به نام خدا

علم داده از تئوری تا عمل

مؤلفان

محمد جواد جعفری

کارشناسی ارشد مهندسی فناوری اطلاعات

پرمان محمد علیزاده

کارشناسی معماری کامپیوتر

علم داده از تئوری تا عمل

مؤلفان: محمد جواد جعفری
پرمان محمد علیزاده
ناشر: مؤسسه فرهنگی هنری دیباگران تهران
حروفچینی و صفحه‌آرایی: شبنم هاشم زاده
طرح روی جلد: داریوش فرسایی
چاپ: درج عقیق
نوبت چاپ: سوم
تاریخ نشر: ۱۳۹۸
تیراژ: ۱۰۰جلد
قیمت: ۹۵۰۰۰۰ریال
شابک: ۰-۵۸۰-۱۲۴-۶۰۰-۹۷۸
نشانی واحد فروش: تهران، میدان انقلاب، خ کارگر جنوبی، روبروی پاساژ مهستان، پلاک ۱۲۵۱
تلفن: ۶۶۴۱۰۰۴۶-۲۲۰۸۵۱۱۱
کد پستی: ۱۳۱۴۹۸۳۱۸۵
فروشگاههای اینترنتی:
www.mftbook.ir
www.dibagarantehran.com
نشانی تلگرام:@mftbook
نشانی اینستاگرام : Dibagaran_publishing

سرشناسه : جعفری،محمد جواد، ۱۳۶۷-
عنوان و نام پدید آور : علم داده از تئوری تا عمل / مولفان : محمد جواد جعفری،پرمان محمد علیزاده
مشخصات نشر: تهران- دیباگران تهران -۱۳۹۶
مشخصات ظاهری:۳۷۰ ص. مصور.
شابک :۰-۵۸۰-۱۲۴-۶۰۰-۹۷۸
وضعیت فهرستنویسی : فیپا
موضوع : داده کاوی
موضوع : Data mining
شناسه افزوده :محمد علیزاده،پرمان ،۱۳۷۴-
رده بندی کنگره: ۱۳۹۶ ۶ ج ۲ د/۷۶/۹ QA
رده بندی دیویی : ۰۰۶/۳۱۲
شماره کتابشناسی ملی : ۴۶۶۳۸۷۷

فهرست مطالب

مقدمه ناشر

خط مشی کیفیت انتشارات مؤسسه فرهنگی هنری دیباگران تهران در عرصه کتاب هایی است که بتواند خواسته های به روز جامعه فرهنگی و علمی کشور را تا حد امکان پوشش دهد.

حمد و سپاس ایزد منان را که با الطاف بیکران خود این توفیق را به ما ارزانی داشت تا بتوانیم در راه ارتقای دانش عمومی و فرهنگی این مرز و بوم در زمینه چاپ و نشر کتب علمی دانشگاهی، علوم پایه و به ویژه علوم کامپیوتر و انفورماتیک گام‌هایی هرچند کوچک برداشته و در انجام رسالتی که بر عهده داریم، مؤثر واقع شویم.

گستردگی علوم و توسعه روزافزون آن، شرایطی را به وجود آورده که هر روز شاهد تحولات اساسی چشمگیری در سطح جهان هستیم. این گسترش و توسعه نیاز به منابع مختلف از جمله کتاب را به عنوان قدیمی‌ترین و راحت‌ترین راه دستیابی به اطلاعات و اطلاع‌رسانی، بیش از پیش روشن می‌نماید.

در این راستا، واحد انتشارات مؤسسه فرهنگی هنری دیباگران تهران با همکاری جمعی از اساتید، مؤلفان، مترجمان، متخصصان، پژوهشگران، محققان و نیز پرسنل ورزیده و ماهر در زمینه امور نشر درصدد هستند تا با تلاش‌های مستمر خود برای رفع کمبودها و نیازهای موجود، منابعی پُربار، معتبر و با کیفیت مناسب در اختیار علاقمندان قرار دهند.

کتابی که در دست دارید با همت **"آقایان محمد جواد جعفری- پرمان محمد علیزاده "** و تلاش جمعی از همکاران انتشارات میسر گشته که شایسته است از یکایک این گرامیان تشکر و قدردانی کنیم.

کارشناسی و نظارت بر محتوا: زهره قزلباش

در خاتمه ضمن سپاسگزاری از شما دانش‌پژوه گرامی درخواست می‌نماید با مراجعه به آدرس dibagaran.mft.info (ارتباط با مشتری) فرم نظرسنجی را برای کتابی که در دست دارید تکمیل و ارسال نموده، انتشارات دیباگران تهران را که جلب رضایت و وفاداری مشتریان را هدف خود می‌داند، یاری فرمایید.

امیدواریم همواره بهتر از گذشته خدمات و محصولات خود را تقدیم حضورتان نماییم.

مدیر انتشارات

مؤسسه فرهنگی هنری دیباگران تهران
Publishing@mftmail.com

مقدمه مولف

مرور زندگی دانشمندان بزرگ چه در قرون اخیر و چه دورتر منبع الهام بخش مناسبی برای انسان امروز است.نکته ای که به وفور در این بررسی ها مشاهده می شود منابع اطلاعاتی و نوع دسسترسی به آن ها توسط انسان است.تعدادی کتاب از گذشتگان یک علم خاص،مقداری مشاهدات محدود و در ادامه تحلیل و نتیجه گیری و ثبت نتایج برای استفاده آیندگان.چالش در این موضوع داشتن منابع اطلاعاتی بیشتر چه از طریق مشاهده و چه از طریق متن های نوشته شده بود.شاید شما هم داستان فردی را شنیده باشید که برای تهیه کتاب گرسنگی می کشید.

برگردیم به عصر حاضر،به قسمت تنظیمات حافظه کامپیوتر رومیزی،تلفن همراه هوشمند و یا سایر تجهیزات همراه خود نگاهی بیندازید.احتمالا تعداد زیادی کتاب با موضوعات مختلف در دسترس شما هست.اگر اهل مطالعه باشید این تعداد ممکن است به صدها عنوان هم برسد.می توانید سری به کتابخانه های اطراف خود بزنید یا یک سرچ ساده در گوگل در رابطه با عنوانی خاص داشته باشید.

دایره المعارف های بسیار بزرگ و کتابخانه های آنلاین را فراموش نکنید!شاید به دلایلی حوصله مطالعه کتاب های قطور را نداشته باشید.بیایید از فضای کتاب و کتاب خواندن خارج شویم.امروز هاوکینز جدیدترین نظریه خود را در صفحه شخصی اش در فضای مجازی منتشر کرد.نظر یکی از هم کلاسی های دوران دانشگاه را در مورد قسمت جدید فیلم جنگ ستارگان در صفحه مربوط به این فیلم خواندم.

خوب بهتر است کمی به کارهای شخصی خود برسیم.با دوستی که قرار است پس از سال ها به دیدنش بروم توسط سیستم آدرس دهی گوگل نقشه محل قرار را مشخص کردیم.همچنین سیستم پیشنهاد دهنده گوگل به ما یک رستوران خوب معرفی کرد.چند عکس یادگاری با گوشی هوشمند خود انداختیم و با سایر دوستان خودمان از طریق اینستاگرام به اشتراک گذاشتیم.البته آن روز همه چیز به خوبی و خوشی پیش نرفت و مجبور شدیم در نهایت به علت مسمومیت غذایی ناشی از حساسیت دوستم به بیمارستان با استفاده از نقشه گوگل برویم.

می توان ساعت ها از حجم بسیار عظیم منابع اطلاعاتی که ثانیه به ثانیه در حال ذخیره شدن و استفاده است صحبت کرد.اطلاعات زیادی از ما در حال ذخیره شدن و اطلاعات زیادی از بیرون در

معرض استفاده ما قرار گرفته اند.همه ی مواردی که گفته شد امروزه نقش کتاب ها و مشاهدات کمیاب گذشته را بازی می کنند که انسان باید برای بدست آوردنشان گرسنگی می کشید.

این مقدار از منابع اطلاعاتی و نیاز به مدیریت درست آن ها چالش جدیدی به نام استفاده از اطلاعات را بوجود آورد.در پس این چالش ها "علم داده" پا به عرصه زندگی گذاشت.سال ها پیش آقایان توماس دونپورت و دی جی پاتیل در سال 2012 در مقاله «علم داده: جذاب‌ترین شغل قرن بیست و یکم» متخصصین علم داده را این طور تعریف می‌کنند: کسانی که می‌دانند چگونه می‌توان از انبوه اطلاعات بدون ساختار پاسخ سوالهای کسب‌وکار را پیدا کرد. استنتون در سال 2013 علم داده را این طور تعریف می‌کند: علم داده رشته در حال ظهوری است که به جمع‌آوری، آماده‌سازی، تحلیل، بصری‌سازی، مدیریت و نگهداشت اطلاعات در حجم بالا می‌پردازد. دریسکول در سال 2014 علم داده را این طور تعریف می‌کند: علم داده مهندسی عمران داده‌هاست.

هدف کلی این کتاب آشناسازی خواننده با دنیای "علم داده" و موضوعات مرتبط با آن است.خواننده پس از مطالعه این کتاب شناخت کلی با زوایای مختلف این دنیای بزرگ خواهد داشت تا پس از مطالعه هر بخش بتواند مسیر خود را در راه تبدیل شدن به یک متخصص داده به درستی دنبال کند.برای افراد غیر مبتدی نیز بخش بندی کتاب بصورتی است که می توانند بخش های مرتبط با حوزه کاری خود را مطالعه کنند هرچند خواندن همه بخش ها جهت داشتن دید بهتر از علم داده مطمِنا خالی از لطف نخواهد بود.در این کتاب پس از بررسی زوایای مختلف این علم به سه بخش اصلی آن می پردازیم که شامل آماده سازی،تحلیل و تصویرسازی اطلاعات می باشد.علم آمار مطمئنا بخشی مهم و جدایی ناپذیر در علوم مرتبط با داده ها است که دراین کتاب توجه ویژه ای به آن شده است.لابلادر کنار مطالب مطرح شده در هر ضمینه ابزارها و زبان های برنامه نویسی کاربردی همراه با مختصری معرفی از نحوه استفاده آن ها عنوان شده است.در فصل های پایانی نگاهی به کاربرد علم داده در محیط زیست،کشف جرائم،خبرنگاری و تجارت الکترونیک انداخته ایم.

این کتاب مقدمه ایست جهت شروع یک راه طولانی که درباره جاده ها،ابزارهای مورد نیاز،مقصد و راه مواجه با هر پستی و بلندی را به شما نشان می دهد که برای لذت بردن و داشتن انرژی و انگیزه کافی در هر سفر دانستن این موارد لازم است.پس از این کتاب در آینده ای نزدیک بصورت خاص سفر خود را با کتاب "کاربرد پایتون در علم داده" ادامه خواهیم داد.