

۲۰۳۷۹۵۱

مجلسه فرهنگی هنری
دیبگران تهران

به نام خدا

مجلسه فرهنگی هنری
دیبگران تهران

کتابخانه داده ها استخراج ذخیره و پردازش

مؤلفان:

دکتر امین نظارات

(عضو هیئت علمی دانشگاه پیام نور)

مهندس محسن رنجبر

مهندس فرزاد ابراهیمی



هرگونه چاپ و تکثیر از محتویات این کتاب بدون اجازه کتبی ناشر ممنوع است. متخلفان به موجب قانون حمایت حقوق مؤلفان، مصنفان و هنرمندان تحت پیگرد قانونی قرار می گیرند.

◀ عنوان کتاب: کلان داده ها - استخراج، ذخیره و پردازش

◀ مولفان: دکتر امین نظارات

◀ مدرس محسن رنجبر

◀ مهندس فرزاد ابراهیمی

◀ ناشر: موسسه فرهنگی هنری دیباگران تهران

◀ صفحه آرایی: نازنین نوری

◀ طراح جلد: داریوش قربانی

◀ نوبت چاپ: اول

◀ تاریخ نشر: ۱۳۹۸

◀ چاپ و صحافی: درج عقیق

◀ تیراژ: ۱۰۰ جلد

◀ قیمت: ۷۲۰۰۰۰ ریال

◀ شابک: ۹۷۸-۶۲۲-۲۱۸-۱۴۰-۶

◀ نشانی واحد فروش: تهران، میدان انقلاب،

خ کارگر جنوبی، روبروی پاساژ مهستان،

پلاک ۱۲۵۱

تلفن: ۲۲۰۸۵۱۱۱-۶۶۴۱۰۰۴۶

◀ فروشگاههای اینترنتی دیباگران تهران :

WWW.MFTBOOK.IR

www.dibagaran-tehran.com

www.mftdibagaran.ir

◀ نشانی تلگرام: @mftbook

اپلیکیشن دیباگران تهران را از سایت های اینترنتی دیباگران دریافت نمایید.

سرشناسه: نظارات، امین، ۱۳۵۹-

عنوان و نام پدیدآور: کلان داده ها استخراج، ذخیره و پردازش

/مولفان: امین نظارات، محسن رنجبر، فرزاد ابراهیمی.

مشخصات نشر: تهران: دیباگران تهران: ۱۳۹۸

مشخصات ظاهری: ۲۲۴ص: مصور.

شابک: ۹۷۸-۶۲۲-۲۱۸-۱۴۰-۶

سعیست فهرست نویسی: فیبا

موضوع: داده های کلان Big Data

موضوع: پایگاه های اطلاعاتی-مدیریت

Database management

شناسه افزوده: ابراهیمی، فرزاد، ۱۳۷۰-

شناسه افزوده: رنجبر، محسن، ۱۳۶۳-

رده بندی کنگره: ۷۶۱۰۱

رده بندی دیویی: ۰۰۵/۷۴

شماره کتابشناسی ملی: ۵۷۰۴۱۶۳

فهرست مطالب

۷ مقدمه ناشر

۹ مقدمه مؤلف

فصل ۱

۱۰ تحلیل داده های بزرگ

۱۳ چالش های علم داده

۱۴ معرفی Apache Spark

فصل ۲

۱۹ مقدمه ای بر تجزیه و تحلیل داده ها با Spark و Scala

۲۰ Scala برای متخصص داده

۲۲ مدل برنامه نویسی Spark

۲۳ ارتباط رکورد

۲۴ شروع کردن: پسته Spark و SparkContext

۳۰ های توزیع شده برگشت پذیر

۳۱ آوردن داده از خوشه به کلاینت

۳۷ انتقال کد از کلاینت به خوشه

۳۸ از RDD به فریم های داده

۴۳ تحلیل داده با DataFrame API

۵۰ آمار خلاصه سریع برای دیتافریم

- ۵۲ reshaping و pivoting دیتافریم‌های
- ۵۸ Join کردن دیتافریم‌ها و انتخاب ویژگی‌ها
- ۶۰ آماده‌سازی مدل‌ها برای محیط‌های تولید
- ۶۲ ارزیابی مدل

فصل ۳

- ۶۶ Audioscrobbler مجموعه داده و موسیقی
- ۶۸ مجموعه داده
- ۶۹ الگوریتم حداقل ربات و وب پیشنهاد دهنده
- ۷۲ آماده‌سازی داده‌ها
- ۷۷ ایجاد اولین مدل
- ۷۹ Broadcast متغیرهای
- ۸۲ بررسی نقطه نظرات
- ۸۵ ارزیابی کیفیت پیشنهاد
- ۸۶ محاسبه AUC
- ۸۹ انتخاب hyperparameter
- ۹۲ ایجاد پیشنهادها

فصل ۴

- ۹۶ پیش‌بینی پوشش جنگل با درخت تصمیم
- ۹۷ پیش‌بینی به سوی رگرسیون
- ۹۸ بردارها و ویژگی‌ها
- ۹۹ درختان تصمیم و جنگل‌ها
- ۱۰۲ مجموعه داده Covtype

۱۰۳.....	پیش پردازش داده‌ها
۱۰۵.....	اولین درخت تصمیم
۱۱۶.....	Hyperparameterهای درخت تصمیم
۱۱۷.....	تنظیم درختان تصمیم
۱۲۴.....	جنگل تصمیم‌گیری تصادفی

فصل ۵

۱۲۹.....	تشخیص ناهنجاری در تراز یک شبکه با خوشه‌بندی K-Means
۱۳۰.....	تشخیص ناهنجاری
۱۳۱.....	خوشه‌بندی K-means
۱۳۲.....	نفوذ شبکه
۱۳۲.....	مجموعه داده KDD Cup 1999
۱۳۴.....	برداشت اول در خوشه‌بندی
۱۳۸.....	انتخاب K
۱۴۱.....	مجازی‌سازی با SparkR
۱۴۷.....	نرمال‌سازی ویژگی
۱۵۰.....	مقادیر غیر عددی
۱۵۱.....	استفاده از برچسب‌ها با آنتروپی
۱۵۳.....	خوشه‌بندی در عمل

فصل ۶

۱۵۸.....	کتابخانه MLlib
۱۵۹.....	Decision Trees
۱۶۴.....	Gradient-Boosted Trees (GBTs)

168	Linear Support Vector Machines (SVMs)
171	Logistic regression
174	Naive Bayes
176	Random Forests
180	Decision Trees
186	Gradient-Boosted Trees (GBTs)
189	Isotonic regression
192	Linear least squares, Lasso, and ridge regression
194	Random Forest Regression
198	K-means
200	Bisecting k-means
202	Gaussian mixture
203	Latent Dirichlet allocation (LDA)
206	Power iteration clustering (PIC)
208	Streaming k-means
210	Collaborative filtering
216	Singular value decomposition (SVD)
218	Principal component analysis (PCA)
220	FP-growth
222	Association Rules
223	PrefixSpan

مقدمه مولف:

سالهاست که از نفت به عنوان ارزشمندترین ماده و ثروت بشر یاد می شود. کشورهای دارنده مخازن عظیم نفتی که در منطقه غرب آسیا قرار گرفته اند از جمله ثروتمندترین کشورهای دنیا هستند. اما آیا به راستی این تعبیر کماکان صحیح است؟ بر اساس تحلیلهای موسسات بزرگ اقتصادی دنیای همچون گارتتر، امروزه "داده از نفت ارزشمندتر است." در سال ۲۰۱۸ که به نام سال حکمرانی داده نام گذاری شده بود، کشورهایی ثروتمندتر بودند که داده بیشتری را در اختیار داشته و از پردازش این حجم عظیم از داده ها تولید ارزش افزوده نموده اند. صاحبان داده صاحبان قدرتند. علم داده که مجموعه ای از علوم آمار، ریاضی، کامپیوتر می باشد از جمله پر طرفدارترین علوم شده است و به همین دلیل حوزه بیشترین درآمد سالانه را دارند. در این کتاب سعی کرده ام از تعاریف و مقدمات رایج، کتابهای کلان داده عبور کرده و به مفاهیم عملیاتی تر در پردازش، استخراج و ذخیره آنها بپردازم. شما در این کتاب به صورت عملی با روشها و ابزارهای پردازش داده های بزرگ آشنا خواهید شد. در چند تمام تلاش نویسندگان در ارائه اطلاعات و مباحث به روز و به بلوغ رسیده بوده است. امید است دلیل رشد بسیار سریع مباحث و ابزارهای این حوزه ممکن است زمانی برسد که تکنولوژی های مطرح شده در این کتاب قدیمی شده باشند. در هر حال هر گونه نظر شما می تواند باعث بالندگی بیشتر و البته کمالها و مطالب مفیدتر در آینده گردد.

راههای ارتباطی با بنده:

aminnezarat@gmail.com و وب سایت www.hpclab.ir می باشد.