

به نام خدا

متن کاوی به کمک بادگیری ماشین

نویسندگان: چا و بی. آگراوال

مترجم:

دکتر مهدی اسماعیلی

متن کاوی به کمک یادگیری ماشین

مترجم: دکتر مهدی اسماعیلی

ناشر: انتشارات آتی‌نگر

ناشر همکار: وینا

صفحه‌آرایی و طراحی: همتا بیداریان

تیراژ: ۵۰۰ نسخه

چاپ اول: ۱۳۹۸

قیمت: ۹۹۰,۰۰۰ ریال

شابک: ۹۷۸-۶۲۲-۶۱۰۲-۵۹-۹

ISBN: 978-622-6102-59-9

حق چاپ برای انتشارات آتی‌نگر محفوظ است.

نشانی دفتر فروش: خیابان جمال‌الدین، روبروی کوچه رشتچی، پلاک ۱۴۴، واحد ۱

نماد: ۶۶۵۶۵۳۳۷

تلفن: ۶۶۵۶۵۳۳۶-۸



www.ati-negar.com * info@ati-negar.com

سرشناسه: آگراوال، چارو سی، ۱۹۷۰-م. Aggarwal, Charu C.

متن کاوی به کمک یادگیری ماشین / نویسنده چارو سی. آگراوال، مترجم: مهدی اسماعیلی

تهران: آتی‌نگر، وینا ۱۳۹۸

۵۹۲ ص: مصور، جدول، نمودار.

ISBN: 978-622-6102-59-9

وضعیت فهرست نویسی: فیا.

یادداشت: عنوان اصلی کتاب: Machine Learning for Text, 2018

یادداشت: کتابنامه.

یادداشت: نمایه.

موضوع: فراگیری ماشینی - Machine learning - متن‌پردازی - Text processing (Computer science)

موضوع: هوش مصنوعی - Artificial intelligence - داده کاوی - Data mining

شناسه افزوده: اسماعیلی، مهدی، ۱۳۵۰-، مترجم

شناسه افزوده: بیداریان، همتا، ۱۳۶۱-، گرافیکست

رده‌بندی کنگره:

رده‌بندی دیویی:

شماره کتابشناسی ملی:

Q۳۲۵/۵

۰۰۶/۳۱

۵۸۱۵۴۴۳

فهرست مطالب

۹	
۱۱	
۱۱	
۱۴	۱-۱ مقدمه.....
۱۶	۱-۲ چه چیزی درباره یادگیری از متن خاص است؟.....
۳۱	۳-۱ مدل‌های تحلیلی برای اسناد متنی.....
۳۱	۴-۱ خلاصه.....
۳۱	۵-۱ کتاب‌شناختی.....
۳۲	۶-۱ تمرین.....
۳۵	
۳۵	۱-۲ مقدمه.....
۳۶	۲-۲ استخراج متن و تبدیل آن به توکن.....
۴۱	۳-۲ استخراج عبارات از توکن‌ها.....
۴۵	۴-۲ نمایش برداری و نرمالسازی.....
۴۷	۵-۲ محاسبه شباهت در متن.....
۵۰	۶-۲ خلاصه.....
۵۱	۷-۲ کتاب‌شناختی.....
۵۱	۸-۲ تمرین‌ها.....
۵۳	
۵۳	۱-۳ مقدمه.....
۵۷	۲-۳ تجزیه مقدار منفرد.....
۶۵	۳-۳ تجزیه نامنفی ماتریس.....
۷۱	۴-۳ تحلیل معنایی نهفته احتمالاتی.....
۷۷	۵-۳ نگاهی اجمالی به تخصیص نهفته دیریکله.....
۸۲	۶-۳ تبدیل‌های غیرخطی و مهندسی ویژگی.....
۹۷	۷-۳ خلاصه.....

۹۸..... ۸-۳ کتاب شناختی

۹۹..... ۹-۳ تمرین ها

۱۰۳..... فصل چهارم: خوشه بندی

۱۰۳..... ۱-۴ مقدمه

۱۰۵..... ۲-۴ انتخاب و مهندسی ویژگی

۱۱۱..... ۳-۴ مدل سازی موضوعی و تجزیه ماتریس

۱۱۶..... ۴-۴ مدل های مولد آمیخته برای خوشه بندی

۱۲۲..... ۵-۴ "k-means" الگوریتم

۱۲۵..... ۶-۴ الگوریتمی نوین برای خوشه بندی سلسله مراتبی

۱۳۲..... ۷-۴ خوشه بندی تاپی

۱۳۴..... ۸-۴ خوشه بندی متا پنگار به خوشه بندی توالی ها

۱۴۱..... ۹-۴ تبدیل خوشه بندی یادگیری ناظر

۱۴۲..... ۱۰-۴ ارزیابی خوشه بندی

۱۴۸..... ۱۱-۴ خلاصه

۱۴۹..... ۱۲-۴ کتاب شناختی

۱۵۰..... ۱۳-۴ تمرین ها

۱۵۳..... فصل پنجم: خوشه بندی و خوشه بندی

۱۵۳..... ۱-۵ مقدمه

۱۵۹..... ۲-۵ انتخاب و مهندسی ویژگی

۱۶۳..... ۳-۵ مدل بیز ساده

۱۷۹..... ۴-۵ رده بند نزدیک ترین همسایه ها

۱۸۸..... ۵-۵ درختان تصمیم و جنگل های تصادفی

۱۹۵..... ۶-۵ رده بندهای مبتنی بر قاعده

۲۰۱..... ۷-۵ خلاصه

۲۰۲..... ۸-۵ کتاب شناختی

۲۰۴..... ۹-۵ تمرین ها

۲۰۷..... فصل ششم: خوشه بندی و خوشه بندی

۲۰۷..... ۱-۶ مقدمه

۲۱۴	۲-۶ رگرسیون و رده‌بندی کمترین مربعات
۲۲۹	۳-۶ ماشین‌های بردار پشتیبان
۲۴۱	۴-۶ رگرسیون لجستیک
۲۴۸	۵-۶ تعمیم‌های غیرخطی از مدل‌های خطی
۲۶۰	۶-۶ خلاصه
۲۶۰	۷-۶ کتاب‌شناختی
۲۶۲	۸-۶ تمرین‌ها

۲۶۵	۱-۷ مقدمه
۲۶۵	۲-۷ موازنه بایاس و واریانس
۲۷۳	۳-۷ اثرات درازنه بایاس و واریانس بر روی کارایی
۲۷۷	۴-۷ بهبود کارایی با استفاده از روش‌های تلفیقی
۲۸۱	۵-۷ ارزیابی رده‌بند
۲۹۵	۶-۷ خلاصه
۲۹۵	۷-۷ کتاب‌شناختی
۲۹۷	۸-۷ تمرین‌ها

۲۹۹	۱-۸ مقدمه
۲۹۹	۲-۸ ترنند تجزیه ماتریس مشترک
۳۰۲	۳-۸ ماشین‌های تجزیه
۳۱۶	۴-۸ تکنیک‌های مدل‌سازی احتمالاتی توأم
۳۲۱	۵-۸ تبدیل به تکنیک‌های گرافکاوی
۳۲۳	۶-۸ خلاصه
۳۲۶	۷-۸ کتاب‌شناختی
۳۲۶	۸-۸ تمرین‌ها

۳۲۹	۱-۹ مقدمه
۳۲۹	۲-۹ شاخص‌بندی و پردازش پرسش

۳۵۵	۳-۹ امتیازدهی با مدل های بازیابی اطلاعات
۳۶۳	۴-۹ خزش وب و کشف منبع
۳۶۹	۵-۹ پردازش پرسش در موتورهای جستجو
۳۷۴	۶-۹ الگوریتم های رتبه بندی مبتنی بر لینک
۳۸۳	۷-۹ خلاصه
۳۸۴	۸-۹ کتاب شناختی
۳۸۶	۹-۹ تمرین ها

۳۸۹	۱-۱۰ مقدمه
۳۹۲	۲-۱۰ مدل های آوری
۳۹۹	۳-۱۰ روش های گزینش
۴۰۰	۴-۱۰ مدل های تجزیه سائری
۴۰۵	۵-۱۰ نمایش فواصل واژه ها با کمک گراف
۴۰۸	۶-۱۰ مدل های عصبی زبان
۴۳۵	۷-۱۰ شبکه های عصبی برگشتی
۴۵۳	۸-۱۰ خلاصه
۴۵۴	۹-۱۰ کتاب شناختی
۴۵۶	۱۰-۱۰ تمرین ها

۴۵۹	۱-۱۱ مقدمه
۴۶۲	۲-۱۱ روش های مبتنی بر واژه های موضوعی برای تلخیص استخراجی
۴۶۸	۳-۱۱ روش های نهفته برای تلخیص استخراجی
۴۷۴	۴-۱۱ یادگیری ماشین برای تلخیص استخراجی
۴۷۹	۵-۱۱ تلخیص چندسندی
۴۷۸	۶-۱۱ تلخیص چکیده های
۴۸۱	۷-۱۱ خلاصه
۴۸۱	۸-۱۱ کتاب شناختی
۴۸۲	۹-۱۱ تمرین ها

۴۸۳.....

۴۸۳..... ۱-۱۲ مقدمه

۴۸۹..... ۲-۱۲ شناسایی موجودیت نامدار

۵۰۶..... ۳-۱۲ استخراج روابط

۵۱۷..... ۴-۱۲ خلاصه

۵۱۸..... ۵-۱۲ کتاب‌شناختی

۵۲۰..... ۶-۱۲ تمرین‌ها

۵۲۳.....

۵۲۳..... ۱-۱۲ مقدمه

۵۲۹..... ۲-۱۳ مبنای اساسات در سطح سند

۵۳۳..... ۳-۱۳ رده‌بندی اساسات در سطح جمله و عبارت

۵۳۶..... ۴-۱۳ نظرکاوی مبتنی بر جملات به عنوان استخراج اطلاعات

۵۴۱..... ۵-۱۳ نظرات هرز

۵۴۵..... ۶-۱۳ خلاصه‌سازی نظرات

۵۴۶..... ۷-۱۳ خلاصه

۵۴۷..... ۸-۱۳ کتاب‌شناختی

۵۴۹..... ۹-۱۳ تمرین‌ها

۵۵۱.....

۵۵۱..... ۱-۱۴ مقدمه

۵۵۲..... ۲-۱۴ تقطیع متن

۵۶۱..... ۳-۱۴ کاوش جریان‌های متنی

۵۶۳..... ۴-۱۴ تشخیص رویداد

۵۷۰..... ۵-۱۴ خلاصه

۵۷۱..... ۶-۱۴ کتاب‌شناختی

۵۷۲..... ۷-۱۴ تمرین‌ها

۵۷۳.....

ناظر روی تو صاحب نظری نیست که نیست
بوی گیسوی تو در هیچ سری نیست که نیست

پیشگفتار مترجم

در سال‌های اخیر به دلیل افزایش متون در محیط‌هایی نظیر وب، رسانه‌های اجتماعی و دیگر پلت‌فرم‌ها، متن‌کاوی از اهمیت ویژه‌ای برخوردار شده است. در تحلیل متن، تکنیک‌هایی از حوزه‌های دیگر نظیر بازیابی اطلاعات، یادگیری ماشین و پردازش زبان طبیعی دیده می‌شود، و برای هر یک از آن‌ها نیز کتاب‌های زیادی نوشته شده است. تمرکز این کتاب بر روی ابزارهای یادگیری ماشین برای اسناد متنی است؛ هر چند در برخی از فصل‌ها رنگ روش‌های حوزه‌های دیگر پررنگتر شده است. به زعم مترجم، کتاب حاضر یکی از ارزنده‌ترین کتاب‌هایی است که در این حوزه به رشته تحریر درآمده است. به همین دلیل از میان کتاب‌های موجود در این حوزه، به انتخاب و ترجمه آن همت گماشتیم.

مطالب کتاب در چهارده فصل گردآوری و تهیه شده است که می‌توان آن را در سه بخش گروه‌بندی کرد. بخش اول یعنی فصل‌های اول تا هشتم کتاب، شامل الگوریتم‌ها و مدل‌ها، پایه برای تحلیل متن است. تجزیه ماتریس، خوشه‌بندی و رده‌بندی، موضوعات اصلی این فصل‌ها را تشکیل می‌دهد. طبیعی است روش‌های ارائه شده در این بخش، برای کار با متن، تطبیق داده شده‌اند. همان‌طور که قبل از این نیز به آن اشاره شد، تحلیل متن ارتباط نزدیکی با حوزه بازیابی اطلاعات دارد. در بخش دوم که تنها شامل فصل‌های دهم تا چهاردهم کتاب، إجمالي از روش‌های بازیابی اطلاعات از دیدگاه متن‌کاوی پرداخته شده است. فصل‌های دهم تا چهاردهم کتاب، بخش سوم کتاب را تشکیل می‌دهند. در این بخش موضوعات پیشرفته‌ای مانند یادگیری عمیق، استخراج اطلاعات، خلاصه‌سازی، نظر‌کاوی، تقطیع متن و تشخیص رویداد بررسی می‌شوند.

تمام تلاش خود را انجام داده‌ایم تا ترجمه کتاب به گونه‌ای انجام شود که خوانندگان محترم آن بتوانند مفاهیم آن را به راحتی درک کنند. بدون شک ممکن است برای برخی از واژه‌های انگلیسی بتوان معادل‌های بهتری یافت. آنچه مسلم است این است که شاید گاهی تبدیل واژه‌ها آنچنان که باید و شاید انجام نشده است؛ اما در ادای جملات و بیان موضوع تلاش فراوان شده است تا خوانندگان گرامی با متنی مبهم و گیج‌کننده روبه‌رو نشوند.

در اینجا لازم می‌دانم از همه اساتید و دانشجویان به خاطر راهنمایی‌های ارزشمندشان در حین آماده‌سازی این کتاب سپاسگزاری کنم. همچنین از مدیریت محترم انتشارات آتی‌نگر و دوست عزیزم جناب رامین مولاناپور نیز به خاطر آماده‌سازی، چاپ و پخش این کتاب تشکر می‌کنم. رهنم محبت بی‌دریغ خانواده‌ام هستم که با فراهم‌سازی محیطی مناسب مرا یاری نمودند. اما با وجود همه سعی و تلاشی که در تمام مراحل آماده‌سازی این کتاب انجام گرفته است، یقین دارم که عاری از اشتباه نیست، چرا که تنها مکتوب بی‌نقص همان معجزه جاوید قرآن کریم است. در آخر ضمن سپاسگزاری از همه کسانی که مرا یاری داده‌اند و با پذیرش مسئولیت هرگونه کاستی احتمالی، امیدوارم که این اندک مفید افتد.

مهدی اسماعیلی

مردادماه ۹۸