

به نام خدا

متن کاوی به کمک
یادگیری ماشین

نویسندگان: چارو سی. آگراوال

مترجم:

دکتر مهدی اسماعیلی

متن کاوی به کمک یادگیری ماشین

مترجم: دکتر مهدی اسماعیلی

ناشر: انتشارات آتی نگر

ناشر همکار: وینا

صفحه آرایبی و طراحی جلد: 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 آتی نگر محفوظ است.

نشانی دفتر فروش: خیابان جمالزاده جنوبی، روبه روی کوچه رشتچی، پلاک ۱۴۴، واحد ۱

نمابر: ۶۶۵۶۵۳۳۷

تلفن: ۸-۶۶۵۶۵۳۳۶



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: آتی نگر، 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 - متن پردازش (Computer science) - Text processing

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

شناسه افزوده: اسماعیلی، مهدی، ۱۳۵۰-، مترجم

شناسه افزوده: بیداریان، همتا، ۱۳۶۱-، گرافیست

رده بندی کنگره:

رده بندی دیویی:

شماره کتابشناسی ملی:

Q۳۲۵/۵

۰۰۶/۳۱

۵۸۱۵۴۴۳

فهرست مطالب

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

۹۸..... کتاب شناختی ۸-۳

۹۹..... تمرین‌ها ۹-۳

فصل چهارم: خوشه‌بندی متن ۱۰۳

۱۰۳..... مقدمه ۱-۴

۱۰۵..... انتخاب و مهندسی ویژگی ۲-۴

۱۱۱..... مدل‌سازی موضوعی و تجزیه ماتریس ۳-۴

۱۱۶..... مدل‌های مولد آمیخته برای خوشه‌بندی ۴-۴

۱۲۲..... الگوریتم k-means ۵-۴

۱۲۵..... الگوریتم‌های خوشه‌بندی سلسله‌مراتبی ۶-۴

۱۳۲..... خوشه‌بندی تلفیقی ۷-۴

۱۳۴..... خوشه‌بندی متن با نگاه به خوشه‌بندی توالی‌ها ۸-۴

۱۴۱..... تبدیل خوشه‌بندی به یادگیری باناظر ۹-۴

۱۴۲..... ارزیابی خوشه‌بندی ۱۰-۴

۱۴۸..... خلاصه ۱۱-۴

۱۴۹..... کتاب‌شناختی ۱۲-۴

۱۵۰..... تمرین‌ها ۱۳-۴

فصل پنجم: رده‌بندی متن: مدل‌های پایه ۱۵۳

۱۵۳..... مقدمه ۱-۵

۱۵۹..... انتخاب و مهندسی ویژگی ۲-۵

۱۶۳..... مدل بیز ساده ۳-۵

۱۷۹..... رده‌بند نزدیک‌ترین همسایه‌ها ۴-۵

۱۸۸..... درختان تصمیم و جنگل‌های تصادفی ۵-۵

۱۹۵..... رده‌بندهای مبتنی بر قاعده ۶-۵

۲۰۱..... خلاصه ۷-۵

۲۰۲..... کتاب‌شناختی ۸-۵

۲۰۴..... تمرین‌ها ۹-۵

فصل ششم: رده‌بندی خطی و رگرسیون برای متن ۲۰۷

۲۰۷..... مقدمه ۱-۶

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

فصل هفتم: کارایی و ارزیابی رده‌بند ۲۶۵

۲۶۵	۱-۷ مقدمه
۲۶۶	۲-۷ موازنه بایاس و واریانس
۲۷۳	۳-۷ اثرات موازنه 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

فصل هشتم: متن کاوی همراه با داده‌های ناهمگن ۲۹۹

۲۹۹	۱-۸ مقدمه
۳۰۲	۲-۸ ترفند تجزیه ماتریس مشترک
۳۱۶	۳-۸ ماشین‌های تجزیه
۳۲۱	۴-۸ تکنیک‌های مدل‌سازی احتمالاتی توأم
۳۲۳	۵-۸ تبدیل به تکنیک‌های گراف‌کاوی
۳۲۶	۶-۸ خلاصه
۳۲۶	۷-۸ کتاب‌شناختی
۳۲۸	۸-۸ تمرین‌ها

فصل نهم: بازیابی اطلاعات و موتورهای جستجو ۳۲۹

۳۲۹	۱-۹ مقدمه
۳۳۰	۲-۹ شاخص‌بندی و پردازش پرسش

۳۵۵	۳-۹ امتیازدهی با مدل‌های بازیابی اطلاعات
۳۶۳	۴-۹ خزش وب و کشف منبع
۳۶۹	۵-۹ پردازش پرسش در موتورهای جستجو
۳۷۴	۶-۹ الگوریتم‌های رتبه‌بندی مبتنی بر لینک
۳۸۳	۷-۹ خلاصه
۳۸۴	۸-۹ کتاب‌شناختی
۳۸۶	۹-۹ تمرین‌ها

فصل دهم: مدل‌سازی توالی متن و یادگیری عمیق ۳۸۹

۳۸۹	۱-۱۰ مقدمه
۳۹۲	۲-۱۰ مدل‌های آماری زبان
۳۹۹	۳-۱۰ روش‌های کرنل
۴۰۰	۴-۱۰ مدل‌های تجزیه ماتریس
۴۰۵	۵-۱۰ نمایش فواصل واژه‌ها با کمک گراف
۴۰۸	۶-۱۰ مدل‌های عصبی زبان
۴۳۵	۷-۱۰ شبکه‌های عصبی برگشتی
۴۵۳	۸-۱۰ خلاصه
۴۵۴	۹-۱۰ کتاب‌شناختی
۴۵۶	۱۰-۱۰ تمرین‌ها

فصل یازدهم: تلخیص متن ۴۵۹

۴۵۹	۱-۱۱ مقدمه
۴۶۲	۲-۱۱ روش‌های مبتنی بر واژه‌های موضوعی برای تلخیص استخراجی
۴۶۸	۳-۱۱ روش‌های نهفته برای تلخیص استخراجی
۴۷۴	۴-۱۱ یادگیری ماشین برای تلخیص استخراجی
۴۷۹	۵-۱۱ تلخیص چندسندی
۴۷۸	۶-۱۱ تلخیص چکیده‌های
۴۸۱	۷-۱۱ خلاصه
۴۸۱	۸-۱۱ کتاب‌شناختی
۴۸۲	۹-۱۱ تمرین‌ها

فصل دوازدهم: استخراج اطلاعات 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

فصل سیزدهم: نظر کاوی و تحلیل احساسات ۵۲۳

۵۲۳	۱-۱۳ مقدمه
۵۲۹	۲-۱۳ رده‌بندی احساسات در سطح سند
۵۳۳	۳-۱۳ رده‌بندی احساسات در سطح جمله و عبارت
۵۳۶	۴-۱۳ نظر کاوی مبتنی بر جنبه 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

فصل چهاردهم: تقطیع متن و تشخیص رویداد 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

منابع 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 الگوریتم‌های 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 مدل‌های 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 فصل نهم کتاب است، به مروری اجمالی 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. رهین 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