متن کاوی:
طبقه‌بندی، خوشه‌بندی و کاربردها

Text Mining:
Classification, Clustering and Applications

www.ketab.ir

تالیف:

سینا دامی

(استادیار گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد تهران غرب)

فرید سروری

(گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد اردبیل)

عباس میرزایی

(استادیار گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد اردبیل)

انتشارات جهاد دانشگاهی استان اردبیل

۱۴۰۰

سرشناسه	: دلمی، سینا، ۱۳۶۵ -
عنوان و نام پدیدآور	: متن کاوی: طبقه‌بندی، خوشه‌بندی و کاربردها =Text mining: classification,clustering and applications/تالیف سینا دلمی، فرید سروری، عباس میرزایی
مشخصات نشر	: اردبیل: جهاد دانشگاهی، واحد استان اردبیل، سازمان انتشارات، ۱۴۰۰
مشخصات ظاهری	: ۳۳۶ ص:
شابک	: ۴۵۰۰۰۰ ریال: ۷-۸۱-۶۲۳۱-۶۲۲-۹۷۸
وضعیت فهرست نویسی	: فیپا
موضوع	: داده‌کاوی
موضوع	: Data mining
موضوع	: تجزیه و تحلیل کلاستر
موضوع	: Cluster analysis
موضوع	: تحلیل ممیز
موضوع	: Discriminant analysis
موضوع	: داده‌کاوی —روش‌های آماری
موضوع	: Data mining – Statistical methods
شناسه افزوده	: سروری، فرید، ۱۳۵۸-
شناسه افزوده	: میرزایی، عباس، ۱۳۶۱ -
شناسه افزوده	: جهاد دانشگاهی. واحد اردبیل
شناسه افزوده	: Jahade Daneshgahi. Vahede Ardabil.
رده‌بندی کنگره	: QA ۷۶/۹
رده‌بندی دیویی	: ۰۰۶/۳۱۲
شماره کتابشناسی ملی	: ۸۴۲۳۵۷۰

www.ketab.ir

متن کاوی (طبقه‌بندی، خوشه‌بندی و کاربردها)

تالیف: سینا دلمی، فرید سروی و عباس میرزایی

صفحه آرایی: حسن قربانژاد

طرح جلد: محمود خروشی

ناشر: انتشارات جهاد دانشگاهی استان اردبیل

چاپ: اول- ۱۴۰۰

شمارگان: ۱۰۰۰ نسخه

قیمت: ۴۵۰۰۰ تومان

شابک: ۹۷۸-۶۲۲-۶۲۳۱-۸۱-۷

واحد استان اردبیل

این اثر، مشمول قانون پشتیبانی مؤلفان و مصنفان و هنرمندان مصوب ۱۳۴۸ است، هر کسی تمام یا قسمتی از این اثر را بدون اجازه مؤلف (ناشر) نشر یا پخش یا عرضه نماید مورد پیگرد قانونی قرار خواهد گرفت.

نشانی انتشارات اردبیل – شهرک کارشناسان – میدان شفا، تلفن: ۳۳۷۴۹۵۰۶، رایانامه ard.chap@chmail.ir

نشانی سازمان انتشارات تهران- خیابان انقلاب اسلامی- خیابان فخررازی- خیابان شهدای ژاندارمری- پلاک ۷۲، تلفن: ۶۶۹۵۲۹۴۸

فهرست مطالب

عنوان | صفحه

پیشگفتار ۱۳

۱- مقدمه‌ای بر متن کاوی ۱۷

۱-۱ مقدمه ۱۸

۱-۲الگوریتمها و کاربردهای متن کاوی ۲۰

۱-۳جهت‌دهی‌های آینده ۲۴

۲- استخراج اطلاعات متن ۲۷

۲-۱مقدمه ۲۸

۲-۲شناسایی موجودیت نام‌دار ۳۲

۲-۲-۱ رویکرد مبتنی بر قاعده ۳۳

۲-۲-۲ رویکرد یادگیری آماری ۳۴

۲-۳ استخراج رابطه ۳۹

۲-۳-۱ طبقه‌بندی مبتنی بر ویژگی ۴۰

۲-۳-۲روش‌های کرنل ۴۴

۲-۳-۳روش‌هایادگیریبا نظارت ضعیف ۴۷

۲-۴استخراج اطلاعات بدون نظارت ۴۸

۲-۴-۱ کشف رابطه و القاء الگو ۴۹

۲-۴-۲ استخراج اطلاعات آزاد ۵۰

۲-۵ارزیابی ۵۱

۲-۶ خلاصه فصل ۵۲

۳- خلاصه‌سازی متن ۵۴

۳-۱خلاصه‌سازی استخراجی ۵۵

۳-۲ رویکردهای بازنمایی موضوع ۵۷

۳-۲-۱ کلمات موضوع ۵۷

www.ketab.ir

۵-۶-۳ طبقه‌بندی شبکه عصبی ۱۶۳
۵-۶-۴ مشاهداتی درباره طبقه‌بندی خطی ۱۶۶
۵-۷ طبقه‌بندی مبتنی بر مجاورت ۱۶۷
۵-۸ طبقه‌بندی متن وب و داده‌های پیوندی ۱۷۰
۵-۹ فرا الگوریتم‌های طبقه‌بندی متن ۱۷۶
۵-۹-۱ طبقه‌بندی گروهی: یادگیری جمعی ۱۷۶
۵-۹-۲ طبقه‌بندی داده‌محور: بوستینگ و بگینگ ۱۷۷
۵-۹-۳ طبقه‌بندی دقت‌محور: بهینه‌سازی سنجش ۱۷۹
۵-۱۰ خلاصه فصل ۱۸۰

۶- متن کاوی در داده‌های جریانی ۱۸۲
۶-۱ مقدمه ۱۸۳
۶-۲ خوشه‌بندی جریان‌های متنی ۱۸۴
۶-۲-۱ شناسایی و ردیابی موضوع در جریان‌های متنی ۱۹۲
۶-۳ طبقه‌بندی جریان‌های متنی ۱۹۷
۶-۴ تحلیل تکامل در جریان‌های متنی ۲۰۰
۶-۵ خلاصه فصل ۲۰۲

۷- متن کاوی در داده‌های فرازبانی ۲۰۳
۷-۱مقدمه ۲۰۴
۷-۲ترجمه ماشینی ۲۰۵
۷-۲-۱ مدل‌های مولد ترجمه و SMT ۲۰۵
۷-۲-۲ مدل‌های مبتنی بر کلمه ۲۰۸
۷-۲-۳ مدل‌های مبتنی بر عبارت ۲۱۰
۷-۲-۴مدل‌های مبتنی بر نحو ۲۱۴
۷-۳ کاوش متون موازی ۲۱۷
۷-۳-۱ کاوش ساختار وب ۲۱۹
۷-۳-۲ تطابق صفحات موازی ۲۲۱
۷-۴ مدل‌های ترجمه در CLIR ۲۲۴
۷-۵ جمع‌آوری و بهره‌برداری از متون تطبیقی ۲۲۶

www.ketab.ir

۶-۷ انتخاب کلمات ترجمه، عبارات و جملات موازی ۲۳۰
۷-۷ کاوش روابط فرازبانی از متون تک‌زبانی ۲۳۲
۸-۷کاوش فراپیوندها ۲۳۵
۹-۷ خلاصه فصل ۲۳۶

۸- متن کاوی در داده‌های چندرسانه‌ای ۲۳۹
۱-۸مقدمه ۲۴۰
۲-۸کاوش متن پیرامون ۲۴۲
۳-۸ برچسب کاوی ۲۴۵
۱-۳-۸ رتبه‌بندی برچسب ۲۴۵
۲-۳-۸پالایش برچسب ۲۴۷
۳-۳-۸ غنی‌سازی برچسب ۲۴۸
۴-۸ کاوش محتوای مشترک متنی و بصری ۲۵۰
۱-۴-۸ رتبه‌بندی بصری ۲۵۱
۵-۸ کاوش محتوای متقابل متنی و بصری ۲۵۴
۶-۸ خلاصه فصل ۲۵۷

۹- متن کاوی در رسانه‌های اجتماعی ۲۶۰
۱-۹ مقدمه ۲۶۱
۲-۹ جنبه‌های مجزای متن در رسانه‌های اجتماعی ۲۶۲
۱-۲-۹ چارچوب عمومی برای تحلیل متن ۲۶۳
۲-۲-۹حساسیت به زمان ۲۶۵
۳-۲-۹ طول کوتاه ۲۶۶
۴-۲-۹ عبارات غیرساخت‌یافته ۲۶۷
۵-۲-۹ اطلاعات اضافی ۲۶۸
۳-۹ تحلیل متن در رسانه‌های اجتماعی ۲۶۸
۱-۳-۹ شناسایی رویداد ۲۶۹
۲-۳-۹ پاسخ‌گویی به سوال مشارکتی ۲۷۱
۳-۳-۹ برچسب‌گذاری اجتماعی ۲۷۲

www.ketab.ir

۹-۳-۴ پرکردن شکاف معنایی ۲۷۳
۹-۳-۵ بهره‌برداری از قدرت اطلاعات اضافی ۲۷۴
۹-۳-۶ تلاش‌های مرتبط ۲۷۷
۹-۴ مثالدنیای واقعی ۲۷۷
۹-۴-۱ استخراج عبارات اولیه ۲۷۸
۹-۴-۲ استخراج ویژگی‌های معنایی ۲۸۰
۹-۴-۳ساخت فضای ویژگی ۲۸۰
۹-۵ خلاصه فصل ۲۸۲

۱۰- عقیده کاوی و تحلیل احساسات ۲۸۵
۱۰-۱ مقدمه ۲۸۶
۱۰-۲ طبقه‌بندی احساسات ۲۹۰
۱۰-۲-۱ طبقه‌بندی احساسات مبتنی بر یادگیریبا نظارت ۲۹۲
۱۰-۲-۲ طبقه‌بندی احساسات مبتنی بر یادگیریبدون نظارت ۲۹۵
۱۰-۳ طبقه‌بندی احساسات و دهیت جمله ۲۹۸
۱۰-۴ گسترش واژه‌نامه عقیده ۲۹۹
۱۰-۴-۱ رویکرد مبتنی بر فرهنگ لغت ۳۰۰
۱۰-۴-۲ رویکرد مبتنی بر پیکرهمتنی ۳۰۰
۱۰-۵ تحلیل احساسات مبتنی بر جنبه ۳۰۲
۱۰-۵-۱ طبقه‌بندی جنبه احساسات ۳۰۳
۱۰-۵-۲ قواعد اساسی عقاید ۳۰۵
۱۰-۵-۳ استخراج جنبه ۳۰۸
۱۰-۵-۴ همزمانی گسترش فرهنگ لغت عقیده و استخراج جنبه ۳۱۰
۱۰-۶ عقیده کاوی تطبیقی ۳۱۱
۱۰-۷ برخی مسایل عقیده کاوی ۳۱۵
۱۰-۸ سودمندی عقاید ۳۱۹
۱۰-۹خلاصه فصل ۳۲۰

۱۱- متن کاوی معنایی مبتنی بر آنتولوژی ۳۲۲
۱۱-۱ مقدمه ۳۲۳

www.ketab.ir

۲-۱۱وب معنایی و آنتولوژی ۳۲۳
۳-۱۱ خوشه‌بندی معنایی متن ۳۲۶
۱-۳-۱۱ تشابه معنایی ۳۲۷
۴-۱۱طبقه‌بندی معنایی متن ۳۲۹
۵-۱۱ خلاصه فصل ۳۳۱
منابع ۳۳۲

www.ketab.ir

پیشگفتار

متن کاوی در سال‌های اخیر به دلیلحجم زیادی از داده‌های متنی که در شبکه‌های اجتماعی مختلف، وبسایت‌ها و سایر برنامه‌های کاربردی اطلاعات محور ایجادشده، بیشتر مورد توجه قرار گرفته است. داده‌هایغیرساخت‌یافته ساده‌ترین شکل داده‌هایی هستند که می‌توانند در هر سناریوییممکن ایجاد شوند. در نتیجه،یک نیازمبرم به طراحی روش‌هاوالگوریتم‌هاییاحساس می‌شود که بتوانند طیف گسترده‌ای از برنامه‌های کاربردی مبتنی بر متن را پردازش کنند. این کتاب مروری بر روش‌هاوالگوریتم‌های مختلف و متداول در حوزه متن کاوی با تمرکز ویژه بر روش‌های طبقه‌بندی، خوشه‌بندی و کاربردهای آن خواهد داشت.

در ابتدا ضمن بیان مقدمه‌ای بر متن کاوی در فصل اول، به مفاهیم استخراج اطلاعات از متن و کارهایی که در جامعه پردازش زبان طبیعی انجام شده است ، در فصل دوم می‌پردازیم. استخراج اطلاعات،کشفاطلاعات ساخت‌یافته از متن غیرساخت‌یافتهیا نیمه‌ساخت‌یافته است. این یک کار مهم در استخراج متن است و به طور گسترده در مجامع مختلف پژوهش‌هایی از جمله پردازش زبان طبیعی، بازیابی اطلاعات و استخراج وب مورد مطالعه قرار گرفتهاست. دو وظیفه اساسی استخراج اطلاعات به عنوان شناساییموجودیت‌های اسمی و استخراج روابط معنایی بین آن‌هت استکه در این فصل بدان پرداخته می‌شود.

وظیفه مهم دیگر، استخراج بخش‌های مهم از یک یا چند متن است که خلاصه‌سازی متن نام دارد. نوعی از خلاصه‌سازی متن را می‌توان به عنوان یکی از اهداف اصلی استخراج اطلاعات دانست.با توجه به این‌که اسناد زیادی در وب موجود است که بیشتر آن‌ها محتوای غیرضروری دارند، اهمیت خلاصه‌سازی متن به‌منظور کاهش زمان پردازش و مطالعه محسوس‌تر می‌شود.فصل سوم، به خلاصه‌سازی خودکار متن می‌پردازد که یکی از وظایف مهم و کاربردی در زمینه متن کاوی است. هدف از خلاصه‌سازی متن، یافتن خلاصه‌ای از محتویات متن به صورت فشرده و مطابق با اطلاعات موردنیاز کاربر استبه‌گونه‌ای که مفاهیم کلی و محتوای اطلاعاتی متن حفظ شود.تا کنونروش‌های بی‌شماری برای شناسایی محتوای مهم در خلاصه‌سازی متن توسعه یافته است. در این فصل، مروری کلیبررویروش‌های موجودخلاصه‌سازی متنارائه می‌دهیم. همچنین به برخی از مسائل و راه‌حل‌های پیشنهادی که رویکردهاییادگیری ماشین را برایاین وظیفه به چالش کشیدهاست، اشاره می‌کنیم.

در فصل چهارم و فصل پنجم، بترتیب به دو دسته‌بندی مهم متن کاوی، یعنی خوشه‌بندی متن و طبقه‌بندیمتن می‌پردازیم. ما در فصل چهارم، یک بررسیجامع از مساله‌ی خوشه‌بندی متن ارائه می‌کنیم. انواع روش‌ها، مزایای نسبی و چالش‌های کلیدیمساله‌ی خوشه‌بندی برروی داده‌های متنیمورد مطالعه و بحث قرار می‌گیرند. در فصل پنجم، مروری جامع بر طیف گسترده‌ای از الگوریتم‌های طبقه‌بندی متن خواهیم داشت و به کاربردهای بی‌شمارینظیرطبقه‌بندی پرونده‌های

www.ketab.ir

سلامت بیماران، نامه‌های الکترونیکی، مقالات علمی، متون خبری،صفحات وب و شبکه‌های اجتماعی اشاره می‌شود.

بخش اعظمی از داده‌های متنی که دائما در طی زمان با کاربردهایی همچون شبکه‌های اجتماعی تولید می‌شوند، منجر به جریان‌های متنیکلان می‌گردد.عمدتا این جریان‌های کلان متن از طریق تعاملات فردی در مقیاس عظیم یا از طریق تکوین ساختارمند انواع محتوای خاص توسط سازمان‌های اختصاصی ایجاد می‌شوند. بارزترین نمونه، جریان‌های متنیاخباراست که توسط سرویس‌های خبری تولید می‌شود. چنین جریان‌هایی،چالش‌های بی‌سابقه‌ای را برای الگوریتم‌های داده کاوی مطرح کرده است. ما در فصل ششم، الگوریتم‌های متن کاوی در داده‌هایجریانیرا برای انواع مسائل موجود در داده کاوی از قبیل خوشه‌بندی، طبقه‌بندی وتحلیل تکامل بررسی می‌کنیم.

بازیابی اطلاعات از میان چندین زبان مختلف مانند ترجمه متن کامل، عملی است که نیازمند نوعی انتقال دانش میان زبان‌ها می‌باشد. برای ساختن ابزارهای ترجمه باکیفیت وجود منابع ترجمه‌ی قوی امری حیاتی است. این در حالیست که منابع دست‌ساز دارای محدودیت‌هایی از قبیل شمول و سازگاری با طیف گسترده‌ای از کاربردها می‌باشند. متن کاوی در داده‌های فرازبانی ما را قادر می‌سازد تا این منابع دستی را تکمیل کنیم. فصل هفتم،به توصیف مباحثیمی‌پردازد که در جست‌وجوی استخراج دانش فرازبانی از داده‌های متنی و به‌خصوصداده‌های موجود در وب هستند.

همان‌طور که می‌دانیم حجم بسیاری از داده‌هایچندرسانه‌اینظیر متن، تصویر و ویدیودروب دردسترس است. معمولاً این رسانه‌ها با ترکیبی از انواع فرادادهها نظیر متن پیرامون، برچسب‌های کاربر، امتیازها، نظرات و غیره همراه می‌شود. کاوش این فرادادههای متنی در تسهیل پردازش و مدیریت اطلاعات چندرسانه‌ای موثر است. فصل هشتم ارزیابی جامعی از متن کاوی در داده‌هایچندرسانه‌ای فراهم می‌کند.

رشد سریع رسانه‌های اجتماعی آنلاین در قالب تولید محتوای مشارکتی، فرصت‌ها و چالش‌های جدیدی را برای تولیدکنندگان و مصرف‌کنندگان اطلاعات ایجاد می‌کند. با وجود تعداد زیادی از داده‌های تولید شده توسط سرویس‌های مختلف رسانه‌های اجتماعی، تحلیل متن، یک روش مؤثر برای پاسخگویی به نیازهای متنوع اطلاعاتی را فراهم می‌کند. در فصل نهم، ضمن معرفی پیشینه‌ی تحلیل متن و جنبه‌های متمایز داده‌های متنی در رسانه‌های اجتماعی، در مورد پیشرفت نتایج استفاده از متن کاوی در رسانه‌های اجتماعی از دیدگاه‌های مختلف بحث می‌کنیم.

یکی دیگر از جنبه‌های مهم رشد انفجاری رسانه‌های اجتماعی در وب،استفاده فزاینده‌یافراد یاسازمان‌ها از نظرات و عقاید عمومی برگرفتهاز این رسانه‌هابرای تصمیم‌گیریاست. کاوش متون مرتبط با عقاید و افکار مردم در وب به دلیل گسترش رسانه‌های متنوع، به عنوان یکی از کاربردهای

www.ketab.ir

مهم متن کاوی مطرح است. هر وب‌سایت به طور معمول شامل حجم عظیمی از متن نظرات است که همیشه در پست‌ها و وبلاگ‌های انجمن‌هایمرتبط رمزگشایی نمی‌شود. معمولاً یک خواننده‌ی انسانی در شناسایی وب‌سایت‌های مربوطه و خلاصه‌سازی دقیق اطلاعات و نظرات موجود در آن‌ها با مشکل روبه‌رو خواهد شد. علاوه بر این، همواره تحلیل انسانی از اطلاعات متنی در معرض تعصبات قابل‌توجهی قرار می‌گیرد؛ به عنوان مثال، افراد اغلب به عقایدیکه خودشان ترجیحمی‌دهند تا با چیزی سازگارتر باشد، بیشتر توجه می‌کنند.آن‌ها همچنین به دلیل محدودیت‌های ذهنی و جسمی خود در پردازش حجم زیادی از اطلاعات با مشکل روبه‌رومی‌شوند که استفاده از سیستم‌هایکاوش عقایدو خلاصه‌سازینظرات را جهت تحلیل احساساتبرای آن‌ها ملزم می‌کند.ازاین‌رو، فصل دهم را به مسائل مربوط به عقیده کاوی و تحلیل احساساتاختصاص دادیم.

در نهایت، پس از پرداختن به مسایل مربوط به متن کاوی در وب یک‌طرفه (وب‌سایت‌ها) و وب دوطرفه (رسانه‌های اجتماعی)، به استقبالمهم‌ترین مساله‌ی متن کاوی در نسل بعدی وب یعنی وب معناییمی‌رویم. هدف از وب معنایی این است که اطلاعات در وب نه تنها برای انسان‌ها قابل فهم باشد، بلکه ماشین‌ها نیز توانایی درک و پردازش آن‌ها را داشته باشند.یکی از مفاهیم اساسی وب معنایی، آنتولوژی است که در فصل یازدهم بعد از آشنایی با مفاهیم وب معنایی و آنتولوژی، مسائل مربوط بهمتن کاوی معنایی مبتنی بر آنتولوژی مورد بررسی قرار می‌گیرد.

بدیهی است که هر اثری کامل و بدون نقص نیست از این‌رو، از خوانندگان علاقمند خواهشمندیم که نظرات و بازخورد سازنده‌ی خود را با ما از طریق پست الکترونیکیdami@wtiau.ac.irبه اشتراک بگذارند و در آینده نیز ما را همراهی کنند.در ادامه از تمامی دوستان و همکاران عزیزی که همراه و یاری‌گرمان بودند، صمیمانه سپاسگزاریم.

www.ketab.ir

سینا دامی

بهار ۱۴۰۰