به‌نام خداوند ن وَ اَلْقَلَم

علم داده کاربردی

تحلیل داده‌های واقعی با اسکریپت‌نویسی بَـش

تألیف:

ابوالفضل خورشیدی

نیاز دانش

سرشناسه	: خورشیدی، ابوالفضل، ۱۳۶۸ -
عنوان و نام پدیدآور	: علم داده کاربردی تحلیل داده‌های واقعی با اسکریپت‌نویسی بش / تالیف ابوالفضل خورشیدی.
مشخصات نشر	:تهران: نیاز دانش، ۱۳۹۷.
مشخصات ظاهری	:۳۳۶ص.: جدول، نمودار.
شابک	:978-600-8906-37-7
وضعیت فهرست‌نویسی	:فیپا
یادداشت	:کتابنامه.
موضوع	: بش (زبان برنامه‌نویسی کامپیوتر)
موضوع	:Bash (Computer program language)
موضوع	:زبان‌های نوشتاری (کامپیوتر) -- آزمون‌ها و تمرین‌ها
موضوع	:Scripting languages (Computer science) -- Examinations, questions, etc.
رده‌بندی کنگره	: QA۷۶/۷۳ /۱۳۹۷ ۹خ۵ب
رده‌بندی دیویی	: ۰۰۵/۱۳۳
شماره کتابشناسی ملی	: ۵۴۸۱۵۵۵

نیاز دانش

نام کتاب	:	علم داده کاربردی-تحلیل داده‌های واقعی با اسکریپت نویسی بَش
نویسنده	:	ابوالفضل خورشیدی
طراح جلد،حروف‌چینی و صفحه آرایی	:	فاطمه سادات کسائیان زاده مهابادی
مدیر اجرایی – ناظر بر چاپ	:	حمیدرضا ا محمد شیرازی – محمد شمس
ناشر	:	نیاز دانش
نوبت چاپ	:	اول – ۱۳۹۷
شمارگان	:	۱۰۰
قیمت	:	۳۶۰۰۰۰ ریال

شابک: ۷-۳۷-۸۹۰۶-۶۰۰- ۹۷۸ **ISBN:978-600-8906- 37-7**

آدرس انتشارات : تهران – میدان انقلاب – خیابان ۱۲ فروردین – تقاطع وحید نظری – پلاک ۲۵۵- طبقه ۱- واحد ۲

تماس با انتشارات: ۰۹۱۲۷۰۷۳۹۳۵-۶۶۴۷۸۱۰۸-۶۶۴۷۸۱۰۶-۰۲۱

www.Niaze-Danesh.com

مشاوره جهت نشر: ۲۱۰۶۷۰۹ – ۰۹۱۲

فهرست مطالب

پیش‌گفتار

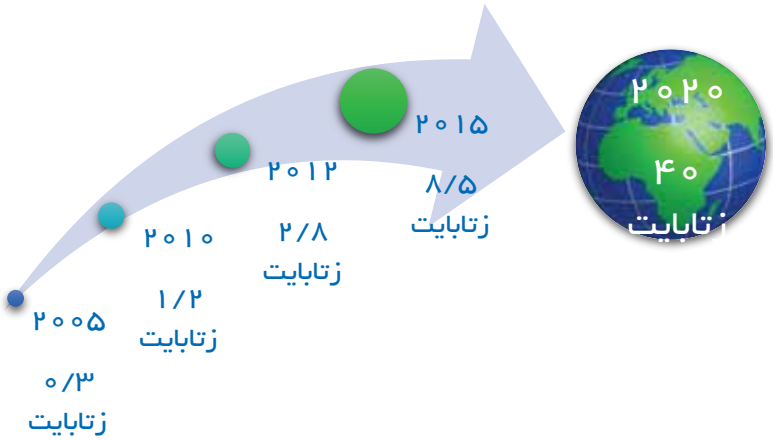
قبل از هر چیز لازم است تا از شما خواننده‌ی گرامی تشکر کنم و جای بسی خرسندی است که تصمیم گرفته‌اید تا بخشی از وقت ارزشمندتان را برای مطالعه‌ی این کتاب سرمایه‌گذاری نمایید. در این کتاب گام به گام همراه شما خواهیم بود تا با رویکردی عملگرایانه به ماجراجویی در دنیای علم داده بپردازیم و بتوانیم بینش‌های ناشناخته را کشف و آینده را پیش‌بینی کنیم. برای آن‌که علم داده برای تمامی افراد با هر سطح دانش و تجربه‌ی کاری فراگیر شود، لازم است تا در وهله‌ی نخست، داده‌ها را ساده، در دسترس و ملموس کنیم. برای رسیدن به این مهم، در این کتاب اتفاقات و مسائلی را که در کار و زندگی روزمره با آن‌ها سر و کار داریم از دریچه‌ای متفاوت مورد بررسی قرار داده‌ایم و تلاش کرده‌ایم با کمک ابزارهای ساده و کوچک، امّا قدرتمند، نگاهی نوین، علمی و مبتنی‌بر کدنویسی به چالش‌های زندگی واقعی داشته باشیم.

واقعیت این است که بینش‌های ناشناخته به صورت داده‌های خام و بعضاً حجیم و کلان ذخیره می‌شوند. پیش‌بینی آینده‌ی یک کسب‌وکار نیازمند به کارگیری این داده‌ها می‌باشد. به بیان دیگر، داده‌ها نقش بسیار مهمی درتصمیم‌سازی، مدیریت بهتر منابع و سودآوری برای یک شرکت یا سازمان داشته و هم‌چنین موجب ارائه‌ی خدمات بهتر به مشتریان شده و رضایت و به تبع آن وفاداری مشتری را به همراه خواهند داشت. مطالعات مؤسسه IDC[1] نشان می‌دهد که تا سال ۲۰۲۰ حدود ۴۰ زتابایت (۴۰ هزار میلیارد گیگابایت)[2] داده در سراسر دنیا تولید می‌شود که ۳۰۰ برابر بیشتر از حجم داده‌ی دنیا در سال ۲۰۰۵ است [۱]. در این مطالعه اشاره شده است که بیش‌تر این داده‌ها نه توسط انسان بلکه توسط ماشین‌ها تولید خواهد شد. گوشی‌های تلفن همراه، شبکه‌های

[1] International Data Corporation

[2] $40\ ZB = 40 \times 10^{21}\ bytes$

حسگری، دوربین‌ها، دستگاه‌های کنترل تردد، وسائل حمل و نقل، وب سایت‌ها و فروشگاه‌های اینترنتی، لاگ‌های نرم‌افزاری و ... ماشین‌هایی هستند که در افزایش حجم داده نقش بسزایی دارند.

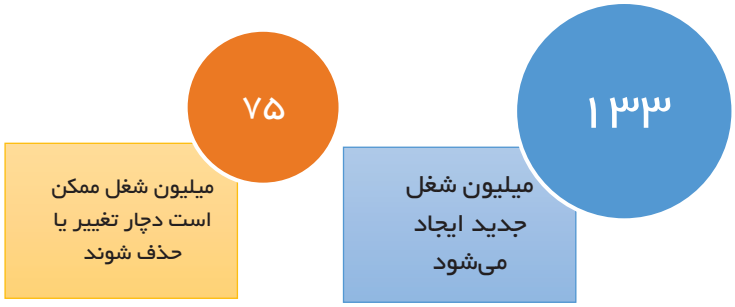


تعداد **سلول**‌های بدن انسان حدود **۴۰** هزار میلیارد تخمین زده می‌شود.
اگر هر **بیت** را معادل یک سلول بگیریم، **۴۰** زتابایت معادل مجموع سلول‌های کل **جمعیت بشر** بر روی کره زمین می‌باشد؛ یعنی کل سلول‌های بدن ۸ میلیارد انسان.

شرکت DOMO با تمرکز بروی داده‌های تولید شده در سراسر دنیا گزارشی منتشر کرده است که نشـــان می‌دهد این تولیدکنندگان جهانی داده در حال تولید ۲/۵ اگزابایت (۲/۵ میلیارد گیگابایت) داده جدید در هر روز هستند. پیش‌بینی می‌شود این حجم تولید داده با رشد اینترنت اشیاء (IoT) شتاب فزاینده‌ای به خود بگیرد [۲].

۲٫۵ اگزابایت (میلیارد گیگابایت) معادل داده‌ی پر شده بر روی **۱۰** میلیون **دیسک بلوری** است.
اگر این تعداد دیسک را روی یکدیگر قرار دهیم ارتفاع آن **۴** برابر ارتفاع **برج ایفل** می‌شود.

بر اساس پیش‌بینی‌های مجمع جهانی اقتصاد (WEF)[۱] تا سال ۲۰۲۲، حدود ۱۳۳ میلیون جایگاه شغلی وجود خواهد داشت که از این میان، صنایع مختلف به تحلیلگران و دانشمندان داده بیشتر از هر متخصص دیگری نیاز دارند.



دانشمندان و تحلیگران داده و متخصصان هوش مصنوعی و یادگیری ماشین در رتبه‌های **اوّل** و **دوم** فهرست شغل‌های پر تقاضا تا **۲۰۲۲** قرار دارند. همچنین متخصصان کلان‌داده، متخصصان فناوری‌های نوین و هوشمند و متخصصان مخابرات و انتقال دیجیتال در بین **۱۰** شغل برتر تا سال **۲۰۲۲** حضور دارند.

علم داده به عنوان جذاب‌ترین شغل قرن ۲۱ معرفی شده است [۳]. این حوزه با یک نرخ نجومی در حال رشد بوده و شرکت‌ها و سازمان‌های متعددی به دنبال استخدام نیروهای متخصص هستند تا بتوانند معدن طلای داده‌ها را کاوش کرده و بدین ترتیب با کمک داده‌ها موجب هدایت بهتر و مؤثرتر کسب وکار شوند. IBM پیش‌بینی کرده است که تا سال ۲۰۲۰ تعداد جایگاه‌های شغلی مرتبط با علم داده در ایالت متحده آمریکا به میزان ۳۶۴۰۰۰ شغل افزایش یافته و به عدد ۲،۷۲۰،۰۰۰ شغل خواهد رسید.

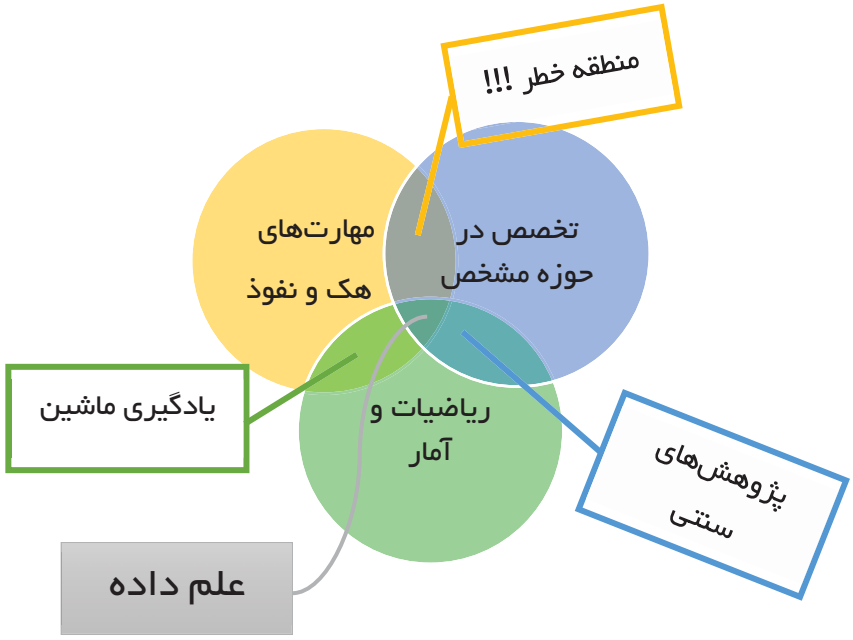
از سوی دیگر، در سال‌های اخیر با پدیده‌ی رو به رشد دیگری مواجه بوده‌ایم، و همچنان هستیم، و آن ظهور رشته‌ی دانشگاهی علوم داده در دانشگاه‌های معتبر و بزرگ آمریکا چون هاروارد[۴] ، استنفورد[۵] و کالیفرنیا برکلی [۶] و ... می‌باشد. در کشورمان نیز حرکت‌های امیدبخشی را شاهد بوده‌ایم و آن تصویب گرایش داده‌کاوی زیرمجموعه رشته‌ی علوم کامپیوتر [۷] و گرایش علوم داده زیر مجموعه رشته‌ی ریاضی کاربردی [۸] است.

[۱] World Economic Forum

تا سال ۲۰۲۰ تعداد ۳۶۴۰۰۰ شغل جدید در حوزه علم داده در ایالت متحده آمریکا ایجاد می‌شود. حداقل حقوق سالیانه ۸۰۲۶۵ دلار بوده و ۸۱ درصد این مشاغل به سابقه کار ۳ تا ۵ سال یا بیشتر نیاز دارند.

بعد از ارائه‌ی این آمار و ارقام در مورد این حوزه علمی و فضـــای کاری جدید آن، به ســـراغ تعریف علم داده می‌رویم. واقعیتی که در صـنعت وجود دارد این اسـت که علم داده آن‌قدر جدید است که شرکت‌ها هنوز یک مسیر مشخص و ثابتی برای تربیت نیروی متخصص این حوزه ندارند. این حوزه هنوز در دوران طفولیت خود بســر می‌برد و لذا تعاریف متنوعی از مرزهای آن وجود دارد. شاید مهم‌ترین سوالی که ذهن شما را درگیر نموده است این باشد که چگونه در حوزه علم داده مشـــغول به کار شـــویم؟ اجازه دهید این ســـوال را این‌گونه مطرح کنیم که مباحث نظری و مهارت‌هایی که باید بر آن‌ها مسلط شویم، چیست؟ خوب، پاسخ این است که کمی از همه چیز، یعنی ریاضیات، آمار، یادگیری ماشین، علوم کامپیوتر، هوش تجاری، مدل سازی و برنامه‌نویسی و ...) و البته یک مهارت اصـلی و ضـروری و آن تخصـص در یک حوزه کاری خاص، برای مثال: اقتصاد، مالی، علوم اجتماعی، بیوانفورماتیک، مهندسی رباتیک، مدیریت سرور، فیزیک، پزشکی و ... . نمودار وِن تعریف شـده توسـط درو کانوی[۱] به خوبی نشـان می‌دهد که علم داده چیسـت و چرا دانشمندان و تحلیلگران داده از ارزش و اهمیت بالایی در دنیایی امروز برخوردار هستند.

[۱] Drew Conway's Data Science Venn Diagram



منظور از «منطقه خطر» در این نمودار شاید به هکرها و کدنویسانی اشاره دارد که با نفوذ به سیستم‌های کامپیوتری، امنیت آن‌ها را به مخاطره می‌اندازند.

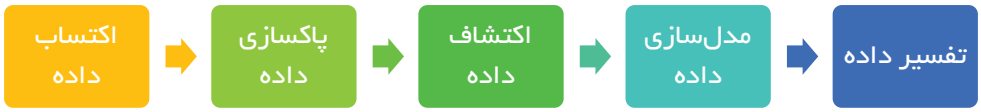
با توجه به این نمودار مشخص است که علم داده از تقاطع سه حوزه اصلی تشکیل شده است:

- **ریاضیات و آمار**: استفاده از فرمول‌ها و روابط برای مدل‌سازی مسئله و تحلیل آن.
- **برنامه‌نویسی**: توانایی کدنویسی و توسعه برنامه‌ها برای ایجاد خروجی مطلوب بر روی کامپیوتر.
- **دانش وابسته به حوزه کاری**: یعنی همان شناخت و درک درست حوزه‌ی مسئله (رباتیک، پزشکی، مالی، علوم اجتماعی و ...).

خوب است در این‌جا مقایسه‌ای بین آمار و علم داده داشته باشیم. در واقع به این فکر کنید که آیا علم داده همان آمار است؟ یا آیا علم داده بدون آمار نیز ممکن است؟ در مقاله‌ی «۵۰ سال علم داده» [۹] به اختلاف سلیقه‌ی بین آماردان‌ها و دانشمندان علوم کامپیوتر در مورد واژه‌ی علم داده به تفصیل پرداخته شده است. به نظر برخی آماردان‌ها دانشمند داده صرفاً یک واژه‌ی جذاب برای

آماردان بوده و علم داده نامی است که در سال‌های اخیر در بازارکار و صنعت برای علم آمار بکار برده می شـــود. با این وجود، همانطور که در این کتاب خواهیم دید، علم داده بدون آمار، نه تنها امکان‌پذیر بوده بلکه لذت‌بخش‌تر خواهد بود. به بیان روشـن‌تر علم آمار فقط یک قسـمت اندک، اما بسیار مهم، از علوم داده را تشکیل می‌دهد.

رویکردی که در این کتاب در پیش گرفته شده است یک رویکرد کاملاً عملگرایانه می‌باشد. لذا برای این کتاب یک تعریف خیلی کاربردی از علم داده را که توسط میسون و ویگینز در سال ۲۰۱۰ ارائه شـده اسـت [۱۲-۱۰]، برگزیده‌ایم. آن‌ها علم داده را بر اسـاس پنج گام زیر توصـیف کرده‌اند: ۱- اکتساب داده ، ۲-پاکسازی داده ، ۳- اکتشاف داده ، ۴- مدل سازی داده ، ۵- تفسیر داده.



تعریف شماره یک از علم داده در عمل. در این تعریف مرحله «پاکسازی داده» جزو مراحل اصلی فرایند علم داده می‌باشد.

آن‌چه که بسـیار اهمیت دارد این اسـت که حوزه‌ای که مسـئله در آن تعریف شـده اسـت را به خوبی بشـناسـیم تا درک درسـتی از مسـئله برای ما حاصـل شـود. این مرحله شـامل مطرح کردن سـؤالات مهم و قابل توجه، هم‌فکری و مشـورت با افراد متخصـص آن حوزه، تهیه‌ی مجموعه داده‌های مورد نیاز، تعریف متغیرهای کلیدی و مهم‌تر از همه تعیین هدف پروژه، یعنی آن‌چه که باید پیش‌بینی یا کشـف شـود، می‌باشـد. لذا می‌توانیم قبل از این مراحل پنج‌گانه یک پیش-مرحله جدید با عنوان طرح سـوالات قابل توجه اضـافه کنیم. در [۱۳, ۱۴] این مراحل پنج‌گانه اندکی متفاوت تعریف شـده‌اند: ۱- طرح پرسـش ، ۲- اکتسـاب داده ، ۳- اکتشـاف داده ، ۴- مدل‌سـازی داده ، ۵- مباحثه و تصویرسازی داده.

تعریف شماره دو از علم داده در عمل. تفاوت این تعریف با تعریف قبلی اضافه شدن یک پیش–مرحله با عنوان «طرح پرسش» می باشد.

این تعریف پاکسازی داده را جزو مراحل اصلی ذکر نکرده است اما همانطور که در [۱۵] بیان شده است ۸۰ درصد کار در یک پروژه علم داده مربوط به مرحله‌ی پاکسازی داده‌ها می‌باشد. از آن جا که خط فرمان یونیکس ابزارهای متعدد و متنوعی را برای این منظور در اختیار ما قرار می‌دهد، در این کتاب این مرحله به تفسیر و در قالب پروژه‌های عملی مختلف توضیح داده می شود. بنابراین اگر بگوییم این کتاب لازمه‌ی پیش‌برد بیش از ۸۰ درصد از هر پروژه‌ی علم داده است، سخن به گزافه نگفته‌ایم. خوب است در اینجا به قسمتی از سخنرانی دکتر سید جواد کاظمی‌تبار در همایش آشنایی با علوم داده که آذرماه ۹۶ در دانشگاه شریف [۱۶] ایراد گردیده است، اشاره کنیم. ایشان در شروع صحبت درباره اهمیت اسکریپت‌نویسی خط فرمان و استفاده از عبارت‌های باقاعده چنین گفتند:

«... ذهنیتی که در دانشگاه در مورد داده‌کاوی وجود دارد این است که داده‌کاوی به طریق خاصی انجام می‌شود و حتما باید از روش‌های آماری و یادگیری ماشین در آن استفاده کنیم. امّا زمانی‌که وارد یک پروژه واقعی می‌شوید، آنجاست که متوجه می‌شوید لزوماً این گونه نیست. بلکه داده‌کاوی یعنی در درجه‌ی اوّل یادبگیریم با داده‌ها راحت باشیم، با داده‌ها بازی کنیم و از حجم زیاد آن‌ها نترسیم... در بین دوستان وطنی که در این حوزه مشغول‌اند، کم‌تر دیده‌ام که به متن‌کاوی توجه کنند. این در حالی بود که در شرکت‌های بزرگ دنیا اولین چیزی که از شما خواسته می‌شود تا یاد بگیرید، متن‌کاوی و کار با عبارت های باقاعده (Regex) است. برای یادگیری عبارت های باقاعده مجبورید تا یک زبان اسکریپت‌نویسی بیاموزید. چرا که اولاً این داده‌ها فایل‌های متنی (مثلاً CSV) هستند و ثانیاً شما اغلب اجازه‌ی ذخیره آن‌ها را روی سیستم خود نداشته و باید به یک سرور لینوکسی متصل و از طریق ترمینال آن روی پروژه کار کنید. لذا قواعد متن‌کاوی و نیز دستورات یونیکس را باید به خوبی فرا گیرید... شما برای آن‌که متخصص داده شوید، لازم نیست مهندس کامپیوتر بشوید. امّا باید یک سری ابزارها را به طور هوشمندانه انتخاب کرده و یاد بگیرید و یکی از مهم‌ترین آن‌ها یادگیری یک زبان اسکریپت‌نویسی است».

بدون هیچ داده‌ای، علم داده‌ی چندانی وجود نخواهد داشت. بنابراین گام نخست بدست آوردن مجموعه داده‌ها می‌باشد. داده‌ها از منابع مختلف و به شیوه‌های متنوعی بدست می‌آیند. برای مثال لازم است تا در آغاز هر پروژه یک یا چند گام زیر را انجام دهیم:

- دانلود داده روی یک سرور یا صفحه وب (معمولاً فایل‌های فشرده هستند)
- بازیابی داده از روی یک دیتابیس (برای مثال MySQL) و یا API (برای مثال تلگرام، توئیتر)
- استخراج داده از یک فایل صفحه گسترده (برای مثال CSV)
- خودمان داده‌ها را تولید کنیم، برای مثال خواندن اطلاعات سنسورها و یا استفاده از شیوه‌های سنتی مثل جمع‌آوری پرسش‌نامه.

گام بعدی پاکسازی داده‌ها می‌باشد. معمولاً داده‌ی اکتساب شده دارای مقادیر از دست رفته، خطاها، نویسه‌های عجیب و غریب، سطرهای خالی و یا حتی ستون‌های زاید می‌باشد. منظور از ستون‌های زاید، ستون‌هایی است که در تحلیل مورد نظر علاقه‌ای به آن‌ها نداریم. کارهایی که معمولا در این قسمت انجام می‌گیرد شامل موارد زیر است:

- تبدیل فایل‌ها از یک فرمت (پسوند) به فرمت دیگر
- پالایش سطور و استخراج ستون‌های خاص
- جایگذاری مقادیر جدید و تصحیح مقادیر از دست‌رفته

در گام سوم سراغ اکتشاف بر روی داده‌ها می‌رویم. اینجا جایی است که برای اولین بار با داده‌ها و چیزی که بیان‌گر آن هستند سروکار پیدا می‌کنیم. شناخت داده، بدست آوردن آماره‌های آن و تصویرسازی داده‌های خام در این قسمت انجام می‌گیرد. در این مرحله با کمک روش‌های آماری و نمودارها به دنبال یافتن الگوها و مشخصه‌های قابل توجه در مجموعه داده هستیم. اغلب یک نمودار ساده از داده‌های خام بینش‌های مهم و بسیار مفیدی در اختیار قرار می‌دهد که ادامه مسیر پروژه تحلیل داده را به ما دیکته می‌کند. در این کتاب از روش‌های آماری ساده و از طریق ترسیم نمودارها و با کمک فیلترهای متنی یونیکس مرحله‌ی اکتشاف داده را انجام می‌دهیم. علاوه

بر این‌ها در این مرحله با استفاده از کدنویسی کیفیت توصیف مجموعه داده را بهبود داده و آن را برای مرحله‌ی مدل‌سازی آماده می‌کنیم.

در گام چهارم سراغ توصیف ریاضی داده‌ها و یا اینکه چه چیزی را پیش‌بینی می‌کنند، با کمک مدل‌سازی آماری می‌رویم. در این قسمت روش‌های یادگیری ماشین چون خوشه‌بندی، دسته‌بندی، رگرسیون و کاهش ابعاد بروی داده‌ها اعمال می شود. علاوه بر انتخاب و برازش مدل، در این مرحله از معیارهای ریاضی برای سنجش اعتبار مدل انتخابی بهره می‌بریم. برای این منظور داده‌ها به گروه‌های داده آموزش، داده آزمایش و داده اعتبارسنجی متقابل[۱] تقسیم می‌شوند. هنگامی‌که مهارت کافی در یادگیری ماشین پیدا کنید، قادر خواهید بود تا الگوریتم‌های مناسب برای یک کاربرد خاص را براحتی شناسایی نمایید. در این کتاب به جزء گام چهارم، تمامی گام‌های دیگر در پروژه‌های مورد بررسی وجود دارند. دلیل این امر آن است که خط فرمان یونیکس برای پیاده‌سازی مدل‌های آماری و یادگیری ماشین از ابتدا مناسب نیست. لذا مرسوم است تا برای پیاده‌سازی الگوریتم‌های یادگیری ماشین در پروژه‌های علم داده از زبان‌هایی چون پایتون و R استفاده می‌شود.

گام آخر تفسیر داده‌ها و تحلیل نتایج است. این مرحله شامل نتیجه‌گیری، بررسی و ارزیابی نتایج و بحث و تبادل نظر با صاحبنظران و متخصصان شرکت یا سازمان می‌باشد تا مدیران بتوانند تصمیمات مناسبی اتخاذ کرده و موجب رونق و شکوفایی کسب وکار شان شوند. شما باید قادر باشید تا نتایج را به بهترین شکل که بیانگر واقعیت‌های موجود باشد، تصویر سازی نمایید. علاوه بر این شما باید مهارت داستان سرایی داده[۲] را نیز کسب نمایید. اهیمت این قسمت تا آنجا ست که کتاب‌هایی با همین عنوان چاپ شده‌اند. بر اساس این واقعیت که اکثر مدیران دارای دانش آمار و علم داده نیستند، لازم است تا هر آنچه در دل داده‌ها نهفته است در قالب یک داستان جذاب و واقعی بیان شده تا توسط یک فرد تجاری قابل درک باشند.

---

[۱] cross-validation set

[۲] data storytelling

تا اینجا به تفصیل در مورد علم داده و اهمیت آن صحبت کردیم. در بخش پایانی این پیش‌گفتار به فصل‌بندی این کتاب می‌پردازیم. آن‌چه که ما در این کتاب روی آن تمرکز کرده‌ایم، پردازش و کار روی داده‌های متنی است. اما دلیل اهمیت زبان بَش در حوزه علم داده و یا به بیان دیگر دلیل کاربرد بَش در تحلیل داده‌های متنی را باید در تاریخچه‌ی سیستم یونیکس ریشه‌یابی کرد. در سال‌های اوّلیه‌ی توسعه‌ی سیستم یونیکس، علاقه‌ی شدیدی روی پردازش متن با کمک این سیستم وجود داشت؛ به طوری که اوّلین محصول تولید شده بر اساس سیستم یونیکس، یک ابزار پردازش متن برای استفاده در دپارتمان ثبت اختراعات آزمایشگاه بِل بود.

این کتاب در دو بخش تدوین شده است: ۱- **اسکریپت‌نویسی بَش** (Bash) و ۲- **تحلیل چالش‌های دنیای واقعی**. بخش اوّل کتاب به مبانی اسکریپت نویسی بَش، عبارت‌های با قاعده، و پردازش متن با کمک فیلترهای یونیکس می‌پردازد. این بخش دانش کافی برای مطالعه و درک این کتاب و پروژه‌هایی که در بخش دوم معرفی می‌شوند، در اختیار شما قرار می‌دهد. با این وجود خواننده‌ی علاقه‌مند برای مطالعه‌ی مباحث پیشرفته‌تر خط فرمان می‌تواند به مراجعی که در انتهای کتاب معرفی شده‌اند، مراجعه نماید. این بخش شامل ۸ فصل ابتدایی کتاب بوده که محتویات هر فصل در زیر خلاصه شده است. در انتهای هر فصل یک قسمت خودآزمایی همراه با جواب قرار داده شده است.

در **فصل اوّل** ابتدا تاریخچه‌ی یونیکس و بَش ارائه می‌شود و سپس با معرفی چند دستور ساده، شما را با محیط خط فرمان آشنا می‌کنیم. در انتهای این فصل از شما می‌خواهیم تا در مورد استفاده از دستورات کوچک و ساده خط فرمان برای ساختن برنامه‌های کاربردی فکر کنید. به طور خاص‌تر از شما خواسته می‌شود تا مسئله‌ی ساده و پرکاربرد تغییر نام گروهی هزاران فایل را با کمک ابزارهای موجود، حل نمایید.

فصل دوم در مورد اجراهای مختلف دستورات، نوع دستورات، فایل‌های راه‌انداز پوسته و نحوه‌ی صدور متغیر به آن‌ها، می‌باشد. فایل profile. و bashrc. که به ترتیب محل نگهداری

متغیرهای محیطی و محل نگهداری توابع و جانشین‌ها هستند، به طور کامل بررسی می شوند. در ادامه فصل اسکریپت سنتی «سلام دنیا» را تمرین می‌کنیم و گام‌به‌گام با معرفی دستورات جدید بَش، ویژگی‌های جدیدی به آن اضافه می‌کنیم. بعلاوه در این فصل به متغیرهای محیطی در محیط ویندوز و نحوه ویرایش آن‌ها اشاره می‌شود.

در **فصل سوم** ابزارهای ضروری برای مدیریت یک سیستم یونیکس معرفی می‌شوند. در ابتدا مباحث فایل، پروسه و سیگنال بررسی می‌شوند. پروسه‌های پیش‌زمینه و پس‌زمینه، پروسه‌های والد و فرزند، و زیر-پوسته (زیر-شِل) از جمله مواردی هستند که در این قسمت توضیح داده شده اند. در قسمت انتهایی فصل به مبحث بسیار مهم مجوزهای دسترسی و حق مالکیت فایل‌ها و پوشه‌ها پرداخته می‌شود. تغییر مجوزها به دو روش نمادین و هشت‌هشتی آموزش داده شده است.

ورودی‌ها و خروجی‌ها در زبان بَش موضوع **فصل چهارم** کتاب است. خواندن ورودی صفحه کلید، چاپ بر روی ترمینال، بازهدایت به یک فایل/ از یک فایل، و نیز بازهدایت خروجی یک دستور به ورودی دستور دیگر در این فصل به طور مفصل بحث شده‌اند. در ابتدا ابزارهای یونیکس برای مرور و مشاهده محتویات فایل‌های متنی معرفی می‌شوند. سپس برای خواندن ورودی دو روش مرسوم آموزش داده شده است. دستور read برای خواندن ورودی کاربر و ذخیره آن در یک متغیر استفاده می‌شود. عملگر «سند اینجایی» یکی دیگر از شیوه‌های خواندن اطلاعات در زبان بَش است. تمامی خطوط مشخص شده بین کلمه‌ی جداکننده یک سند اینجایی یکی پس از دیگری خوانده می‌شوند. برای چاپ روی خروجی استاندارد که به صورت پیش‌فرض ترمینال است، از دو دستور echo و printf بهره می‌بریم. جریان‌های ورودی، خروجی و خطای استاندارد، که به ترتیب با stdin، stdout و stderr مشخص می‌شوند، با کمک عملگرهای بازهدایت و واصف فایل‌ها ارجاع داده می‌شوند. واصف فایل و نحوه تغییر آن‌ها توسط دستور exec در ادامه بحث می‌شوند. در انتهای فصل نیز به لوله یا «پایپ» برای خوراندن ورودی یک دستور به خروجی دستور، دیگر به همراه چندین مثال توضیح داده می‌شود.

فصل پنجم یکی از مهم‌ترین فصل‌های این کتاب است که به انواع «جانشینی» در بَش می‌پردازد. متغیرها و توابع نیز در این فصل بحث شده اند. در ابتدا سراغ متغیر ها و نحوه تنظیم و حذف آن ها می‌رویم. در بَش برای تعریف یک متغیر نیازی به اعلان نوع آن نیست. هم‌چنین مقادیر متغیرها به صورت یک رشته ذخیره می‌شوند. در این قسمت انواع متغیرهای اسکالر، آرایه‌ها و اعداد صحیح بحث شده‌اند. بعلاوه به روش‌های مختلف بسط یک متغیر آرایه و نیز آرایه‌های شرکت پذیر می‌پردازیم. متغیر های زبان بَش در سه دسته مختلف یعنی، متغیر های محیطی، متغیرهای محلی و متغیرهای پو سته مقایسه می شوند. محیط به مجموعه‌ای از متغیرها و توابع پو سته اطلاق می‌شود. تابع و محدوده‌ی تعریف متغیرها موضوع قسمت بعدی است. در این قسمت پارامترهای مکانی ا ستفاده شده در داخل بدنه توابع نیز بحث می شوند. قسمت پایانی این فصل به انواع مختلف جانشینی یا بسط در بَش اختصاص یافته است. اهمیت استفاده بجا از نقل‌قول‌های یگانه، دوگانه و نقل‌قول برگشتی در ابتدا توضیح داده می‌شود. در ادامه سراغ عملگرهای ریاضی و شیوه‌های مختلف ارزیابی عبارت‌های ریا ضی می‌رویم. سپس در مورد جانشینی د ستور، جانشینی نام‌فایل و جداسازی کلمات بحث می کنیم. مبحث مهم بسط پارامتر یا متغیر در انتهای فصل آورده شده است. این قسمت مکمل قسمت بسط آرایه‌ها می‌باشد. در این قسمت شکل‌های مختلف بسط یک پارامتر به تفصیل توضیح داده شده است. در پایان اشاره‌ای به متغیرهای ویژه زبان بَش می کنیم.

فصل ششم در مورد ساختارهای کنترل جریان می‌با شد. در واقع این ویژگی ا ست که یک زبان برنامه‌نویسی را می‌سازد. در ابتدا وضعیت خروج و آزمایش بولی توضیح داده می‌شود. و ضعیت خروج یا نتیجه‌ی یک آزمایش بولی در ساختارهای if، while و until و نیز عملگرهای NOT، AND و OR مورد استفاده قرار می گیرد. دستور test و یا معادل آن دستور [..] برای ارزیابی فایل‌ها، رشته و مقادیر عددی مورد استفاده قرار می گیرد. کروشه‌های دوگانه [[..]] و پرانتزهای دوگانه ((..)) به ترتیب برای ارزیابی عبارت‌های باقاعده تعمیم‌یافته و عبارت‌های ریاضی کاربرد دارند. در قسمت دوم این فصل عبارت‌های شرطی بحث می‌شوند. نتایج بدست آمده از

آزمایش یک عبارت، برای تصمیم‌گیری در مورد ادامه‌ی جریان یک اسکریپت استفاده می شود. شکل‌های مختلف ساختار if و نیز عبارت case در این قسمت توضیح داده شده است. قسمت پایانی این فصل به بحث در مورد حلقه‌های for ، while و until اختصاص دارد.

در **فصل هفتم** به سراغ فیلترهای متنی مقدماتی در بَش و عبارت‌های باقاعده می‌رویم. عبارت‌های با قاعده یک حالت قوی‌تری از تطابق الگو نسبت به بسط نام فایل هستند و می‌توانند محدوده‌ی خیلی وسیع‌تری از الگوها را با دقت بیش‌تری بیان کنند. ابزارهای زیادی را در یونیکس می‌توان یافت که قدرت‌شان را از عبارت‌های با قاعده می‌گیرند، از جمله grep، sed و awk. دستور grep به همراه سایر ابزارهای پردازش متن مقدماتی در ابتدای این فصل بحث می‌شوند. قسمت دوم این فصل به عبارت‌های با قاعده اختصاص دارد. انواع عبارت‌های با قاعده از جمله عبارت با قاعده مقدماتی و تعمیم‌یافته با یکدیگر مقایسه و کاربردهای هر یک با ذکر مثال توضیح داده می‌شود. کلاس نویسه، لنگرها و ارجاعات برگشتی سایر مباحثی هستند که در این قسمت بررسی می‌شوند.

فصل هشتم به دو ابزار پیشرفته یونیکس برای پردازش متن، یعنی sed و awk اختصاص دارد. ویرایشگر sed بر خلاف ویرایشگرهای مرسوم با صفحه‌کلید می‌توانید با آن در تعامل باشید، یک ویرایش‌گر جریان است. به این معنی که قبل از پردازش یک فایل متنی، قوانین و دستورات لازم برای ویرایش آن را باید وارد کنیم. این ابزار بوسیله‌ی محدوده آدرس تعریف شده برای آن تعیین می کند روی کدام خطوط باید تغییرات صورت گیرد. دستوراتی که برای ویرایش یک جریان داده می توان استفاده کرد شامل جانشینی، حذف، چاپ و انتقال است. awk یک زبان برنامه‌نویسی کامل مبتنی‌بر فیلد بوده و از ساختارهای برنامه‌نویسی کامل‌تری نسبت دستوررات ساده‌ی ویرایشگر sed برخوردار است. در این زبان برنامه‌نویسی شما می‌توانید با تعریف متغیرها، داده‌ها را ذخیره نمایید. از ساختارهای شرطی و حلقه‌ها کمک بگیرید. بر روی رشته کار کنید. عملیات ریاضی انجام دهید. و در نهایت گزارش‌های قالب‌بندی شده اختصاصی تهیه نمایید.

بخش اوّل کتاب همچون پلی است که فاصله‌ی بین علم و عمل را می‌پوشاند. بعد از یادگیری ابزارها و دستورات معرفی‌شده در این بخش، شما آماده‌اید تا آموخته‌هایتان را روی پروژه‌های واقعی بخش دوم کتاب اعمال نمایید. در فصول نهم تا پانزدهم، گام‌های علم داده را در چالش‌های واقعی بکار می‌بریم. فصل شانزدهم نیز شامل ۸ چالش دیگر خواهد بود که از حوزه‌های مختلف در صنعت، خدمات و تجارت انتخاب شده و از خواننده خواسته شده تا با تفکر و کار تحلیلی بر روی این پروژه‌ها دانش خود در این حوزه تثبیت نماید.

اگر می‌خواهید بر روی داده‌های تولید شده توسط برنامه‌های سفارش آنلاین غذا مثل اسنپ‌فود، چیلیوری، ریحون و یا ...کارکنید، **فصل نهم** برای شما نوشته شده است. در این فصل چون دسترسی به داده رستوران‌های ایرانی مقدور نبود، از یک مجموعه داده آزاد موجود در اینترنت برای انجام تحلیل بهره برده‌ایم. پروژه تعریف شده در این فصل به تحلیل عادت‌های غذایی مشتریان یک رستوران زنجیره‌ای در ایالت متحده آمریکا می‌پردازد. در حال حاضر نه تنها در تهران بلکه در اکثر شهرهای کشور، امکان سفارش غذا به صورت آنلاین از طریق برنامه‌های کاربردی تلفن‌های همراه و یا وب‌سایت‌های بی‌شماری که برای این منظور توسعه پیدا کرده‌اند، میسر شده است. صاحبان چنین کسب و کارهایی، اعم از توسعه‌دهندگان برنامه‌های کاربردی و نیز صاحبان رستوران‌ها، می‌توانند با کمک علم داده خدمات مطلوب‌تری (غذای با کیفیت، قیمت مناسب، کاهش زمان تحویل و بسته‌بندی سالم) را به مشتریان خود ارائه دهند. آن‌ها می‌توانند با در نظر گرفتن ذائقه و سلیقه‌ی مشتریان، ضمن بهبود رضایت و به تبع آن کسب وفادرای مشتریان، به موفقیت کسب‌وکار خود کمک کنند.

اگر قصد ادامه تحصیل در خارج از کشور را دارید و می‌خواهید بر اساس واقعیت‌های نهفته و مجهول در دل داده‌ها، بهترین دانشگاها را مشخص نمایید، **فصل دهم** برای شماست. نظام رتبه‌بندی دانشگاه‌ها در عصر حاضر بسیار متداول بوده و توسط موسسه‌های مختلفی در سراسر دنیا انجام می‌پذیرد. معمولاً شاخص‌های ارزیابی شامل مواردی چون جوایز معتبر دریافت شده توسط دانشجویان و اساتید، تعداد مقالات، تعداد ارجاع‌ها (استناد به مقالات)، کیفیت هیئت علمی، و

نسبت هیئت علمی به دانشجویان هستند. در این فصل بروی یک مجموعه داده آزاد که شامل رتبه‌بندی دانشگاه‌های کشور ایالات متحده آمریکا در سال ۲۰۱۷ میلادی است، کار می‌کنیم.

اگر به تحقیق بروی رفتار افراد در شبکه‌های اجتماعی و تاثیرگذاری سایرین و موضوعات مختلف بر نظر و عقاید آن‌ها علاقه دارید، مطمئناً **فصل یازدهم** برای شما جذاب خواهد بود. تلفیق علم داده، علوم اجتماعی و شبکه‌های پیچیده حوزه‌ی جدیدی با نام علوم اجتماعی محاسباتی یا واقعیت کاوی را بوجود آورده است، که البته موضوع این فصل نیست. حوزه دیگر مرتبط با شبکه‌های اجتماعی، بازاریابی و تبلیغات شبکه‌های اجتماعی است. این شبکه‌ها نقش مهمی در شکل‌دهی و آینده‌ی کسب و کارها بازی می‌کنند. لذا لازم است با تجزیه و تحلیل دقیق داده‌های صفحه‌ی شرکت یا سازمان در شبکه‌های اجتماعی، آینده کسب وکار را پیش‌بینی نماییم. در این فصل ابتدا ابزاری برای استخراج داده‌های موردنیاز از صفحات فیس‌بوک معرفی می‌شود و سپس بر اساس دادهای تولیدشده، به بررسی پست‌های یک صفحه خاص می‌پردازیم.

اگر به کاربرد علم داده در حوزه های سیاسی و اجتماعی و به طور خاص مبحث مهم امنیت اجتماعی در یک کشور علاقه‌مند هستید، مطالعه‌ی **فصل دوازدهم** بسیار برای شما مفید و جالب خواهد بود. تا کنون مؤسسات مختلفی در سراسر دنیا، به مقایسه و امتیازبندی شهرهای جهان بر اساس معیارهایی چون امنیت، آزادی کسب‌وکار، آموزش، حمل‌ونقل، تامین اجتماعی، درآمد سرانه، سلامت و درمان، آب‌وهوا و اوقات فراغت پرداخته‌اند. یکی از معیارهای مهم در این ارزشیابی‌ها، معیار محیط سیاسی و اجتماعی می‌باشد که به ثبات سیاسی، ارتکاب جرم و جنایت و اجرای قانون اشاره دارد. در این فصل چون به داده‌های ملی در حوزه‌ی امنیت اجتماعی دسترسی نداشتیم، از یک مجموعه داده که مربوط به کشور استرالیا می‌باشد، استفاده کرده‌ایم. با کمک این مجموعه داده و با بررسی جرائم و جنایات انجام شده، نرخ ناامنی در شهرهای مختلف را پیدا می‌کنیم.

فصل سیزدهم برای اهالی فرهنگ و ادب نگاشته شده است و سعی دارد تا با رویکردی عملگرایانه و مبتنی‌بر کدنویسی، آثار منظوم و نمایشنامه‌های دوره‌ی رنسانس در اروپا را از زاویه‌ی متفاوت مورد نقد و واکاوی قرار دهد. به طور خاص، در این فصل به بررسی و تحلیل آثار ادبی دوره‌ی رنسانس در انگلستان می‌پردازیم و می‌خواهیم دریابیم که شاعران معروفی هم‌چون شکسپیر بیشتر از چه واژگانی در اشعار و نمایش‌نامه‌های خود بهره برده‌اند. مجموعه داده‌ای که برای انجام این پروژه مورد استفاده قرار گرفته است، نتیجه‌ی یک کار پژوهشی در دانشگاه نیوکاسل استرالیا می‌باشد.

اگر به بانکداری، بورس و بازارهای مالی علاقه دارید، فصل **چهاردهم** با یک مثال ساده قابلیت‌های بَش در تولید و مدیریت صورت‌های مالی را به شما نشان می‌دهد. صورت‌های مالی از جمله معروف‌ترین داده‌هایی هستند که بعد از پایان هر عملیات تجاری و غیرتجاری تولید یا بروزرسانی می‌شوند. در واقع داده‌های موجود در یک صورت حساب مالی مهم‌ترین و اساسی‌ترین اطلاعات را برای مدیریت و هدایت هر کسب وکاری فراهم می‌آورند. در حوزه بانکداری یکی از عملیاتی که هر روز و یا در هر ساعت انجام می‌گیرد، بالانس حساب جاری می‌باشد. در این فصل بر اساس یک صورت مالی که به طور دستی تولید شده است، میزان واریز، برداشت و چک‌های صادره از حساب جاری را مدیریت می‌کنیم. پس از پایان این فصل قادر خواهید بود تا برای صورت‌های مالی بزرگ و حجیمی که توسط سامانه‌های اینترنتی بانکی یا بورس تولید می‌شوند، برنامه‌هایی مشابه به زبان بَش بنویسید.

فصل پانزدهم یک فصل کاملاً منحصر به فرد و ویژه بوده و علاقمندان خاص خود را در حوزه‌های هوش مصنوعی، علوم کامپیوتر و علوم و فنون قرآن کریم دارد. عنوان این فصل را قرآن‌کاوی یا قرآن پژوهی رایانه‌ای انتخاب کرده‌ایم تا به خوبی منعکس‌کننده‌ی آنچه در این اینجا به دنبال آن هستیم، باشد. به بیان دیگر، رویکردی نوین و مبتنی‌بر کدنویسی در حوزه علوم و فنون قرآن کریم معرفی می شود. در این فصل بجای استفاده از متن انگلیسی قرآن، روی متن عربی آن

متمرکز شـــده‌ایم تا بدین وســـیله بینش و دانش دقیق و کاملی از تحلیل کلام الله مجید در اختیار ما قرار بگیرید.

آخرین فصل کتاب ، یعنی **فصل شانزدهم**، به پروژه‌هایی برای تمرین بیشتر اختصاص داده شـــده اســـت. در انتخاب پروژه‌ها ســـعی شـــده اســـت تا چالش‌های متنوعی از حوزه‌های گوناگون انتخاب شود. تمرین‌های متعدد مطرح شده در این فصل، با این هدف انتخاب شده‌اند که خواننده را به شوق آورده و به چالش برانگیزد تا ضمن مطرح نمودن سوالات جالب و اساسی، به شرح و توصـیف بیشـتر مسـائل بپردازد، و در نهایت بتواند با تحلیل‌های آگاهانه و منطقی، ضـمن افزایش بینش خود نسبت به موضوع مورد بررسی، آینده را پیش‌بینی نماید. این تمرین‌ها مثال‌های واقعی از کاربرد علم داده در صنعت، تجارت و تجارب زندگی روزانه می‌باشد.

کتاب «**علم داده کاربردی با کمک اسکریپت‌نویسی بَش**» اولین کتاب فارسی می‌باشد که در مورد زبان بَش نوشته شده است. با توجه به اهمیت سیستم‌های یونیکس و کاربرد روزافزون آن‌ها چه در صنعت و چه در دانشگاه، نگارش کتابی که به طور اختصاصی به زبان برنامه‌نویسی خط فرمان بپردازد، لازم و ضروری بود. بنابراین مطالعه‌ی بخش اوّل این کتاب علاوه بر متخصصان علم داده برای سایر متخصصان و پژوهشگران در علوم مختلف از جمله هوش مصنوعی، رباتیک، فیزیک و ریاضـــی بســـیار مفید خواهد بود. بعد از مطالعه‌ی مطالب این کتاب، خواننده باید پایه‌ی خوبی برای درک علم داده و کاربردهای آن یافته و قادر به انجام پروژه‌های واقعی و چالش‌های پیچیده‌تر با کمک اســـکریپت‌نویســـی بَش و قواعد متن‌کاوی آن باشـــد. بعلاوه و مهم‌تر از همه، خواننده باید بتواند گام‌های اجرایی این کتاب را بدرستی پیاده سازی کرده تا بدین ترتیب، کشف بینش‌های نهفته در دل داده‌ها او را به واقعیت کاوی و یا پیش‌بینی آینده سوق دهد.

از شـما خواننده محترم تقاضـا دارم پس از مطالعه و اسـتفاده از این کتاب نظرات و پیشـنهادات خود در مورد این اثر را به نشـــانی ایمیل *akhorshidi@live.com* ارســـال کرده و یا از طریق صفحه‌ی اختصاصی کتاب در سایت ناشر، کاربردهای این اثر در پروژه‌های کاری و شخصی خود

را با سایر خوانندگان به اشتراک بگذارید. گرچه که ما تمام سعی خود را به کار گرفته‌ایم تا اثری عاری از نقص را تقدیم حضورتان کنیم، اما اشتباهات هیچ وقت قابل اجتناب نبوده و نیستند. لذا خواهشمندیم اشتباهات ظاهری و باطنی اثر را برای ما مکاتبه نمایید. پیشاپیش از حسن نظر و دقت شما سپاسگزاریم. کدها، فایل‌ها، مجموعه داده‌ها، مثال‌ها، پاسخ تمرینات و حل پروژه‌های کتاب در مخزن اختصاصی زیر قرار داده می‌شود.

https://github.com/akhorshidi/Bash4DSBook

برای دریافت فایل‌های کتاب کافی است روی دکمه Clone or download کلیک نموده و گزینه Download Zip را انتخاب نمایید. همچنین می‌توانید از طریق وب‌سایت انتشارات و صفحه اختصاصی این کتاب به این محتویات دسترسی پیدا کنید.

در پایان از همسر عزیزم خانم فاطمه سادات کسائیان‌زاده مهابادی به خاطر سعی و بردباری‌های زیادی که در مدت زمان نگارش این کتاب از خود نشان دادند، و نیز زحمت حروف‌چینی، صفحه‌آرایی و ویرایش ادبی این اثر را متحمل شدند، صمیمانه سپاسگزاری می‌کنم. از کلیه برنامه‌نویسان و اساتید ارجمندی که با ارائه پیشنهادات و نقطه نظرات خود، سطح کیفی این اثر را ارتقاء داده‌اند، خالصانه تقدیر و تشکر می‌کنم. از همکاری کلیه‌ی کارکنان نشر ... بویژه مدیریت محترم آن جناب آقای ... کمال تشکر را دارم. سرانجام، این اثر را تقدیم به همه‌ی مردان و زنانی می‌کنم که برای اعتلای نام این سرزمین خالصانه تلاش می کنند. و این قول حتمی خداوند هست که می‌فرماید:

« وَأَنْ لَيْسَ لِلْإِنْسَانِ إِلَّا مَا سَعَى » نجم/۳۹.