까? 왜 데이터가 커지면 일반적인 단일 시스템이 아닌 맵리듀스를 사용하는 하둡 같은 분산머신을 사용해야 할까?

학부생 시절 사례 기반 추론(case-based reasoning) 기법을 공부한 적이 있다. 영어 원서의 내용도 어려웠지만 따라 해보니 그냥 모듈만 쓰면 매우 쉽게 두 인스턴스 간의 유사도가 나왔다. 영어에다 설명을 어렵게 써놔서 그렇지 한국 사람 대부분이 고등학교 때 배우는 피타고라스 정리를 조금만 더 들여다보면 이해할 수 있는 유클리드 거리로, 두 인스턴스 값들의 유사도를 구하는 기법이었다. 처음부터 프로그래밍으로 바닥부터 작성해 보았다면 훨씬 더 빨리 이해했을 텐데...

이 책은 그런 책이다. 데이터와 관련해서 가장 넓은 개념이라고 볼 수 있는 '데이터 과학'에서 쓰이는 모든 기초적인 기법을 파이썬을 이용해서 가장 기초적인

것부터 구현한다. 정말 거의 모든 데이터 과학의 영역을 다룬다. 파이썬, 통계학, 머신러닝, 네트워크 분석, 데이터베이스, 데이터 엔지니어링 그리고 심지어 2판에서는 데이터에 대한 윤리까지 말이다. 데이터 과학에 입문하면서 모든 내용을 톺아볼 때 이것만큼 좋은 책이 있을까?

조엘의 서술 방식 역시 처음 데이터 과학을 배우는 독자들에게 큰 도움이 될 것이다. 조엘은 복잡한 개념을 유머와 해학으로 쉽게 풀어 설명하는 재능이 뛰어나다. 자연어처리(natural language processing, NLP) 영역에서 뛰어난 연구를 하는 학자면서 학회에서 많은 이에게 감명을 주는 워크숍을 진행하는 자애로운 교육자이기도 하다. 세 분의 번역자 역시 독자들에게 조금 더 좋은 내용을 전달해 주기 위해 고심했다. 현재 파파고 기계 번역팀의 리더이기도 한 박은정 님이 이 책의 초판을 번역할 때, 기존의 우리말 번역에서는 의미 구별이 쉽지 않은 단어들(예를 들어 regularization과 normalization은 모두 '정규화'라고 번역하는 게 맞는가?)을 함께 고민하면서 밥 먹으러 갔던 기억이 있다. 번역의 생산성보다는 용어에 대한 올바른 이해를 제공하기 위해 늘 고민하던 분들의 작품이다.

처음 데이터 과학을 배우는 사람이 내게 책을 한 권만 추천해달라고 한다면 이 책을 권하고 싶다. 이 책은 어려운 내용을 쉽게 설명하고 있으며, 우리가 어릴 적부터 배워왔던 한국의 수학교육이 머신러닝 분야에서 얼마나 쓸모 있는지 이해할 수 있게 해준다. 독자는 처음 봤을 때 다소 어렵게 느껴지는 데이터 과학의 수식들이 코드로 표현하면 얼마나 간단한지 이해할 수 있게 될 것이다. 이후에 공부를 이어갈 때도 좋은 초석이 될 만한 기초 코드들이 많이 포함되어 있다. 데이터 과학을 공부하려 한다면 이 책을 반드시 소장할 것을 권한다.

최성철, 가천대 산업경영공학과 교수

옮긴이 서문

1판이 나온 지 3년이 넘었습니다. 그간 데이터 과학 업계에는 새로운 기술이 많이 탄생했고, 그러한 기술을 토대로 세상에 실질적인 영향을 끼치기도 했습니다. 한편으로는 빠르게 발전하는 기술을 좇아가기가 쉽지 않다고도 느끼지만, 그럴 때일수록 기본을 다지는 것이 중요하다고 생각합니다. 2판은 1판에서 훑었던 데이터 과학과 기계학습의 중요한 알고리즘들에, 1판 이후 탄생한 기술 중 핵심적인 몇 가지(딥러닝 등)를 더해서 기초를 탄탄히 쌓을 수 있게 해줍니다.

데이터 과학은 기술적으로 발전했을 뿐 아니라, 더욱더 많은 학생 그리고 연구자와 대중으로부터 관심을 얻기 시작하면서 한국어의 용어 면에서도 많은 논의와 개선이 있었습니다. 예를 들어 'likelihood'의 경우 2016년 당시에는 '우도' 또는 '가능도'라고 번역되곤 했고 그 어느 쪽도 정착했다고 보기 어려운 상황이었지만, 지금은 많은 번역서가 '가능도'라는 용어를 채택하고 있고, 대중도 큰 어색함 없이 이 용어를 받아들이고 있는 것 같습니다. 한편으로는 아직 마땅히 자리를 잡지 못했거나 옮긴이들이 모호하다고 판단한 용어 몇 가지는 음차 했습니다. 예를 들어 'gradient'의 경우 'gradient descent'의 맥락에서 논의될 때는 '경사'로 번역했지만, 그 외에 홀로 사용된 경우에는 '그래디언트'라고 표기했습니다.

이 책의 1판을 통해 새로운 인연을 맺게 되었기 때문에 저희에게는 매우 의미가 있는 책입니다. 훌륭한 책의 번역을 다시 한번 맡게 되어 영광입니다. 한편으로 부족한 번역임에도 기꺼이 책을 추천해 주신 분들께 깊은 감사를 드립니다. 훌륭한 데이터 과학자들을 양성해서 세상에 내보내고 계신 최성철 교수님, 냉철한 시각으로 한국의 데이터 과학 업계를 이끌어 주고 계신 권정민 님, 그리고 혜안 넘치는 블로그 포스트와 각종 발표 자료로 많은 이에게 영감을 주는 데이터 과학계의 떠오르는 샛별 변성윤 님(*https://zzsza.github.io*), 감사합니다.

3년 뒤, 5년 뒤, 10년 뒤의 데이터 과학 업계는 또 어떻게 변해있을까요? 그때는 우리와 같은 옮긴이가 필요하지 않고 기계번역으로 번역서가 나오게 될까요? 혹시 좋은 기술서를 기계가 쓰고 있지는 않을까요? 다가오는 미래가 무척 기대됩니다. 이 번역서를 통해 그 미래를 더 많은 분과 함께 지켜보고, 만들어 갈 수 있게 되길 바랍니다.

김한결, 하성주, 박은정 드림

2판 서문

나는《밑바닥부터 시작하는 데이터 과학》초판이 아주 자랑스럽다. 내가 원하는 방향대로 책이 나와 주었기 때문이다. 하지만 수년간 데이터 과학이 발전하고, 파이썬 개발 환경이 진보하고, 개인적으로는 개발자와 교육자로서 성장하다 보니 이상적인 데이터 과학 책에 관한 나의 생각이 바뀌었다.

인생에 있어 재도전이란 없다. 하지만 글에는 2판이라는 것이 있다.

그렇기에 이번에는 모든 코드와 예시를 파이썬 3.6으로 수정하였으며 타입 어노테이션 등 새로운 기능을 활용하였다. 깔끔한 코드 작성에 대해 강조하는 내용을 같이 엮어 보았다. 초판의 간단한 예시들을 '진짜' 데이터셋을 사용하여 조금 더 현실적인 예시로 교체했다. 오늘날의 데이터 과학자들이 다루어야 할 딥러닝, 통계, 자연어 처리 등의 주제를 다루는 내용을 추가했다. (시의성이 떨어진다고 생각되는 내용도 제거했다.) 세심하게 책을 다시 살피며 버그를 고치고, 설명이 명쾌하도록 다시 쓰고, 몇몇 농담을 새롭게 바꿨다.

초판은 훌륭한 책이었지만, 이번 판은 그보다 더 낫다. 즐겁게 읽기를 바란다!

조엘 그루스(시애틀, 2019)

표기 관례

이 책에서는 다음의 표기 관례를 따른다.

이탤릭

URL, 메일 주소, 책 제목(원저작물)에 사용한다.

고정폭 글꼴

코드의 표현 및 문단 내의 프로그램 구성 요소인 변수, 함수명, 데이터 타입, 환경 변수, 문장(statement), 키워드 등을 표현한다.

볼드체

강조하는 용어에 사용한다.

이 책에 쓰인 표시

	팁 혹은 제안을 나타낸다.
	일반적인 주석을 뜻한다.
	경고 혹은 주의를 나타낸다.

예시 코드 사용하기

예시 코드, 문제 풀이 등 이 책에서 사용되는 자료는 *https://github.com/joelgrus/data-science-from-scratch*에서 내려받을 수 있다.

이 책은 업무에 실질적인 도움을 줄 목적으로 쓰였기 때문에, 예시 코드를 자신의 프로그램이나 문서에 재사용해도 문제가 없다. 가령, 예시 코드를 짜깁기해서 프로그램을 짜는 것은 내 허락을 받을 필요가 없다. 특정 질문에 관해 답변할 때 이 책을 인용하면서 예시 코드를 활용하는 것도 따로 허락을 받을 필요가 없다. 하지만 책에 수록된 예시들을 판매하거나 재배포하는 경우와 예시 코드의 상당량을 다른 문서에 활용할 때는 허락이 필요하다.

출처(attribution) 표기는 권고하지만, 강제하지는 않는다. 출처는 일반적으로 제목, 저자, 출판사 등을 포함하고, ISBN이 있는 경우 병기한다. 예를 들어 《밑바닥부터 시작하는 데이터 과학 2판》(조엘 그루스 지음, 김한결·하성주·박은정 옮김, 인사이트, 2020)과 같은 방식으로 표기하면 된다. 위에서 언급되지 않은 용도로 코드 예시를 사용하고 싶다면 *permissions@oreilly.com*으로 연락을 주기 바란다.

연락하는 법

이 책에 대한 지적이나 질문은 출판사로 연락하길 바란다.

- O'Reilly Media, Inc.
  1005 Gravenstein Highway North
  Sebastopol, CA 95472
- 800-998-9938(미국 혹은 캐나다)
- 707-829-0515(국제)
- 707-829-0104(팩스)

우리는 이 책을 위한 웹페이지에 정오표, 예시 및 기타 정보를 모아두었다. 해당 페이지는 *http://bit.ly/data-science-from-scratch-2e*(영문판) 또는 *https://github.com/insightbook/data-science-from-scratch/wiki/Errata-(2nd-Edition)*(한국어판)으로 접근할 수 있다.

책의 기술적인 부분에 관한 질문이나 논의는 *bookquestions@oreilly.com*으로 메일을 보내면 된다.

감사의 글

가장 먼저, 이 책을 쓰겠다는 제안을 받아 준 (그리고 적정량으로 내용을 줄이자고 조언해 준) Mike Loukides에게 감사드린다. '자꾸 나한테 원고 샘플을 보내서 귀찮게 하는 이 작자는 누구야? 어떻게 물리치지?'라고 충분히 생각할 수 있었음에도 그러지 않아서 감사하다. 출판 과정 내내 옆에서 지도해 주며 혼자서 했을 때보다 훨씬 좋은 책을 만들 수 있게 해준 편집자 Marie Beaugureau에게도 감사하다.

내가 데이터 과학을 배우지 않았다면 이 책을 쓰지 못했을 것이고, Dave Hsu, Igor Tatarinov, John Rauser 등 Farecast 친구들의 영향이 없었다면 데이터 과학을 배우지 않았을 것이다(너무 오래전이어서 당시에는 그것을 데이터 과학이라고 부르지도 않았지만). Coursera와 DataTau에 있는 분들에게도 많은 고마움을 전한다.

베타 리딩과 검토를 해준 분들께도 감사하다. Jay Fundling은 수많은 실수와 불명확한 설명을 짚어 주었고, 그 덕분에 훨씬 정확하고 좋은 책을 쓸 수 있었다. Debashis Ghosh는 내 통계 수식들에 틀린 곳이 없는지 검토해 주는 데 훌륭한 역할을 했다. Andrew Musselman은 내가 자꾸 "파이썬보다 R을 선호하는 사람은 도덕적으로 파렴치한 사람(people who prefer R to Python are moral reprobates)"이라고 주장하는 것을 조금 누그러뜨릴 수 있게 해줬는데, 결과적으로 그건 꽤 괜찮은 조언이었다. Trey Causey, Ryan Matthew Balfanz, Loris Mularoni, Núria Pujol, Rob Jefferson, Mary Pat Campbell, Zach Geary, Denise Mauldin, Jimmy O'Donnell 그리고 Wendy Grus 역시 아주 귀중한 피드백을 주었다. 초판을 읽어 주고 이 책이 더 나은 책이 될 수 있도록 도와준 모두에게 감사한다. 책에 오류가 남아 있다면 그것은 당연히 나의 책임이다.

내게 수많은 새로운 개념을 소개해 주고 멋진 사람들을 만나게 해주며, 스스

로의 자격지심을 만회하기 위해 이 책을 쓰게 해준 트위터의 #datascience 해시태그 사용자들에게도 감사한다. 선형대수에 관한 장을 책에 추가하도록 조언해 준 Trey Causey에게 (다시 한번) 감사드리고, 10장 '데이터 다루기'에서 누락된 중요한 부분들을 지적해 준 Sean J. Taylor에게도 감사드린다.

그 누구보다도 Ganga와 Madeline에게 감사드린다. 책을 쓰는 것보다 유일하게 더 힘든 것은 책을 쓰는 사람과 함께 사는 것인데, 이들의 도움이 아니었다면 결코 이 책을 쓰지 못했을 것이다.

초판 서문

데이터 과학

'21세기의 가장 섹시한 직업'은 데이터 과학자라고 한다.[1] (이 말을 한 사람은 소방서에 가보지 않았나 보다.[2]) 실제로 데이터 과학은 많은 인기를 누리며 성장하고 있고, 몇몇 애널리스트는 앞으로 10년 안에 수십억 명의 데이터 과학자가 더 필요할 것이라고 침 튀기며 주장하기도 한다.

데이터 과학(data science)이란 무엇인가? 이를 모른다면 데이터 과학자도 양성할 수 없으니 먼저 데이터 과학이 무엇인지 이해하는 것이 중요하다. 업계에서 제법 알려진 한 벤다이어그램[3]에 의하면 데이터 과학은 다음 세 가지 영역의 교집합이다.

- 해킹[4] 실력
- 수학 및 통계에 관한 지식
- 도메인 전문성

처음에는 이 책에서 세 가지를 모두 다루려고 했으나, 도메인 전문성을 상세히 설명하기 위해서는 수만 페이지가 필요하다는 것을 깨달았다. 그래서 앞의 두 가지에 집중하기로 했다. 이 책의 목표는 데이터 과학을 시작하는 데 필요한 해킹 실력을 키워 주고, 데이터 과학의 핵심인 수학 및 통계학에 익숙해질 수 있게 도와 주는 것이다.

어쩌면 이마저도 책 한 권에 담기에는 너무 많을 수 있다. 사실 해킹 실력을 키우기 위한 가장 좋은 방법은 직접 이것저것 해킹해 보는 것이다. 이 책을 읽고 나면 필자가 어떤 방식으로 해킹하는지 알게 되겠지만, 막상 그것이 본인에게는 가장 적합한 방법이 아닐 수 있다. 내가 어떤 도구들을 주로 쓰는지도 알게 되겠지만, 마찬가지로 본인에게 가장 좋은 방법은 아닐 수 있다. 또한 데이터 관련

1 *https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century*
2 (옮긴이) 미국에는 소방관이 가장 섹시한 직업이라는 농담이 있다.
3 *http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram*
4 (옮긴이) 여기서 말하는 해킹은 악의적 의도를 담은 크래킹이 아니라, 일반적인 의미의 프로그래밍이라고 할 수 있다.

문제에 어떻게 접근하는지는 알게 되겠지만, 그것이 본인에게 가장 좋은 접근법은 아닐 수 있다. 나의 의도는 (그리고 바람은) 여기서 보여 주는 예시들이 영감이 돼서 자기만의 방법을 찾는 것이다. 이 책의 모든 코드와 데이터는 깃허브(GitHub)에서 볼 수 있다.[5]

비슷한 맥락으로 수학을 배우는 가장 좋은 방법은 수학을 직접 해보는 것이다. 이 책은 결코 수학책이 아니며 '진짜 수학'을 다루지 않을 것이다. 하지만 확률과 통계, 선형대수에 대한 최소한의 이해 없이 데이터 과학을 할 수는 없다. 따라서 우리는 필요할 때마다 수식, 수학적 직관, 수리적 공리 그리고 수학적 아이디어를 최대한 간단하게 이해하려 할 것이다. 이런 것들을 함께 풀어나가는 과정을 겁내지 않았으면 한다.

또한 이 책을 읽고 나서 데이터를 가지고 노는 것이 재밌다는 느낌을 받았으면 좋겠다. 왜냐하면... 데이터를 가지고 노는 것은 재밌으니까! (적어도 연말정산을 하거나 석탄을 캐는 것보다는 훨씬 재밌다.)

왜 '밑바닥'부터인가

잘 알려진 (그리고 덜 알려진) 알고리즘이나 테크닉을 효율적으로 잘 구현한 데이터 과학 관련 라이브러리, 프레임워크, 모듈 그리고 툴킷은 아주아주 많다. 데이터 과학자가 되면 NumPy, scikit-learn, pandas 그리고 그 외에도 아주 많은 다른 라이브러리에 익숙해진다. 이들은 데이터 과학을 하는 데 아주 유용한 도구다. 하지만 한편으로는 데이터 과학의 원리를 제대로 이해하지 못한 상태로 데이터 과학을 할 수 있는 방법이기도 하다.

따라서 이 책에서 우리는 데이터 과학을 '밑바닥'부터 시작해 볼 것이다. 즉, 도구와 알고리즘을 더 명확하게 이해하기 위해서 라이브러리 등을 이용하지 않고 직접 만들어 볼 것이다. 나는 이를 위해 주석을 열심히 달고, 읽기 쉬운 명확한 구현체와 예시를 만드는 데 많은 노력을 들였다. 하지만 우리가 작성할 코드는 이해하기는 쉬워도 효율적이지는 않을 수 있다. 즉, 작은 데이터에는 잘 작동하겠지만 '웹 스케일(web scale)'의 대용량 데이터에는 적합하지 않을 것이다.

대용량의 데이터에 어떤 라이브러리를 쓰면 적합한 알고리즘을 적용할 수 있는지 알려주겠지만, 직접 사용하지는 않을 것이다.

요즘 데이터 과학을 배우기 위한 가장 적합한 언어가 무엇인지에 대한 생산적

5 *https://github.com/joelgrus/data-science-from-scratch*

인 토론이 많다. 어떤 사람들은 그 언어가 R이라고 한다(물론 나는 그 사람들이 틀렸다고 생각한다). 몇몇은 자바 또는 스칼라도 괜찮다고 한다. 그러나 내 생각에는, 당연히 파이썬이 정답이다.

파이썬은 데이터 과학을 배우기 위해 (그리고 수행하기 위해) 적합한 몇 가지 특성이 있다.

- 공짜다.
- 코드를 간편하게 작성할 수 있다(게다가 읽기 쉽다).
- 데이터 과학과 관련해 유용한 라이브러리가 많다.

그렇다고 내가 가장 좋아하는 프로그래밍 언어가 파이썬이라고 말하기는 어렵다. 개인적으로 더 만족스럽거나, 잘 설계되었다고 생각하거나, 또는 단순히 코딩하는 게 더 즐거운 언어들이 있다. 그럼에도 새로운 데이터 과학 프로젝트를 시작할 때마다 나는 매번 파이썬을 쓴다. 실제로 작동하는 프로토타입을 빠르게 만들어야 할 때나, 데이터 과학의 개념을 명확하고 이해하기 쉬운 방식으로 전달하고 싶을 때도 파이썬을 쓴다. 따라서 이 책에서도 파이썬을 사용하게 되었다.

이 책의 목표는 파이썬을 가르치는 것이 아니다(물론 이 책을 읽음으로써 파이썬을 어느 정도 배우기는 할 것이다). 한 개 장 분량의 파이썬 속성 강좌를 통해 기초적인 파이썬 사용법을 다루기는 하겠지만, 파이썬 프로그래밍을 전혀 할 줄 모른다면 (또는 프로그래밍을 전혀 할 줄 모른다면) 초보자를 위한 파이썬 책을 함께 볼 것을 추천한다.

데이터 과학에 대한 설명 역시 비슷한 방식을 취할 것이다. 필요하거나 도움이 될 만한 부분은 깊이 들어가겠지만, 그렇지 않은 경우에는 스스로 내용을 터득해야 (또는 위키피디아를 찾아보도록) 한다.

지난 몇 년간, 나는 여러 명의 데이터 과학자를 지도해왔다. 그 모두가 세상을 변화시키는 슈퍼스타가 되지는 않았지만, 처음 만났을 때보다 분명히 더 나은 데이터 과학자가 되었다고 자부한다. 그리고 그 경험을 통해, 어느 정도의 수학적 능력과 프로그래밍 실력만 있다면 누구나 데이터 과학자가 될 자질을 갖고 있다고 믿게 되었다. 그 외에는 오로지 호기심과 열심히 하겠다는 의지 그리고 이 책만 있으면 된다. 이 책을 선택하게 된 것을 축하한다.

1장

Data Science from Scratch

들어가기

"데이터! 데이터! 데이터!" 그가 조바심내며 말했다.
"진흙 없이 벽돌을 만들 수는 없잖아."
- 아서 코난 도일(Arthur Conan Doyle)

1.1 데이터 시대의 도래

우리는 데이터의 홍수 속에 살고 있다. 웹사이트는 모든 사용자의 클릭 하나하나를 추적하며, 스마트폰은 우리의 위치와 움직이는 속도를 매일 시시각각 기록하고 있다. 자신의 생활을 데이터로 남기려는 사람들은 첨단 만보계를 차고 다니며 심장 박동수, 움직임, 식습관 그리고 수면 습관까지 기록한다. 스마트카(smart car)는 운전 습관을, 스마트홈(smart home)은 생활 습관을, 스마트 마케터(smart marketer)는 고객의 구매 습관 데이터를 수집한다. 인터넷 또한 방대한 양의 지식이 서로 참조되어 있는 거대한 백과사전으로 자리잡았다. 인터넷을 통해 영화, 음악, 스포츠 경기 결과, 핀볼 게임, 밈(memes), 칵테일 등 세부적인 분야의 데이터까지 얻을 수 있으며 지나치게 많은 대상으로부터 나오는, 지나치게 많은 데이터는 우리가 분석해 주기만을 기다리고 있다. (심지어 그들 중 대부분은 거짓 데이터도 아니다!)

이렇게 방대한 데이터 속에는 아무도 생각해 내지 못했던 수많은 질문에 관한 답이 숨어 있다. 이 책에서는 그 답을 찾는 방법들을 알아볼 것이다.

1.2 데이터 과학이란?

데이터 과학자란 컴퓨터 과학자보다는 통계학을 더 잘 알고 통계학자보다는 컴

퓨터 과학을 더 잘 아는 사람이라는 농담이 있다(타당한 농담이라는 뜻은 아니다). 실제로, 여러 이유로 통계학자인 데이터 과학자들도 있으며 소프트웨어 엔지니어와 크게 다른 일을 하지 않는 데이터 과학자들도 있다. 또한 기계학습 전문가인 사람도 있으며 기계학습의 '기'자도 모르는 사람이 있다. 몇몇은 대단한 논문 실적을 가지고 있는 박사지만, 논문 한 편 읽어본 적도 없는 사람들 또한 있다(물론 그런 사람들은 스스로 좀 부끄러워해야 한다고 생각한다). 간단히 말해, 데이터 과학자를 어떻게 정의하든 그 정의에 결코 부합되지 않는 실무자들은 항상 존재한다.

그러한 어려움에도 불구하고 데이터 과학자에 관한 정의를 내려보자. 일단, 지저분한 데이터에서 통찰(insight), 즉 유용한 규칙을 발견하려는 사람이라고 해볼까? 실제로 세상에는 데이터를 통찰력으로 바꾸려는 사람들이 한가득이다.

데이팅 서비스 OkCupid는 회원들에게 수천 개의 질문에 관한 답변을 받아 그들에게 가장 잘 맞는 짝을 찾아준다. 게다가 그 결과를 분석해서 상대방이 첫 데이트에서 나와 잠자리를 가질 가능성이 얼마나 되는지[1] 알아볼 수 있는 질문들을 찾아냈다.

페이스북은 표면적으로는 사용자의 친구들을 쉽게 찾고 연결해 주기 위해 사용자의 고향과 현재 위치 정보를 요청한다. 하지만 그뿐 아니라 이 데이터를 분석해서 국제 이주 패턴을 찾거나[2] 각 미식축구팀 팬들의 주요 거주 지역을 찾기[3]도 한다.

Target[4]은 온라인과 오프라인 구매 내역 데이터로 예측 모델을 만들어 어떤 고객이 임신했는지 추정하고[5], 해당 고객들을 대상으로 아기 용품을 광고한다.

2012년 오바마 대선 캠프는 수십 명의 데이터 과학자를 고용했다. 이들은 데이터마이닝과 실험을 통해 더욱 집중적인 관심을 요하는 유권자들을 걸러내고, 기부자들이 가장 좋아할 만한 최적의 선거 모금 프로그램과 분위기를 선택하는 데 도움을 주었다. 그리고 선거 참여율이 높아질 경우 당선에 도움이 되는 지역을 찾아냈다. 2016년 트럼프 대선 캠프 또한 다양한 종류의 온라인 광고에서 발생하는 데이터를 분석하여[6] 반응이 좋은 캠페인 광고와 그렇지 못한 광고를 구

1 *https://www.gwern.net/docs/psychology/okcupid/thebestquestionsforafirstdate.html*
2 *http://on.fb.me/1EQTq3A*
3 *http://on.fb.me/1EQTvnO*
4 (옮긴이) 미국의 커다란 슈퍼마켓 체인점.
5 *http://nyti.ms/1EQTznL*
6 *https://www.wired.com/2016/11/facebook-won-trump-election-not-just-fake-news/*

별했다.

이러한 데이터 과학의 용도에 실망할 수도 있지만, 몇몇 데이터 과학자는 정부를 더 효율적으로 돌아가게 한다거나[7], 집이 없는 사람들을 돕거나[8], 공중 보건을 개선하는 등, 좋은 의도의 행동을 위해 자신의 능력을 이용하기도 한다. 하지만 사용자의 광고 클릭 수를 늘릴 수 있는 최적의 방법을 분석해 보는 게 개인의 경력을 해치진 않을 것이다.

1.3 동기부여를 위한 상상: 데이텀 주식회사

축하한다! 여러분은 데이터 과학자들의 최대 소셜 네트워크, '데이텀(Datum) 주식회사'의 책임 데이터 과학자로 고용되었다.

이 책의 초판이 나왔을 때, 데이터 과학자들을 위한 소셜 네트워크는 흥미롭지만 비현실적인 예시라고 생각했었다. 하지만 실제로 이러한 사이트들이 생겨나고 있고, 이 책의 원고료보다 훨씬 많은 금액을 벤처 캐피털에서 투자받고 있다. 터무니 없는 데이터 과학 예시에서 조차 얻어갈 것은 항상 있는가 보다.

데이텀은 데이터 과학자들을 위한 소셜 네트워크임에도 불구하고, 아직 내부 시스템에 데이터 과학을 도입해 보지는 않았다. (회사의 입장에서 변명을 하자면 데이텀은 아직 제품을 만드는 데도 투자해 본 적이 없다.) 그것을 바꾸는 것이 여러분의 역할이다! 이 책에서 우리는 업무 중 맞닥뜨리는 몇 가지 문제를 해결하면서 데이터 과학에 관한 개념들을 하나씩 익혀나갈 것이다. 사용자들이 직접 제공한 데이터를 들여다보거나, 그들이 사이트 안에서 돌아다니면서 생성한 데이터를 다뤄 보고, 때로는 우리가 직접 설계한 실험에서 나온 데이터를 사용하기도 할 것이다.

데이텀에는 '우리가 만든 게 아니면 안돼(not-invented-here)'[9] 정신이 만연하기 때문에 모든 코드를 밑바닥부터 직접 작성할 것이다. 결국 책이 끝나갈 때쯤에는 데이터 과학을 제법 탄탄하게 이해하고 있을 것이며, 회사 안팎에서 더욱 안정적으로 실력을 발휘할 수 있게 될 것이다.

7 *http://bit.ly/1EQTGiW*

8 *http://www.dssgfellowship.org//2014/08/20/tracking-the-paths-of-homelessness/*

9 (옮긴이) 'not-invented-here(NIH)'는 다른 개인 혹은 집단의 저작물을 사용하지 않는 폐쇄적 행위를 일컫는 부정적인 말이다. 프로그래머들은 이러한 행동을 '바퀴를 재발명하기(reinventing the wheel)'라고 표현하기도 한다.

하여튼, 입사를 진심으로 환영한다. 그리고 행운을 빈다! (금요일에는 청바지를 입는 것이 허용되고, 화장실은 복도 끝에서 오른쪽으로 돌면 있다.)

1.3.1 핵심 인물 찾기

오늘은 첫 출근일이다. 네트워크 사업부 부사장은 사용자들에 관해 매우 궁금해한다. 그동안 궁금한 점을 물어 볼 사람이 없어서 아쉬워하고 있었는데, 이제 물어 볼 데가 생겨 아주 신이 나 있다.

부사장이 특히 궁금해 하는 것은, 데이터 과학의 핵심 인물이 누구인지 알아내는 것이다. 그것을 알아내기 위해 그는 여러분에게 데이텀 네트워크 데이터 전체를 넘겨 주었다. (실제 상황에서는 필요한 데이터를 바로 넘겨 주는 경우는 흔치 않다. 9장 '파이썬으로 데이터 수집하기'에서는 실생활에서 데이터를 수집하는 방법을 알아본다.)

어떻게 생긴 데이터인지 살펴보자. 먼저 딕셔너리 형태로 구성된 사용자 명단이 있다. 각 사용자는 숫자로 된 고유 번호인 `id`와 이름을 나타내는 `name`으로 구성되어 있다(엄청난 우연의 일치로 `id`와 `name`의 영어 발음의 운율이 딱 맞아떨어진다).

```
users = [
    { "id": 0, "name": "Hero" },
    { "id": 1, "name": "Dunn" },
    { "id": 2, "name": "Sue" },
    { "id": 3, "name": "Chi" },
    { "id": 4, "name": "Thor" },
    { "id": 5, "name": "Clive" },
    { "id": 6, "name": "Hicks" },
    { "id": 7, "name": "Devin" },
    { "id": 8, "name": "Kate" },
    { "id": 9, "name": "Klein" }
]
```

그리고 `id`의 쌍으로 구성된 친구 관계 데이터인 friendship_pairs도 있다.

```
friendship_pairs = [(0, 1), (0, 2), (1, 2), (1, 3), (2, 3), (3, 4),
                    (4, 5), (5, 6), (5, 7), (6, 8), (7, 8), (8, 9)]
```

예를 들어 `(0, 1)`은 `id`가 0인 데이터 과학자 Hero와 `id`가 1인 데이터 과학자 Dunn이 서로 친구라는 것을 의미한다. 그림 1-1은 이 데이터가 나타내는 네트워크를 보여 주고 있다.

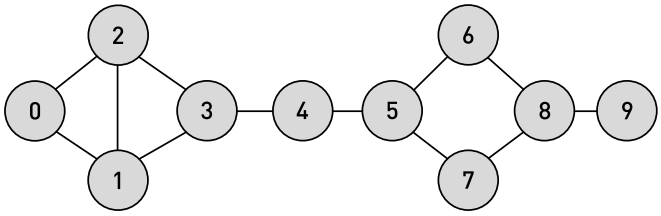


그림 1-1 데이텀의 사용자 네트워크

친구 관계를 쌍의 리스트로 표현하는 것이 이것을 다루는 가장 쉬운 방법은 아니다. 가령, id가 1인 사용자의 모든 친구를 찾으려면 모든 쌍을 순회하여 1이 포함되어 있는 쌍을 구해야 한다. 만약 엄청나게 많은 쌍이 주어졌다면 특정 사용자의 모든 친구를 찾기 위해 굉장히 오랜 시간이 걸릴 것이다.

대신, 사용자 id를 키(key)로 사용하고 해당 사용자의 모든 친구 목록을 값(value)으로 구성한 딕셔너리를 생성해 보자. (딕셔너리를 통한 데이터 탐색은 매우 빠르다).

☑ 2장에 '파이썬 속성 강좌'가 준비되어 있으니 당장은 코드에 너무 집착하지 말자. 지금은 우리가 하려는 것이 무엇인지 파악하는 데 집중하길 바란다.

사용자의 친구 목록을 딕셔너리로 생성하기 위해서는 여전히 모든 쌍을 탐색해야 한다. 하지만 처음에 단 한 번만 탐색하면 이후에는 훨씬 빠르게 각 사용자별 친구 목록을 찾아볼 수 있다.

```
# 사용자별로 비어 있는 친구 목록 리스트를 지정하여 딕셔너리를 초기화
friendships = {user["id"]: [] for user in users}

# friendship_pairs 내 쌍을 차례대로 살펴보면서 딕셔너리 안에 추가
for i, j in friendship_pairs:
    friendships[i].append(j)  # j를 사용자 i의 친구로 추가
    friendships[j].append(i)  # i를 사용자 j의 친구로 추가
```

이렇게 각 사용자의 친구 목록을 딕셔너리로 만들면 '네트워크상에서 각 사용자의 평균 연결 수는 몇 개인가?'와 같이 네트워크의 특성에 관한 질문에 답할 수 있다.

이 질문에 답하기 위해 먼저 friendships 안 모든 리스트의 길이를 더해서 총 연결 수를 구해 보자.

```
def number_of_friends(user):
    """user의 친구는 몇 명일까?"""
    user_id = user["id"]
```

```
    friend_ids = friendships[user_id]
    return len(friend_ids)

total_connections = sum(number_of_friends(user)
                        for user in users)        # 24
```

이제 단순히 이 합을 사용자의 수로 나누면 된다.

```
num_users = len(users)                            # 총 사용자 리스트의 길이
avg_connections = total_connections / num_users   # 24 / 10 == 2.4
```

다음으로 연결 수가 가장 많은 사람, 즉 친구가 가장 많은 사람이 누군지 알아보자.

사용자의 수가 많지 않으므로 '친구가 제일 많은 사람'부터 '제일 적은 사람' 순으로 사용자를 정렬해 보자.

```
# (user_id, number_of_friends)로 구성된 리스트 생성
num_friends_by_id = [(user["id"], number_of_friends(user))
                     for user in users]

num_friends_by_id.sort(                           # 정렬해 보자.
       key=lambda id_and_friends: id_and_friends[1],   # num_friends 기준으로
       reverse=True)                              # 제일 큰 숫자부터 제일 작은 숫자순으로

# (user_id, num_friends) 쌍으로 구성되어 있다.
# [(1, 3), (2, 3), (3, 3), (5, 3), (8, 3),
#  (0, 2), (4, 2), (6, 2), (7, 2), (9, 1)]
```

우리는 방금 네트워크에서 중심 역할을 하는 사람들을 찾아냈다. 실제로 우리가 방금 계산한 것을 네트워크 용어로는 **연결 중심성**(degree centrality)이라고 부른다(그림 1-2).

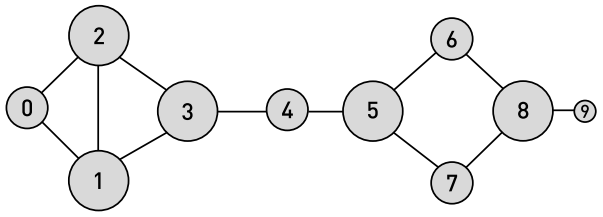


그림 1-2 연결 차수(degree)의 크기가 반영된 데이텀 네트워크

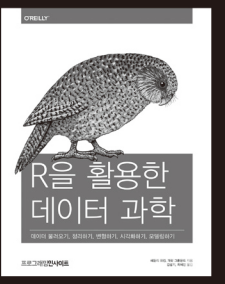
이 지수는 계산하기 쉽다는 장점이 있지만, 항상 기대하는 결과를 가져다 주지는 않는다. 예를 들어 데이텀 네트워크를 살펴보면 Dunn(id: 1)은 세 개의 연

파이썬을 활용한 데이터 길들이기

데이터 전처리 효율화 전략

캐서린 자멀, 재클린 카질 지음
이정윤, 이제원, 임훤 옮김 | 536쪽

데이터 수집, 정제, 가공과 같은 반복적인 전처리 과정을 파이썬 프로그래밍을 통해 효과적으로 작업하는 방법이 담겨 있다. 매번 되풀이되는 데이터 분석 초기 단계를 좀 더 효율적으로 작업하고 싶은 독자라면 데이터 분석 능력을 한 단계 업그레이드할 수 있을 것이다.

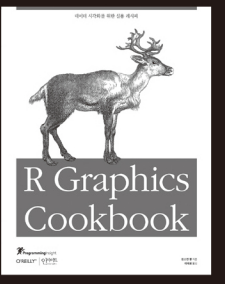


R을 활용한 데이터 과학

데이터 불러오기, 정리하기, 변형하기, 시각화하기, 모델링하기

해들리 위컴, 개럿 그롤문드 지음
김설기, 최혜민 옮김 | 495쪽

R을 활용하여 원 데이터로부터 지식과 통찰을 끌어내면서 데이터 과학의 분석 기법을 알려 주는 길잡이 책이다. R, RStudio와 R 패키지 모음인 tidyverse를 중심으로, 데이터 분석을 빠르고 능숙하고 수행할 수 있도록 이끌어 준다.

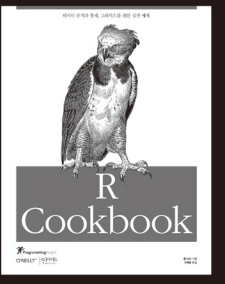


R Graphics Cookbook

데이터 시각화를 위한 실용 레시피

윈스턴 챙 지음 | 이제원 옮김 | 400쪽

R 기반 환경에서 그래프를 그려야 하는 독자를 위해 150개가 넘는 예제를 소개한다. 단순히 그래프 그리는 방법을 넘어, 그래프의 각 요소를 필요에 따라 삭제하거나 수정하는 방법도 알려 준다.

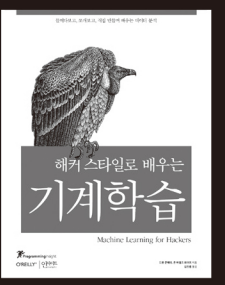


R Cookbook

데이터 분석과 통계, 그래픽스를 위한 실전 예제

폴 티터 지음 | 이제원 옮김 | 유충현 감수 | 520쪽

이 책은 R 사용자가 마주치는 여러 문제와 해결 방법을 200개가 넘는 예제를 통해 알려 준다. R을 처음 접하는 독자에겐 기초부터 한 단계씩 배우는 안내서가 되며, R을 잘 아는 이에게는 항상 곁에 두어야 할 참고 자료가 될 것이다.



해커 스타일로 배우는 기계학습

들여다보고, 쪼개보고, 직접 만들며 배우는 데이터 분석

드류 콘웨이, 존 마일즈 화이트 지음
김진홍 옮김 | 328쪽

이 책에는 기계학습의 12개 기법에 대한 훌륭한 사례가 담겨 있다. 이론을 설명하는 대신 과정에 초점을 맞췄기 때문에, 프로그래밍을 조금 할 줄 알고 정량적으로 사고할 줄 아는 사람이면 누구나 이해하기 쉽다.

Data Science from Scratch 2/E
First Principles with Python

데이터 과학의 기본 개념을 잡는 입문서!

이 책은 라이브러리나 프레임워크와 같은 도구를 사용하지 않고 '밑바닥부터' 만들어 보며 데이터 과학과 관련된 알고리즘을 알려 주는 기본서다. 데이터 과학을 배우기 위해 꼭 필요한 기본적인 내용을 중심으로, 책 전반에 걸쳐 수학과 통계학 기초가 녹아든 데이터 과학의 주요 기술을 다룬다.
이번 2판에서는 모든 코드와 예시가 파이썬 3.6으로 수정됐고 타입 어노테이션 등 새로운 기능을 활용한다. 오늘날의 데이터 과학자들이 다루어야 할 딥러닝, 통계, 자연어 처리 등의 주제도 추가됐다. 데이터 분석에 필요한 해킹과 수학·통계학 이론을 살펴보려는 프로그래머라면 기초부터 잡아 주는 강력한 도구로써 이 책의 일독을 권한다.

데이터 과학을 처음 배울 때 어떤 책을 보면 좋은지 사람들이 물어보면 늘 《밑바닥부터 시작하는 데이터 과학》을 추천한다. "사실 데이터 과학은 어렵지 않다"라고 얘기해 주는 책이기 때문이다. 독자는 처음 볼 때는 어렵게 느껴지는 데이터 과학의 수식들이 코드로 표현하면 얼마나 간단한지 알게 될 것이다. 또 저·역자가 데이터 과학을 얼마나 깊이 이해하고 있는지도 느낄 수 있다. 파이썬을 막 끝내고 처음으로 데이터 과학을 공부하려 한다면 반드시 소장할 것을 권한다.

최성철, 가천대 산업경영공학과 교수

수많은 데이터 과학 관련 책 사이에서 헤매고 있는, 데이터 과학에 발을 들이고자 하는 사람이라면 일단 이 책을 집어 들어도 좋을 것이다. 데이터에 대한 해박한 지식과 실무 감각을 모두 갖춘 저자는 데이터 과학 업무에서 실제로 꼭 필요한 지식을 잘 간추렸다 파이썬 코드로 많은 내용을 효과적으로 설명하여, 데이터 분석의 기본 개념을 다지면서 흥미를 키우도록 도와준다.

권정민, ODK Media, 데이터 과학자

데이터 과학 분야를 처음 접하는 분들이 보면 매우 좋은 책이다. 파이썬, 데이터 시각화, 선형대수, 통계, 데이터 전처리, 기계학습, 데이터베이스, 맵리듀스, 데이터 윤리 등 광범위한 주제를 명확히 설명하여, 데이터 과학에 대한 큰 그림을 그릴 수 있다. 특히 각 장 끝에 있는 "더 공부해 보고 싶다면"을 꼭 읽어보자. 사이킷런이나 텐서플로를 접하기 전에 먼저 이 책을 읽는 것을 추천한다.

변성윤, 쏘카 데이터 그룹, 머신러닝 엔지니어 & 데이터 과학자

이 책의 대상 독자

- 데이터 과학에 필요한 기초와 프로그래밍 두 마리 토끼를 모두 잡고 싶은 전공자
- 수학·통계학 이론이 녹아든 데이터 과학 관련 알고리즘을 살펴보려는 데이터 과학자

이 책에서 다루는 내용

- 선형대수, 통계, 확률에 관한 기초와 데이터 과학에서 활용하는 법 배우기
- 데이터를 수집, 탐색, 정제, 가공, 조작하기
- 기계학습의 원리 탐색하기
- k-근접 이웃, 나이브 베이즈, 선형 회귀 분석, 로지스틱 회귀 분석, 의사결정나무, 인공신경망, 군집화 등 구현하기
- 추천 시스템, 자연어 처리, 네트워크 분석, 맵리듀스, 데이터베이스 등 살펴보기
- 속성으로 파이썬 배우기

블로그 http://blog.insightbook.co.kr
트위터 @insightbook
페이스북 https://www.facebook.com/insightbook

이 책에 수록된 예제는 https://github.com/insightbook/data-science-from-scratch/에서 내려받을 수 있습니다.

값 33,000원
93000
9 788966 262571
ISBN 978-89-6626-257-1

인사이트 O'REILLY®

《밑바닥부터 시작하는 데이터 과학 2판》 미리보기