داده‌کاوی رسانه‌های اجتماعی

«توییتر، وبلاگ، ویکی‌پدیا»

مؤلفین

گابو زابو
گانگور پولاتکان
اسکار بویکین
آنتونیو چاکیوپولو

مترجم

ایوب ترکیان

نیاز دانش

عنوان و نام پدیدآور	:	**داده‌کاوی رسانه‌های اجتماعی: (توییتر، وبلاگ، ویکی‌پدیا)**/مولفین گابو زابو...[و دیگران]؛ مترجم ایوب ترکیان.
مشخصات نشر	:	تهران: نیاز دانش، ۱۳۹۷.
مشخصات ظاهری	:	۳۶۰ ص.
شابک	:	978-600-8906-40-7
وضعیت فهرست‌نویسی	:	فیپا
یادداشت	:	عنوان اصلی:.Social Media Data Mining and Analytics,2019
یادداشت	:	مولفین گابو زابو، گانگور پولاتکان، اسکار بویکین، آنتونیو چاکیوپولو.
موضوع	:	برنامه‌ریزی سازمانی -- داده‌پردازی — Business planning -- Data processing
موضوع	:	مصرف‌کنندگان -- رفتار-- آینده‌نگری — Consumer behavior-- Forecasting
موضوع	:	داده‌کاوی — Data mining
موضوع	:	مصرف‌کننده‌شناسی -- داده‌پردازی — Consumer profiling -- Data processing
موضوع	:	رسانه‌های اجتماعی -- داده‌پردازی — Social media -- Data processing
شناسه افزوده	:	سابو، گابور — Szabó, Gábor
شناسه افزوده	:	ترکیان، ایوب، ۱۳۳۷، مترجم
رده‌بندی کنگره	:	HD۳۰/۲۱۳/د۲ ۱۳۹۷
رده‌بندی دیویی	:	۶۵۸/۴۰۳۸۰۱۱
شماره کتابشناسی ملی	:	۵۵۳۰۵۱۲

نیاز دانش

نام کتاب	:	**داده‌کاوی رسانه‌های اجتماعی «توییتر، وبلاگ، ویکی‌پدیا»**
مؤلفین	:	گابو زابو - گانگور پولاتکان - اسکار بویکین - آنتونیو چاکیوپولو
مترجم	:	ایوب ترکیان
مدیر اجرایی ـ ناظر بر چاپ	:	حمیدرضا ا محمد شیرازی - محمد شمس
ناشر	:	نیاز دانش
صفحه‌آرا	:	واحد تولید انتشارات نیازدانش
نوبت چاپ	:	اول - ۱۳۹۷
شمارگان	:	۱۰۰ نسخه
قیمت	:	۵۰۰۰۰۰ ریال

شابک: ۷-۴۰-۸۹۰۶-۶۰۰-۹۷۸ **ISBN:978-600-8906-40-7**

آدرس انتشارات: تهران، میدان انقلاب، خیابان ۱۲ فروردین، تقاطع وحید نظری، پلاک ۲۵۵، طبقه ۱، واحد ۲
۰۹۱۲۷۰۷۳۹۳۵-۶۶۴۷۸۱۰۸-۶۶۴۷۸۱۰۶-۰۲۱

www.Niaze-Danesh.com

مشاوره جهت نشر: ۲۱۰۶۷۰۹ – ۰۹۱۲

فهرست مطالب

عنوان ---------- **شماره صفحه**

فصل ۱ / کاربران: چه‌کسی رسانه‌های اجتماعی ---------- ۲۹

سنجش تغییرپذیری رفتار کاربر در ویکی‌پدیا ---------- ۲۹
تنوع فعالیت‌های کاربر ---------- ۳۰
منشأ توزیع فعالیت کاربر ---------- ۴۱
تبعات قاعده توان ---------- ۵۰
دنباله دراز در فعالیت‌های انسانی ---------- ۵۶
دنباله دراز در همه جا: قاعده ۸۰/۲۰ (قاعده p/q) ---------- ۵۹
رفتار برخط روی توییتر ---------- ۶۴
بازیابی توییت‌ها برای کاربران ---------- ۶۵
سبدبندی لگاریتمی ---------- ۶۷
فعالیت‌های کاربر در توییتر ---------- ۶۹
خلاصه ---------- ۷۱

فصل ۲ / شبکه‌ها: چگونگی رسانه‌های اجتماعی ---------- ۷۳

انواع و ویژگی‌های شبکه‌های اجتماعی ---------- ۷۵
شبکه‌های صریح ---------- ۷۵
گراف‌های جهت‌دار و بدون‌جهت ---------- ۷۷
ویژگی‌های گره و ضلع ---------- ۷۸
گراف‌های وزندار ---------- ۷۹
شبکه‌های غیرصریح ---------- ۸۱
مصورسازی شبکه‌ها ---------- ۸۴
درجه شبکه ---------- ۸۸
شمارش تعداد ارتباط ---------- ۹۰
دنباله دراز در ارتباطات کاربر ---------- ۹۲
ورای مدل شبکه ایدهال ---------- ۹۶
صید هم‌بستگی‌ها ---------- ۹۸
مثلث‌های محلی و خوشه‌بندی ---------- ۹۹

هم‌سنخی ۱۰۶
خلاصه ۱۱۱

فصل ۳ / فرایندهای زمانی: چه‌موقعی رسانه‌های اجتماعی ۱۱۳

مدل‌های سنتی و رویدادهای زمانی ۱۱۳
وقوع همگن رویداد در زمان ۱۱۵
زمان‌های بین‌رویدادی ۱۱۸
مقایسه با فرایند فاقدحافظه ۱۲۳
خودهم‌بستگی ۱۲۶
انحراف از بی‌حافظه‌گی ۱۲۹
تناوب زمانی در فعالیت کاربر ۱۳۱
فعالیت‌های فورانی افراد ۱۳۷
هم‌بستگی‌ها و فوران‌ها ۱۴۴
نمونه‌برداری مخزنی ۱۴۵
سنجه‌های پیش‌یابی در زمان ۱۴۹
روندیابی ۱۵۲
یافتن فصل‌گرایی ۱۵۶
پیش‌یابی سری زمانی با ARIMA ۱۵۸
بخش خودرگرسیو (“AR”) ۱۵۸
بخش میانگین متحرک (“MA”) ۱۵۹
مدل ARIMA کامل (p, d, q) ۱۶۰
خلاصه ۱۶۳

فصل ۴ / محتوا: چیستی رسانه‌های اجتماعی ۱۶۵

تعریف محتوا: تمرکز بر متن و داده‌های بدون ساختار ۱۶۵
تولید ویژگی از متن ۱۶۷
آمارهای پایه وقوع ترم در متن ۱۷۰
استفاده از ویژگی‌های محتوا برای شناسایی موضوعات ۱۷۱
مورد اقبال بودن موضوعات ۱۸۲
تنوع علایق کاربر انفرادی ۱۸۵
استخراج اطلاعات بُعد کم از متن بُعد زیاد ۱۸۸
مدل‌سازی موضوعی ۱۹۰
مدل‌سازی موضوعی بدون سرپرست ۱۹۲
استنباط ۱۹۳
نمایش تجربی ۱۹۵
مدل‌سازی موضوعی با سرپرست ۲۰۱

استنباط ۲۰۳
نمایش تجربی ۲۰۴
مدل‌سازی موضوعی رابطه‌ای ۲۰۸
استنباط ۲۱۱
نمایش تجربی ۲۱۲
خلاصه ۲۱۶

فصل ۵ / پردازش مجموعه داده‌های بزرگ ۲۱۷

MapReduce: ساختار دادن به عملیات موازی و سری ۲۱۷
شمارش واژه‌ها ۲۱۹
چولگی: نفرین آخرین کاهش‌گر ۲۲۳
جریان‌های MapReduce چندمرحله‌ای ۲۲۵
پخشینه‌گی ۲۲۷
تلفیق جریان‌های داده‌ها ۲۲۸
وصل کردن دو منبع داده ۲۲۸
وصل کردن در مقابل مجموعه داده‌های کوچک ۲۳۳
مدل‌های MapReduce مقیاس بزرگ ۲۳۴
الگوهای برنامه‌سازی MapReduce ۲۳۵
برنامه‌های MapReduce استاتیک ۲۳۶
برنامه‌های MapReduce تکرارشونده ۲۴۴
درجه‌صفحه برای درجه‌بندی در گراف‌ها ۲۴۴
خوشه‌بندی کامینز ۲۴۹
برنامه‌های MapReduce افزایشی ۲۵۳
برنامه‌های MapReduce زمانی ۲۵۳
انباشتن و مکعب‌سازی داده‌ها ۲۵۵
بسط برنامه‌های انباشتن ۲۶۲
چالش‌های پردازش داده‌های دنباله دراز ۲۶۳
نمونه‌برداری و تخمین ۲۶۵
HyperLogLog ۲۶۹
مثال HyperLogLog ۲۷۲
HyperLogLog روی داده‌های Stack Exchange ۲۷۳
عملکرد HLL روی مِه‌داده‌ها ۲۷۴
فیلترهای بلوم ۲۷۵
مثال فیلتر بلوم ۲۷۹
فیلتر بلوم به‌عنوان دانش عضویت پیش‌محاسبه شده ۲۸۱
بلوم روی مجموعه داده‌های اجتماعی بزرگ ۲۸۲
ترسیم Count-Min ۲۸۴

ساختار Count-Min– مثال Hitterهای سنگین ۲۸۷
ساختار Count-Min– مثال درصد بالا ۲۸۸
انباشت ساختارهای داده تخمینی ۲۸۸
خلاصه تخمین‌ها ۲۹۰
خلاصه ۲۹۰

فصل ۶ / یادگیری، نگاشت، و توصیه ۲۹۳

سرویس‌های رسانه‌های اجتماعی برخط ۲۹۴
موتورهای جستجو ۲۹۴
درگیرشدن با محتوا ۲۹۵
تعامل با دنیای واقعی ۲۹۷
تعاملات با افراد ۲۹۷
فرمولاسیون مسئله ۲۹۹
یادگیری و نگاشت ۳۰۱
فاکتورگیری ماتریس ۳۰۴
یادگیری، آموزش ۳۰۶
فرو– و فرابرازش ۳۰۶
تنظیم در فاکتورگیری ماتریس ۳۰۹
فاکتورگیری ماتریس غیرمنفی و تنکی ۳۰۹
نمایش روی درجه‌بندی فیلم ۳۱۰
تفسیر کلیشه‌های یادگیری شده ۳۱۴
تحلیل کاوشی ۳۱۸
پیش‌بینی و توصیه ۳۲۳
ارزیابی ۳۲۶
مروری بر متدولوژی‌ها ۳۲۸
رویکردهای نزدیکترین همسایه ۳۲۸
رویکردهای مبتنی بر یادگیری با سرپرست ۳۲۹
پیش‌بینی درجه‌بندی فیلم با رگرسیون لجیستیک ۳۳۰
سوژه‌های رایج با ویژگی‌ها ۳۳۹
کاربردهای زمینه خاص ۳۴۰
خلاصه ۳۴۱

فصل ۷ / نتیجه‌گیری ۳۴۳

پایداری عجیب الگوی تعامل انسانی ۳۴۳
میانگین‌ها، انحراف معیارها، و نمونه‌برداری ۳۴۶
حذف داده‌های پرت ۳۵۴

مقدمه

این کتاب در مورد استفاده از داده‌ها برای شناخت نحوه استفاده سرویس‌های رسانه‌های اجتماعی است. از زمان توسعه وب ۲، سایت‌ها و خدماتی که به کاربران خود توان تغییر فعال و سهیم‌شدن در محتوای خدمات را می‌دهند، بسیار مورد اقبال قرار گرفته است. منشأ رسانه‌های اجتماعی در شبکه‌های اجتماعی و خدمات ارتباطی گروهی اولیه، شامل سامانه‌های بولتن‌بورد (BBS) دهه ۱۹۸۰، گروه‌های خبری Usenet، و Geocities در دهه ۹۰ است؛ این گروه‌ها حول گرایش‌های موضوعی سازمان‌دهی می‌شدند و برای کاربران، ارتباطات ایمیلی یا چت فراهم می‌کردند. شبکه ارتباطات اطلاعات جهانی موسوم به اینترنت، سامانه شبکه سطح بالاتری را ایجاد کرد: وب ایجاد ارتباط در بین افراد و گروه‌های با گرایش‌های فکری مشابه. اگر چه ایده بنیادی وصل کردن افراد در سطح جهانی از آن موقع تغییر زیادی نکرده، دامنه و تأثیر سرویس‌های رسانه‌های اجتماعی در حد غیر قابل وصفی رشد کرده است. اگر چه طبیعی است که قسمت عمده‌ای از گفتگوها هنوز در «دنیای واقعی» رخ می‌دهد، تغییر جهت به سمت تبادل اطلاعات الکترونیکی در سطح اندرکنش‌های انسانی، قوی‌تر شده است. رشد سریع دستگاه‌های همراه و قابلیت اتصال، «اینترنت را در جیب افراد قرار داده»، و همراه آن، امکان برقراری ارتباط با دوستان، خانواده‌ها، و شرکت‌های تجاری مورد نظر، در هر زمان و در هر مکان، فراهم شده است.

جای تعجب ندارد که مجموعه‌ای از خدمات شکل گرفته و نیازهای ارتباط و اشتراک فراهم شده، فضای عمومی و خصوصی را متحول کرده است. از طریق این خدمات، امکان شناخت نحوه تفکر دیگران در مورد سیاست، برند، فراورده، و غیره، فراهم شده است. با به اشتراک گذاشتن شناسنامه‌دار یا بدون نام و نشان، امکان طرح آزاد تفکرات انسان‌ها در مقایسه با رسانه‌های سنتی، بیش از پیش فراهم گردیده است. با توجه به اینکه هر کس در صورت انتخاب، می‌تواند پیام خویش را به دیگران منتقل کند، مسئولیت مجموعه‌های خدمت‌رسانی نیز برای در

دسترس قرار دادن محتوای مربوط و جالب به کاربران، مطرح است.

چیزی که در مورد همه این خدمات مشترک است، وابستگی آنها به افراد بوده، و آنها فقط نقش تسهیل‌گری را ایفا می‌کنند. این امر بدین معنی است که تا حدی، نظم ریاضی کشف شده در اثر تحلیل داده‌های کاربری، انعکاسی از رفتار افراد است. بنابراین، انتظار می‌رود در هنگام کار با این مجموعه داده‌ها، نکات و چالش‌های مشابهی مشاهده گردد. هدف این کتاب، نمایان کردن این نظم و تبیین رویکردهای کمک به شناخت نحوه توجه کاربران به این خدمات، از دریچه داده‌های جمع‌آوری شده توسط سامانه‌های این خدمات، است.

سنجش اندرکنش‌های انسانی

نیروی محرکه رسانه‌های اجتماعی، اندرکنش افراد حول محتوایی است که خدمت برخط ارایه می‌دهد. به‌عنوان مثال، شبکه‌سازی اجتماعی، ارتباط افراد با یکدیگر، به اشتراک‌گذاری تصویر و چندرسانه، گزارشات خبری، محتوای وب، و اطلاعات مختلف دیگر، را تسهیل می‌کند. در رایج‌ترین سناریوی کاربرد این خدمات، افراد برای به‌روز شدن در مورد دوستان، خویشاوندان، و برای به اشتراک گذاشتن چیزی در مورد زندگی خود با آنها، از فیسبوک استفاده می‌کنند. برای توییتر، چون ارتباط لزومی ندارد دوطرفه باشد، کاربران می‌توانند در مورد نحوه فکر، به اشتراک‌گذاری، یا ارتباط یک فرد با دیگران، اطلاع پیدا کنند. با لینکدین، شبکه اجتماعی متخصصین، هدف، وصل کردن متخصصین با گرایش مشابه به یکدیگر از طریق شبکه و گروه‌های آنها، و وصل کردن شرکت‌های استخدام‌گر با متقاضیان کار است.

سرویس‌های رسانه‌ای دیگری وجود داشته که در آن، جنبه شبکه‌سازی تعاملات اجتماعی، به‌جای ایجاد یا لذت از محتوای اشتراکی (مثلاً، ویکی‌پدیا، یوتیوب، یا اینستاگرام)، بیشتر نقش تسهیل‌گری دارد. اگر چه ارتباط بین کاربران می‌تواند وجود داشته باشد، آنجا هدف، قابل مدیریت کردن کشف محتوا برای کاربران و کارآمد کردن ایجاد محتوا (مثلاً، مقالات ویکی‌پدیا)، است.

البته، سایت‌ها و خدمت‌رسانی‌های رسانه‌های اجتماعی زیاد دیگری وجود داشته که گرایش یا زمینه خاصی را هدف‌سازی می‌کنند (هنر، موسیقی، عکس‌برداری، مراکز علمی، اماکن جغرافیایی، مذاهب، سرگرمی‌ها، و غیره). این وضعیت نشانگر این امر است که بیشترین تمایل کاربران برخط برای ارتباط، مبتنی بر همسانی گرایش‌ها یا اشتراکات است.

صرف نظر از زمینه‌های به شدت گسترده تمرکز، یک جنبه این خدمات مشترک است: فلسفه وجودی آنها، وجود *کاربر* و *مخاطب* است. این جنبه، چیزی است که آنها را از محل‌های «از قبل ایجادشده» اینترنت استاتیک، نظیر سایت‌های خبری رسانه‌های سنتی، صفحه خانگی شرکت‌ها،

دفترچه‌ها، و تقریباً هر منبع وب ایجادشده به‌صورت متمرکز توسط گروه نسبتاً کوچک مسئولین ایجاد محتوا («کوچک»، حداقل در مقایسه با میلیون‌ها فردی که از سرویس‌های رسانه‌های اجتماعی استفاده می‌کنند)، متمایز می‌نماید. حاصل دینامیک جمعی این میلیون‌ها کاربر رسانه‌های اجتماعی، چیزی است که در هنگام بررسی عمیق الگوهای کاربری این خدمات، قابل مشاهده بوده، و این جنبه‌ای است که شناخت آن در این کتاب، مدنظر است.

رفتار برخط از طریق جمع‌آوری داده

در هنگام جمع‌آوری داده‌های کاربری سرویس‌های رسانه‌های اجتماعی، نیم‌نگاهی به رفتار آماری بسیاری از انسان‌های با انگیزه یا انتظارات یا هدف مشابه، وجود دارد. به‌طور طبیعی، نحوه سازمان‌دهی خدمات ذی‌ربط و شیوه بروز محتوای آنها، نقش زیادی در مورد اقلام مشاهده شده در سوابق فعالیت‌های کاربران دارد. سوابق دسترسی و استفاده در پایگاه‌های داده این سامانه‌ها ذخیره شده و بنابراین، الگوهای آماری اندرکنش با دیگران و محتوای میزبان‌های خدمات در این روندها قابل مشاهده است. (به شرط وجود این‌گونه الگوها، و طبعاً فعالیت روزانه کاربران به شیوه‌ای کاملاً نامنسجم و تصادفی رخ نمی‌دهد، مشاهده می‌گردد که، همان‌گونه که شاید مورد انتظار است، نظم آماری به وفور در همه جا وجود دارد).

خوش‌بختانه، در اکثر موارد، طراحی سامانه‌ها به حدی متفاوت نیست که ویژگی‌های رفتاری کاملاً متفاوتی از کاربران بروز داشته باشد. برای تبیین این امر، به عنوان مثال، فرض کنید که در نظر است یک آیتم ساده سنجش شود: فرکانس بازگشت کاربران به سامانه در هفته و انجام نوعی از فعالیت. برای هر کاربر، این فقط یک عدد است که دامنه‌ای از صفر تا (در تئوری) بی‌نهایت دارد. البته، در یک زمان محدود، میزان کاربری، بی‌نهایت نبوده ولی هنوز می‌تواند عدد بزرگی باشد. بنابراین، برای سنجش تعداد فعالیت، آیا می‌توان انتظار داشت که نتایج *آماری* متفاوتی برای دو سامانه مختلف حاصل گردد: کاربرهای بارگذاری‌کننده ویدئو بر روی کانال‌های یوتیوب و کاربران آپلودکننده عکس به حساب‌های Flickr؟

بدیهی است که پاسخ، یک بله محکم است. اگر به توزیع تعداد دفعات استفاده افراد نگاه شود، البته مشاهده می‌گردد که درصد کاربران یوتیوب آپلودکننده یک ویدئو در هفته از درصد کاربران فلیکر آپلودکننده یک تصویر در هفته، متفاوت است. این وضعیت امری طبیعی است، چون این دو سامانه جمعیت‌های متفاوت با سناریوهای کاربری مختلفی را جذب کرده، و در نتیجه، توزیع‌ها متفاوت خواهند بود. با این حال، چیزی که شاید سرراست است، این‌است که در اکثر سامانه‌های برخط مورد بررسی محققین، *رفتار آماری کیفی مشابهی* برای این توزیع‌ها

مشاهده می‌گردد.

«کیفی» بدین معنی است که اگر چه پارامترهای دقیق مدل کاربری ممکن است برای دو سامانه ذی‌ربط متفاوت باشد، خود مدل، که از آن طریق بهترین شرح رفتار کاربر در هر دو سامانه قابل انجام است، هنوز مشابه یا خیلی مشابه (با شاید تغییرپذیری جزئی) است.

خبر خوش در این مورد این است که به میزان معقول می‌توان مطمئن بود که آیتم مورد سنجش با داده‌های سوابق فعالیت، در واقع رفتار انسانی نهشته‌ای است که نیروی محرکه ایجاد، نفوذ، اشتراک‌گذاری، و غیره روی این سایت‌ها است. مورد دیگر این خبر خوب این است که برون‌یابی قابل انجام بوده، و در صورت مواجهه با خدمات جدید عملگر بر روی محتوای تولید شده توسط کاربر، حدس حساب‌شده‌ای در مورد اینکه چه چیزی در آن قابل سنجش است، می‌توان داشت. بنابراین، اگر چیزی غیرقابل انتظار در نمودارها مشاهده شد که از الگوی کلی قبلی متفاوت است، بایستی به دنبال یک دلیل خاص سایت تحت بررسی بود که باید بیشتر مورد بررسی قرار گیرد.

بنابراین تا حدی، روش‌ها و نتایج مطرح شده در این کتاب، در صورت تبعیت سامانه خدمات از رفتار انسانی نهشته مشابه، می‌تواند بالقوه به سیستم کاملاً جدید به خوبی قابل کاربرد باشد. با چند استثناء، این حالت برای سامانه‌های سرویس‌های رسانه اجتماعی که مورد پژوهش قرار گرفته‌اند، صحیح بوده، و بنابراین، تمایل است که این‌گونه فکر شود، این سامانه‌ها نکاتی در مورد رفتار انسانی تأمین می‌کنند. آنگاه، فرصت مشاهده و شرح اینکه انسان‌ها به‌صورت جمعی رفتاری مشابه دارند، مسبوق به سابقه نیست؛ علت، اثرانگشت دیجیتال باقی مانده در سوابق سامانه‌ها است. (حریم خصوصی، البته، یک دغدغه عملی معتبر است، ولی تمایل بیشتر در تصویر کلان بوده و رفتار فردی خاص مدنظر نیست). در بخش‌های بعدی، انواع داده‌های مدنظر در سرویس‌های رسانه‌های اجتماعی مختلف و مجموعه داده‌های مورد استفاده برای مثال‌های این کتاب، مورد بررسی قرار می‌گیرد.

داده‌های ضروری برای جمع‌آوری

سؤالاتی که نهایتاً با داده‌ها باید پاسخ داده شوند، تعیین‌کننده نوع داده‌های مورد نیاز برای جمع‌آوری است، ولی به طور کلی، هر چه میزان داده جمع‌آوری شده بیشتر باشد، پاسخ بهتری برای سوالاتی که در آینده ممکن است ایجاد شود، پیدا خواهد شد. هرگز مشخص نیست که در چه موقع، دقیق‌تر کردن یا بسط تحلیل داده‌ها ممکن است مدنظر باشد و بنابراین، اگر یک سامانه خدمت‌رسانی قرار باشد طراحی شود، بهتر است آینده‌نگری شده و همه یا بیشتر

اندرکنش‌های کاربران با سامانه و با یکدیگر، ثبت شود. امروزه، ذخیره داده‌ها ارزان بوده و بنابراین، نباید سعی کرد که خیلی زود در استفاده از فضای ذخیره بهینه‌سازی صورت گرفته، و حتی‌المقدور نیازهای آتی بالقوه برای داده‌های بیشتر مدنظر قرار گیرد. به‌طور طبیعی، با تکامل سامانه خدمات و روشن شدن زمینه‌های تمرکز، هرس داده‌های جمع‌آوری شده ممکن گردیده، و در صورت نیاز، می‌توان منابع داده موجود را بازنگری کرد.

برای شناخت بهتر داده‌های فعالیت کاربر مورد نیاز، بعضی از سؤالات حول کاربری رسانه‌های اجتماعی که می‌تواند مدنظر قرار داشته باشد، به شرح زیر است:

- کاربران با بیشترین/کمترین فعالیت کدامند؟ چه تعداد از آنها وجود دارد؟
- نحوه تکامل کاربری با گذشت زمان چگونه است؟ آیا می‌توان از ابتدا، میزان کاربری بخش‌های مختلف (برحسب جغرافی، جمعیت، نوع کاربری) را پیش‌بینی کرد؟
- به چه شیوه می‌توان کاربران را با محتوا تطبیق داد؟ کاربران به کاربران؟ به چه شیوه محتوای نظر مورد توجه کاربر را می‌توان با سرعت کافی استخراج کرد؟
- شبکه‌های کاربران چه شکلی دارد؟ کاربران فعال از شبکه‌های مختلف، با یکدیگر متفاوت هستند؟
- در صورت خروج افراد از سامانه خدمات، علت چیست؟ پیش‌نیازی برای این خروج وجود داشته و قابل پیش‌بینی است؟
- چه عواملی در ملحق شدن کاربران به خدمات دخیل است؟ وجه تمایز کاربران راضی و ناراضی چیست؟
- کاربرانی وجود دارند که از خدمات به هر طریق استفاده می‌کنند؟ آیا هرزکاربر، کاربرد غیرمعقول، و رفتار متقلبانه در کاربران وجود دارد؟
- بخش‌های محتوای با بیشترین «جاذبه» یا «روندسازی» در هر زمان خاص چیست؟ چه زیرمجموعه‌ای از کاربران به آن توجه داشته، احصای آنها به چه شیوه امکان‌پذیر بوده، و در مورد چیست؟
- آیا محتوای خاص مدنظر را می‌توان از داده‌های جاری یا تاریخی تولید شده توسط کاربران، استخراج کرد؟ به‌عنوان مثال، آیا می‌توان کاربرانی که اخیرا واژه خاصی را به‌کار برده یا موضوع خاصی را مطرح کرده‌اند، احصا نمود؟
- چه بخشی از محتوا در بین کاربران «مورد اقبال» است؟ تفاوت‌های اساسی بین اقلام مورد اقبال وجود داشته، و در صورت مثبت بودن، چه میزان؟

در فصول این کتاب بعضی از سوالات مطرح شده و پاسخ‌ها برای خدمات خاص ارایه

می‌گردد. همان‌گونه که ممکن است مشهود باشد، بعضی از این موارد را با انجام آزمایش با کاربران، به‌طور مشخص آزمایشات آزمون الف/ب، می‌توان پاسخ داد (در آزمایش آزمون الف/ب، یک ویژگی نشان داده شده یا یک الگوریتم برای هر مجموعه کاربران الف استفاده شده، و همین کار برای مجموعه کاربر ب انجام می‌شود. با سنجش تفاوت‌های بین فعالیت‌های گروه الف و ب، می‌توان در مورد نوع تأثیر تغییر در ویژگی بر کاربران، تصمیم گرفت). با این حال، با توجه به تمرکز روی داده‌های جمع‌آوری شده در قبل و بیشترین یادگیری ممکن از آنها، این روش قدرت‌مند مورد استفاده به‌طور کلی برای بهینه‌سازی تجربه کاربر روی سامانه خدمات، پوشش داده نمی‌شود.

بر این اساس، چه نوع داده‌هایی از سامانه‌های خدمات ارایه شده یا از دیگر رسانه‌های اجتماعی قابل دسترسی، باید جمع‌آوری شود؟ با بهره‌گیری از سؤالات قبلی، چندین جنبه سوابق داده‌ها برای تحلیل لازم است؟

۱. با ورود کاربران به سامانه خدمات، فعالیت‌های مشخصی انجام می‌شود: خواندن مقالات، مشاهده تصاویر، برچسب‌گذاری عکس‌ها، و به‌روزرسانی وضعیت اشتراک‌گذاری. شناسه (بی نام و نشان شده) کاربران، همراه با توصیف فعالیت‌ها، چیزی است که در هنگام بررسی‌های فعالیت‌های آنها، مورد پرسمان است.
۲. شناخت *هنگام* وقوع فعالیت‌ها نیز لازم است. دقت زیرثانیه‌ای برای جمع‌آوری داده‌ها (میلی- یا میکروثانیه) معمولاً کفایت می‌کند.
۳. بدیهی است که برای هر فعالیت، چندین نوع مختلف بخش‌های فراداده می‌تواند همراه باشد. اگر، به‌عنوان مثال، کاربران به یک پست گرایش داشته یا آنرا دوست دارند، بدیهی است که شناسه یکتای آن پست همراه با فعالیت، باید ذخیره شود.

با توجه به اینکه کاربران می‌توانند فعالیت‌های زیادی در یک مدت خاص داشته باشند، داده‌های خام ثبت شده بدین‌طریق، ممکن است نهایتاً مقدار زیادی از فضای ذخیره را به‌خود اختصاص دهند. این وضعیت می‌تواند زمان زیادی برای پردازش حتی سوالات ساده، را ایجاب کند؛ بر این اساس، همه اطلاعات اکثر سوالات رایج همیشه لازم نیست. بنابراین، در حالت عادی، در محیط تولید، عکسی لحظه‌ای از داده‌های انباشته‌شده، به‌عنوان مثال در مورد حالت فعلی نمودار اجتماعی با همه روابط بین کاربران، تعداد توییت، پست، و عکس‌هایی که ایجاد کرده یا به اشتراک گذاشته، و غیره، از طریق فرایندهای ETL (استخراج، تبدیل، بارگذاری)، گرفته می‌شود. در هنگام تحلیل داده‌ها برای دست‌یابی به بعضی از نکات، این توده داده‌ها، به کرات، اولین منبع اطلاعاتی است که مورد رجوع قرار می‌گیرند.

اگر چه باید در مورد نحوه بهینه ذخیره همه این داده‌ها در پایگاه‌های داده مناسب فکر کـرد، طراحی و پیاده‌سازی این‌گونه قالب‌ها، یک تخصص بوده که خارج از دامنـه بحـث موضـوع ایـن کتاب است. هم‌چنین، در نظر است که تمرکز بر روی روش استخراج نکات از داده‌هـا بـوده، و از داده‌های عمومی موجود در سرویس‌های رسانه‌های اجتماعی برای شرح نحوه انجـام تحلیـل‌هـا، استفاده خواهد شد.

سؤال و جواب با داده‌ها

هدف، طرح چند موقعیت رایج مورد مواجهه در حین تلاش بـرای شـناخت داده‌هـای تولیـدی از سامانه‌های رسانه‌های اجتماعی است. روش معمول مطالعه پدیده‌هـای تجربـی (نـه لزومـاً فقـط مربوط به رسانه‌های اجتماعی)، تبعیت از سنت صدها ساله روش علمی است:

۱. به‌طور عام، سوال پرسیدن اولین گام است. در این برهـه، نیـازی بـه انجـام فـرض‌هـای اضافی در مورد داده‌ها نیست؛ فقط رسمی کردن چیزی است که در مورد شـناخت رفتـار مشخص، مدنظر می‌باشد. به‌عنوان مثال، «دینامیک زمانی بازگشت کـاربران بـه سـامانه خدمات چیست تا بتوان پیش‌بینی کرد چه مدت زمان ماند کاربر طول خواهد کشید؟»
۲. گزینه اختیاری، فرموله کردن فرضیه در مورد خروجی مورد انتظار است. این حالت بـرای صحه‌گذاری معقول بودن پیش‌فرضها قابل استفاده است. هم‌چنین، اگر یک مدل ذهنـی وجود داشته که فکر می‌شود بهترین خروجی کمّی را دارد، چک کردن آن از ایـن طریـق قابل انجام است. پس از فرموله کردن فرضیه، پیش‌بینی می‌شود که در صـورت صـحیح بودن آن، چه نتیجه‌ای باید حاصل گردد. این گام اختیاری است، چون اگر سـاخت مـدل حول سوال مدنظر نبوده، و هدف فقط دست‌یابی به نکات باشد، از این گام می‌توان گـذر کرد. یک فرضیه در مورد سوال ۱ می‌تواند، به عنوان مثال، ایـن باشـد کـه «مسـتقل از اینکه خدمات اخیراً استفاده شده باشـد، کـاربران بـه شـیوه‌ای تصـادفی بـه سـامانه وارد می‌شوند». (درست بودن این فرضیه در خدمات واقعی، در فصل ۳ بررسی خواهد شد).
۳. تعیین رویه قابل تبعیت و داده‌های مورد نیاز برای جمع‌آوری برای پاسخ دادن به سوال ۱. اگر چه برای ابزار محاسباتی و روش‌های موجود خاص، رویـه سرراسـت اسـت، معمـولاً آزادی زیادی در رسانه‌های اجتماعی برای انتخاب مجموعه داده‌های تسـت، وجـود دارد. آیا در نظر است که از بین کاربران نمونه‌برداری صورت گرفته یا داده‌های همـه کـاربران مورد استفاده قرار می‌گیرد؟ چه دوره زمانی استفاده خواهد شد؟ آیا بعضی از فعالیت‌هـایی که نامطلوب تشخیص داده می‌شوند، فیلتر می‌گردند؟ بـدیهی اسـت کـه جـامع بـودن و

کاوش هر چه بهتر داده‌ها،به عنوان مثال با در نظر گرفتن دوره‌های مختلفی از مجموعه داده یا نگاه به گروه‌های کاربر مختلف، برای اطمینان داشتن از نتایج مدنظر است. بـرای سوالاتی که قرار است پاسخ یافت شود (به سوال ۱ رجوع شود)، مهر زمـانی داشـتن هـر فعالیت تولیدی توسط کاربران در یک ماه مشخص، به‌عنوان مثال، و آنگاه تعیین تفـاوت زمانی بین مهرهای زمانی مختلف و تحلیل هم‌بستگی‌های زمانی، مورد نظر می‌باشد.

۴. انجام تحلیل داده‌ها. در حالت ایده‌ال، جمع‌آوری داده قبلاً توسط شـما یـا بـرای اسـتفاده شما صورت گرفته، و نیازی به انتظار برای این امر نیست. اگر هدف، آزمون فرضیه است، انجام آزمون آماری نیز لازم است. اگر فقط دست‌یابی به نکات مدنظر است، نتایج عددی پاسخ سوال مطرح شده است.

مجموعه داده مورد استفاده

برای روشن کردن پـردازش‌هـا و نظـم‌هـای قابـل مشـاهده در رسـانه‌هـای اجتمـاعی ناشـی از اندرکنش‌های انسانی، به‌طور طبیعی، دسترسی به بعضی از داده‌های خروجی قابل دانلود کردن از مکان‌های مختلف روی اینترنت، از این‌گونه سامانه‌ها لازم اسـت. اگـر چـه اکثـر سـرویس‌هـای رسانه‌های اجتماعی، داده‌های خود را خصوصی حفظ می‌کنند (دغدغه‌های حریم خصوصی، دلیـل اصلی بوده ولی علت دیگر، حجیم شدن این داده‌ها نیز هست)، بعضی از این سـرویس‌هـا، نظیـر ویکی‌پدیا، همه داده‌های خویش را در اختیار عمـوم قـرار مـی‌دهنـد. در دیگـر مـوارد، محققـین آکادمیک، داده‌هایی از این سرویس‌ها از طریق خزش یا اشتراک داده، را جمع‌آوری مـی‌کننـد. در بخش‌های زیر، لیستی از منابع داده مورد استفاده در این کتاب آورده شده است. توصیه می‌شـود، خوانندگان مثال‌هایی که این داده‌ها پیش‌نیاز آنهاست، را استفاده نمایند (و آنها را بسط دهند).

چند مجموعه داده که عمومی و به‌طرز گسترده موجود بوده، و داده‌های مربوط به کـاربران و محتوای آنها به سادگی قابل دسترسی است، انتخاب شده، تا نشـان داده شـود در سـرویس‌هـای رسانه‌های اجتماعی واقعی برای سوالاتی که مطرح می‌شود، چه نتایجی قابل انتظار است. اسامی این خدمات آشنا بوده، و در نظر بود این مجموعه داده‌ها برای کاربران و دامنه زمانی مدنظر آنهـا برای سـیر، انـدازه متوسـط تلقـی شـده، و بـدین‌طریـق، بـرای نتیجـه‌گیریهـای معنـادار، قابـل پرداخت‌کاری باشد. مثال‌های عملی نشان داده شده در کتاب را استفاده کنید؛ بـا ایـن هـدف، در بخش‌های زیر، مجموعه داده‌های مورد استفاده شرح داده شده است. بـه‌عنـوان خلاصـه، شـرخ مختصری از مجموعه داده‌های نمونه، در جدول ۱ آورده شده است.

جدول ۱ توصیف و آدرس داده‌های مورد استفاده

SERVICE	MAIN PAGE	DATASET
Wikipedia	wikipedia.org	Revision and page meta information, no actual text
Twitter	twitter.com	Tweets created
Stack Exchange	scifi.stackexchange.com	Questions and answers from Stack Exchange's Science Fiction & Fantasy category
LiveJournal	livejournal.com	Directed social network connections
Cora dataset		Scientific documents from an academic search engine
MovieLens	movielens.org	Sample of movie ratings
Amazon Fine Food Reviews		Historical reviews on Amazon for "Fine Foods"

برای ساده‌تر کردن دسترسی خواننده به داده‌ها، همه داده‌هایی که مثال‌ها بـر روی آنهـا بنـا شده، با اجرای data/download_all.sh از سایت http://booksupport.wiley.com، قابـل دسترسی است. (توجه کنید که به علت بزرگ بودن مجموعه داده‌ها، خصوصاً داده‌های ویکی‌پدیا، کامل کردن دانلود ۶۰–۵۰ گیگابایت داده، مقداری زمان خواهد برد.). محل کد منبـع در انتهـای این مقدمه آورده شده است.

ویکی‌پدیا

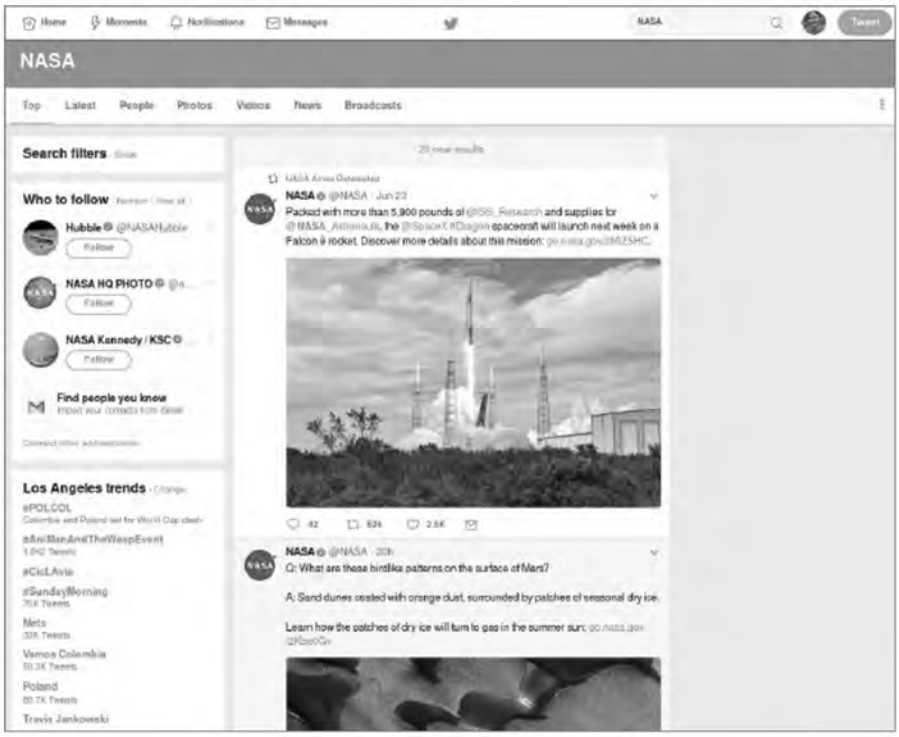
بزرگترین مجموعه داده مورد استفاده در زبان انگلیسی، تاریخچه روایـات چنـدین میلیـون مقالـه موجود در آن است. ویکی‌پدیا، دایره‌المعارف حاصل از تلاش جمعی اسـت و روایـت انگلیسـی آن حاوی تقریباً ۵.۷ میلیون مقاله در ۲۰۱۸ است که حدود ۳۰۰ هزار ویراستار فعال در هر مـاه روی آن فعالیـت مـی‌کننـد (http://en.wikipedia.org/wiki/Wikipedia:Statistics). تصـویر صفحه مقاله «Wikipedia» در شکل ۱ نشان داده شده است.

شکل ۱ نمونه دایره‌المعارف ویکی‌پدیا در مورد ویکی‌پدیا

توییتر

در توییتر (شکل ۲)، کاربران می‌توانند با پیام حداکثر ۱۴۰ کاراکتری (که اخیراً تا ۲۸۰ کاراکتر افزایش یافته است)، وضعیت خاصی را به‌روزرسانی کنند. دیگر کاربرانی که ارسال‌کننده را «دنبال» می‌کنند، این پیام‌های کوتاه را در صفحه خود که موسوم به خطزمانی است، دریافت می‌کنند. تصویر و ویدئوی کوتاه نیز می‌تواند به به‌روزرسانی وضعیت پیوست شود. بسیاری از کاربران منابع خبری، سلبریتی‌ها، یا دوستان و خانواده خود را دنبال (فالو) می‌کنند. غالباً، توییتر یک «شبکه اطلاعات» در نظر گرفته شده که در آن، کاربران می‌توانند از هر کسی که تمایل دارند، به‌روزرسانی دریافت نموده، ولی لزومی ندارد آن فرد نیز، دریافت‌کننده را فالو کند.

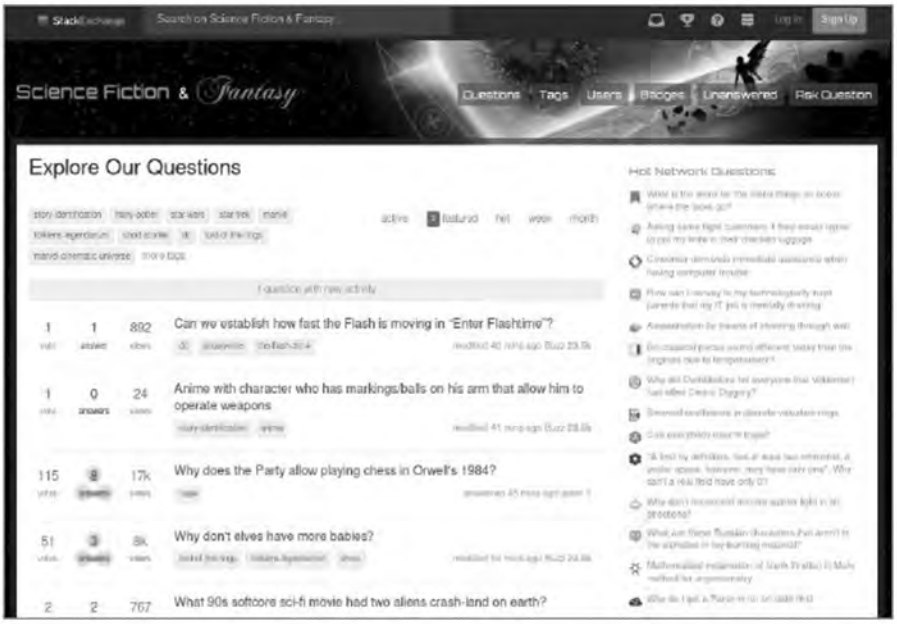
توییت‌ها با استفاده از API توییتر برای تحلیل فعالیت نمونه کاربران در فصل ۱، جمع‌آوری می‌شوند.



شکل ۲ تصویر تیپ جستجوی خطزمانی توییتر. توییت‌ها در بخش اصلی و موضوعات جاری و توصیه‌های «چه کسی دنبال شود»، در کناره قرار گرفته است.

تبادل بسته

Stack Exchange (شکل ۳)، یک شبکه فدراسیونی وب‌سایت‌های با مدل سوال–جوابی است که در آن، کاربران در مورد موضوعات مختلفی سوال مطرح کرده، و کاربران دیگر می‌توانند این سوالات را جواب داده و در مورد سوال و جواب رای بدهند. از این طریق، محتوای با کیفیت بالا (حداقل در نگاه کاربران)، رتبه بالایی کسب می‌کنند. تا ۲۰۱۸، این شبکه از بیش از ۳۵۰ سایت تشکیل شده که موضوعات مختلفی از برنامه‌سازی نرم‌افزار تا نجوم تا بازی‌های خاص را پوشش می‌دهند. در فصل ۴، یکی از سایت‌های موضوعی این سایت، یعنی رده تخیل و فانتزی علمی، استفاده شده و ویژگی‌های پست‌های مختلف ارسال شده کاربران به آنجا، مورد بررسی قرار می‌گیرد.



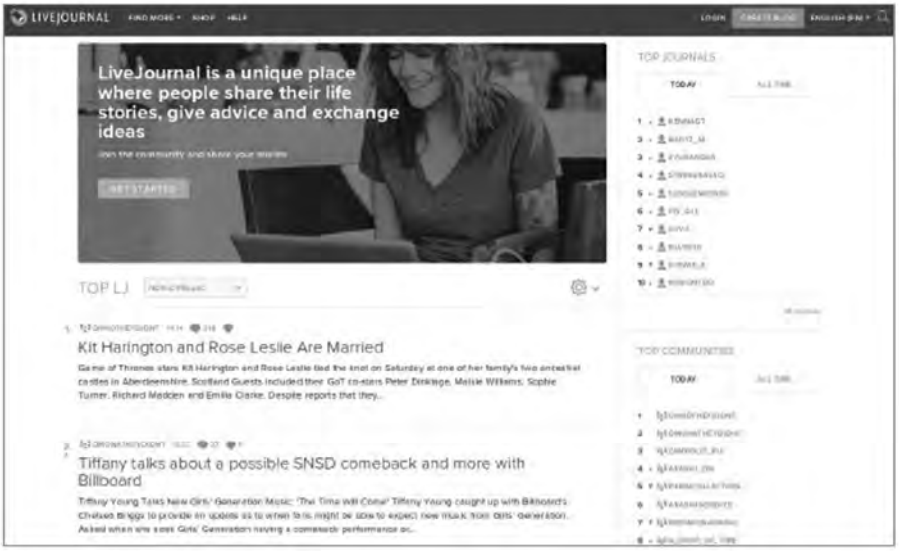
شکل ۳ تبادل بسته یک خدمت‌رسانی با تعداد زیادی زیرسایت موضوعی.

LiveJournal

LiveJournal (شکل ۴)، یک خدمت مخزن ژورنال و بلاگ است که در آن، کاربران می‌توانند ارتباط دوطرفه یا یک‌طرفه با دیگر کاربران را برقرار کنند. دوستان کاربران، می‌توانند ورودی‌های حفاظت‌شده آنها را خوانده، و بالعکس، پست‌های بلاگ دوستان روی «صفحه دوست» آنها ظاهر می‌شود. در فصل ۲، از این مجموعه داده برای مطالعه ساختار ارتباط جهت‌دار شبکه اجتماعی، استفاده خواهد شد.

اسناد علمی Cora

این مجموعه کوچک‌تر بوده و حاوی ۲۴۱۰ سند علمی از موتور جستجوی Cora است. (این موتور جستجو، قبلاً برای انتشارات آکادمیک در علوم کامپیوتر به عنوان اثبات مفهوم استفاده می‌شد). در فصل ۴، از این مجموعه داده برای شرح رویکرد مدل‌سازی موضوعی برای متون زبان طبیعی، استفاده خواهد شد. این مجموعه داده همراه با بسته lda زبان R همراه است؛ دانلود اضافی لازم نیست.



شکل ۴ صفحه اصلی پلاتفرم بلاگ LiveJournal که تشکیل گروه را نیز تسهیل می‌کند.

بررسی غذای ممتاز آمازون

این مجموعه، حاوی نتایج بررسی‌های «غذای ممتاز» آمازون، شامل خلاصه‌های بررسی فراورده، نمرات، و بعضی جزئیات دیگر است. مجموعه داده، برای ۱۰ سال تا اکتبر ۲۰۱۲ است. برای جزئیات بیشتر به https://snap.stanford.edu/data/web-FineFoods.html جوع شود.

رتبه‌بندی فیلم MovieLens

این مجموعه داده، حاوی درجه‌بندی فیلم از خدمات MovieLens (https://movielens.org) در مقیاس ۱ تا ۵ است که توسط ۹۳۸ کاربر برای ۱۶۸۲ فیلم جمع‌آوری شده است. در فصل ۶، در مثال‌ها از این مجموعه داده استفاده شده تا، با استفاده از نحوه درجه‌بندی فیلم‌ها توسط کاربران، پیش‌بینی شود چگونه کاربران مشابه احتمالاً فیلمی که ندیده‌اند، را بررسی می‌کنند.

زبان‌ها و چارچوب‌های مورد استفاده

مثال‌های این کتاب، عمدتاً در سه زبان برنامه‌سازی و قالب نوشته شده است: R، پایتون، و Scalding. از R برای توانمندی‌های عالی آن در آمار، یادگیری ماشین، و گرافیک؛ از پایتون به

خاطر ساده بودن پیش‌پردازش مجموعه داده‌های بزرگ و برقراری ارتباط با APIهای خدمات و سرعت؛ و اسکالدینگ برای قالب قابل انعطاف و متقن برای انجام محاسبات توزیعی روی MapReduce، استفاده شده است.

به‌طور کلی، باور ما این نیز هست که این ابزار برای داده‌کاوی عالی هستند؛ بنابراین، فرض بر این است که خواننده با آنها آشنا بوده، یا حداقل می‌تواند کد نوشته شده در آنها را بفهمد. این زبان‌ها، مسیر توسعه سریعی برای نمونه‌سازی الگوریتم‌ها و نوشتن تست‌های سریع حول داده‌ها، فراهم کرده، و از طریق پشتیبانی مردم نهاد گسترده، پاسخ تقریباً هر چالش فنی رایج به سادگی در تالارهای برخط، موجود است.

عناوین این نمونه کدها، مربوط به پرونده کد منبع مثال است (مگر اینکه کد خیلی کوتاه باشد). پرونده‌های منبع در زیرفولدر src/chapterX مخزن کد کتاب است که در آن، X مربوط به فصلی است که نمونه‌های کد ظاهر می‌شود.

اسکریپت‌ها باید از فولدری که در آن، مخزن استخراج شده، اجرا شوند و نیازی نیست که دایرکتوری پایه به جایی که قرار دارند، تغییر داده شود: برای مثال، برای دانلود کردن فقط مجموعه ویکی‌پدیا، می‌توان src/chapter1/wikipedia/get_data.sh را اجرا کرد؛ برای پیش‌پردازش مجموعه Stack Exchange، می‌توان python src/chapter4/process_stackexchange_xml.py را اجر کرد (در فصول ذی‌ربط، کاری که این اسکریپت‌ها انجام می‌دهند، شرح داده خواهد شد).

R

R یک زبان برنامه‌سازی آماری است که نه فقط بین آماردان‌ها، بلکه بین متخصصین دیگر رشته‌ها که متمایل به تحلیل داده هستند، مورد اقبال است. علت، گستردگی مجموعه کتابخانه‌هایی است که انجمن‌ها برای آن توسعه داده‌اند: در بررسی صفحه CRAN در مورد کتابخانه‌های موجود (http://cran.r-project.org/web/views/)، رده‌های اکونومتریک، فاینانس، ژنتیک، علوم اجتماعی، فن‌آوری‌های وب، و غیره مشاهده می‌شود. با توجه به اینکه R یک منبع آزاد و باز است، فرهنگ اشتراک کد، باعث شکل‌گیری کتابخانه توسعه داده شده مردم نهاد از سراسر جهان شده است. جمعیت حول R نیز فعال است، و یافتن پاسخ برای موضوعات رایج، ساده است.

برای افراد ناآشنا با این زبان، فرمت ممکن است مقداری مشکل باشد. شیب منحنی یادگیری تند بوده ولی ارزش یادگیری را دارد. آموزش عملی رسمی R موجود در سایت اصلی، (https://cran.r-project.org/doc/manuals/R-intro.pdf)، روش خوبی برای آشنا شدن با

زبان برای فهم مثال‌های این کتاب است. یک مزیت R، مکانیسم ذخیره داده موسوم به قالب‌های داده است که در آنها، اسناد مرتبط داده‌ها را می‌توان در ستون‌های با نام ماتریس (با این استثناء که ستون‌ها می‌توانند بردارهای با نوع سلیقه‌ای، به‌جای فقط ارقام عددی، داشته باشند)، قابل ذخیره است.

دانلود و نصب سیستم R پایه، سرراست بوده و برای لینوکس، مک، و ویندوز، موجود است. مستندات صفحه نصب پروژه مناسب بوده و نیازی نیست که در اینجا گام‌ها تکرار شود، اگر چه اگر گام‌های بخش «الزامات سیستم برای اجرای مثال‌ها» در زیر در این مقدمه دنبال شود، نیازی به حتی نصب دستی آن نیز نیست. با این حال، یک نکته قابل ذکر اینکه استفاده از سیستم یکپارچه توسعه R (IDE)، مزایای خاص خود را دارد: اگر چه R یک خط فرمان دارد، استفاده از صفحه گرافیکی اندرکنش با کاربر بسیار ساده‌تر است. دو گزینه اصلی در اینجا، RStudio (www.rstudio.com) و StatET برای Eclipse (www.walware.de/goto/statet) است. اولی، نصب یک کلیکی و صفحه اندرکنش سرراستی فراهم کرده، در حالی‌که دومی، برای کار با دیگر زبان‌های برنامه‌سازی برای کاربران فعلی Eclipse، یکپارچگی بهتر و قابلیت انعطاف بیشتر، فراهم می‌کند. نصب چند بسته اضافی بر R پایه نیز برای اجرای مثال‌های کد لازم بوده که در جدول ۲ لیست شده است.

پایتون

اگر چه R یک ابزار قدرتمند و جامع حاوی چندین بسته است که توانمندی‌های آن را افزایش می‌دهد، برای بعضی از فعالیت‌ها که در هنگام تحلیل استفاده از رسانه‌های اجتماعی رایج است، گزینه‌ای بهینه نیست. غالباً، مجموعه داده‌هایی که از سامانه‌های سرویس‌های رسانه‌های اجتماعی جمع‌آوری می‌شود، باید تمیز، فیلتر، یا تبدیل شوند. در این مورد، R بهینه نیست چون تمرکز آن بر عمل در ساختارهای داخل حافظه‌ای است و بنابراین، به‌عنوان مثال، اگر لازم باشد با یک هفته سال داده فعالیت کاربر با مهر زمانی در R کار شود، به‌طور سنتی کل مجموعه داده را ابتدا باید خوانش کرده و نوعی از فیلتر برای محدود کردن دامنه، اعمال کرد. در بسیاری از مواقع، با توجه به مقدار زیاد داده مورد مواجهه، معمولاً این کار روی RAM رایانه شخصی، محال است.

برای پیش‌پردازش و تجمیع مجموعه داده‌های متوسط تا بزرگ برای تحلیل بیشتر، گزینه‌های بهتر موجود است. زبان برنامه‌سازی دیگر برای مثال‌های کد، پایتون است که، از لحاظ زمان صرف شده برای توسعه اسکریپت‌ها، بسیار کارآمد است. هم‌چنین، گسترده‌ترین

کاربرد در سرتاسر جهان را داشته، و جامعه شکل‌گرفته حول آن نیز، بسیار فعـال اسـت. مجـدداً، مشابه R، مجموعه عظیمی از ماژول‌ها برای آن وجود داشته که همه، منبع بـاز بـوده، و بررسـی نحوه عمل کدها به سادگی قابل انجام است. ماژول‌های مورد استفاده در این کتاب در جـدول ۳ نشان داده شده است.

جدول ۲ بسته‌های R مورد استفاده در مثال‌های کد

R PACKAGE	FUNCTIONALITY WE USE
ggplot2 scales	Creates pretty plots using an intuitive syntax for building up the graphs from layers
reshape2	Restructures data frames and switches between long and wide tabular data representations
plyr	Groups data by column values, performs some aggregations on the chunks, and puts the results back together
forecast	Time series forecasting
Matrix	Package for sparse matrices
NMF	Non-negative matrix factorization functions for matrix completion
glmnet	Efficiently solves the logistic regression problem
ROCR	Visualizes performance metrics for prediction tasks
tm	Creates term-document matrices from natural text
ggdendro	Plots dendrograms
Wordcloud dendextend	Create word clouds to visualize frequent terms in documents
entropy	Calculates the entropies of distributions
lda	Implements the Latent Dirichlet Allocation model for topic detection in texts
rPython	Calls Python functions from R—the best of both worlds

جدول ۳ بسته‌های پایتون مورد استفاده در مثال‌های کد

PYTHON MODULE	FUNCTIONALITY WE USE
matplotlib	Plots from Python
networkx python-igraph	Store, traverse, and plot graphs
nltk	Tokenizes documents and finds the stems of words
beautifulsoup4	Preprocesses text containing HTML tags
tweepy	Fetches tweets from the Twitter API
scipy numpy	Generic-purpose scientific & numeric libraries

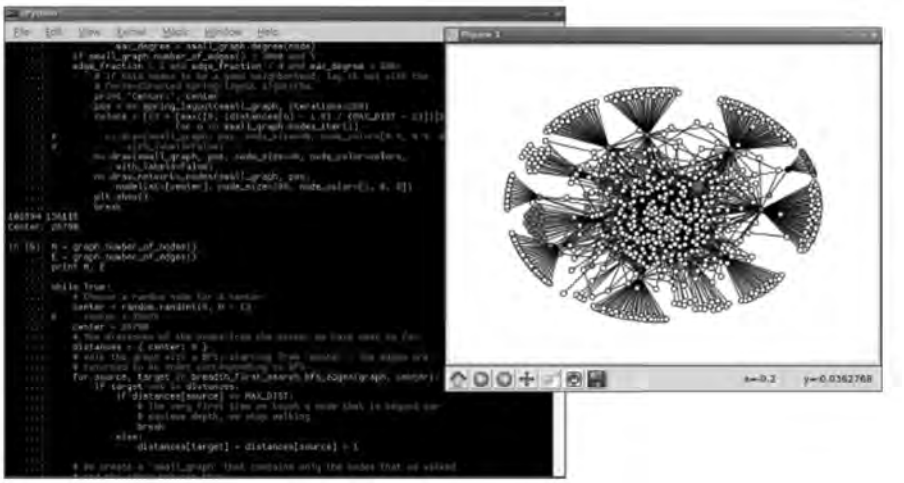
پایتون با استفاده از کتابخانه‌هایی نظیر SciPy, NumPy, matplotlib، و pandas، قادر به اجرای الگوریتم‌های یادگیری ماشین و پلاتفرم‌های آنالیتیک هست. با این حال، با توجه به پوشش عمده این فعالیتها با R در این کتاب، فقط توان‌مندی‌های هسته اصلی زبان و تعداد کمی ماژول اضافی، استفاده خواهد شد. همان‌گونه که ذکر شد، پایتون برای پردازش جاری مجموعه داده‌های با اندازه متوسط، در هنگام نیاز به انجام تبدیل‌های ساده روی آنها، عالی است. هم‌چنین، اگر فعالیت در «سطح پایین‌تری» باشد، نزدیک مسئله برنامه‌سازی رویه‌ای سنتی، پایتون غالباً ابزاری بهتر از R است. فرمت آن، شبه کد بوده، و بنابراین باور ما این است که، اگر خواننده با مفاهیم پایه لیست‌ها و اصطلاحات در پایتون آشنا باشد، حتی در غیاب تجربه با پایتون، مثال‌های کد قابل فهم هستند. (در غیر این‌صورت، آموزش عملی رسمی پایتون در www.python.org/2/tutorial/، نقطه شروع خوبی است). برای اجرای مثال‌های کد، روایت ۲.۷ پایتون لازم است.

در تحلیل داده‌ها، غالباً چندین مرحله محاسبه وجود داشته که روی یکدیگر بنا نهاده می‌شوند. یک مثال کوتاه را در نظر بگیرید: در نظر است که توزیع طول کوتاه‌ترین مسیر در یک شبکه اجتماعی کوچک در بین همه گره‌هایی که درجه آنها (تعداد همسایه)، بیشتر از ۱ است، تعیین گردد. در این مورد، مراحل عبارتند از: بارگذاری شبکه؛ فیلتر کردن برای گره‌های با درجه بیشتر از ۱؛ محاسبه همه کوتاه‌ترین مسیرها؛ و ایجاد هیستوگرام برای نتایج. اگر شبکه بزرگ باشد، محاسبات کوتاه‌ترین مسیر، مثل شاید خواندن و ساختن شبکه از پرونده، ممکن است خیلی طول بکشد. ساده‌ترین راه، نوشتن یک اسکریپت پایتون با همه گام‌ها، و اجرای آن برای یک بار است. با این حال، غالباً اشتباه رخ داده یا در حین تحلیل، مشاهده می‌شود که چیزی فراموش شده و گنجانده نشده است. به‌عنوان مثال، پس از ساختن هیستوگرام، تصمیم گرفته می‌شود که خروجی نه فقط بر روی صفحه نمایش، بلکه باید در فایل نوشته شود. در این مورد، اسکریپت باید تغییر و اجرا شده، و همه محاسبات پرهزینه، مجدداً انجام گردند.

به همین خاطر، تقریباً همیشه استفاده از کنسول پایتون تعاملی بهتر است که در آن، می‌توان در حین نگه‌داشتن همه متغیرها در حافظه، فرمان‌های پایتون را صادر کرد. کنسول پایتون نهادینه برای این کار خوب است (بارگذاری شده با اجرای python). با این حال، یک روایت قدرت‌مندتر آن، ipython (https://ipython.org) و دفترچه یادداشت Jupyter (http://jupyter.org) است که، مجموعه‌ای از دیگر روتین‌های کمکی نظیر اتمام نام متغیر، جستجوی تاریخچه فرمان، رسم نهفته و تعاملی، و محاسبه موازی (http://ipython.org/ipython-doc/dev/parallel/) را فراهم می‌کند. اگر چه در این کتاب

از آن استفاده نمی‌شود، دومی برای محاسباتی که روی یک هسته CPU زمان زیادی نیاز دارد، مفید است.

گزینه‌های مختلفی برای IDE پایتون نیز وجود داشته که برای لیست آن به این سایت رجوع شود: https://wiki.python.org/moin/IntegratedDevelopmentEnvironments. در شکل ۵، یک ipython کنسول‌پایه ساده نشان داده شده است.



شکل ۵ نمونه روند فعالیت IPython تعاملی با نمودار

Scalding

در فصل ۵ کتاب، رویکردهای الگوریتمی پردازش مجموعه داده‌های بزرگ مورد مواجهه در تقریباً همه موارد تحلیل داده‌های سرویس‌های رسانه‌های اجتماعی، مطرح شده است. برای اکثر سوالات نمونه در کتاب، اجرای کد روی یک پردازشگر تکی روی رایانه شخصی، کفایت کرده، ولی ممکن است چندین ساعت طول بکشد. در عمل، تقریباً همیشه در موقع کار با داده‌های فعالیت تولیدی توسط چند میلیون کاربر، باید از راه‌حل‌های محاسبه توزیعی استفاده کرد.

امروزه، سرعت پیشرفت بی سابقه‌ای در توسعه مجموعه ابزار و قالب‌های پردازش داده‌های حجیم را شاهد هستیم که در آن، قالب‌های ابزار، و پایگاه‌های داده قبلی، غالباً ظرف چند سال منسوخ می‌شوند. با این حال، طرحواره MapReduce، به‌عنوان یک مدل غالب برای پردازش ناپیوسته مجموعه داده‌های برزگ روی صدها یا هزاران رایانه، غالب باقی مانده است. علت این امر، قابلیت مقیاس شدن آن به مراکز داده بزرگ و تاب‌آوری آن در مقابل معیوب شدن سرورهای انفرادی بوده، که وقوع آن در شرایط استفاده از تعداد رایانه زیاد به‌صورت مستمر و ۲۴

ساعته در همه مواقع امری اجتناب‌ناپذیر است (پیاده‌سازی جاواپایه منبع باز آن، Hadoop است). این فن‌آوری، نیروی محرکه پردازش داده‌های توزیعی برای مدت زمان زیاد بوده و راه حل‌های ارایه شده توسط آن، اکنون به مرحله بلوغ رسیده است. شناخت نحوه به‌کارگیری این سامانه‌ها در پردازش حجم بالای داده‌های تولیدی کاربران، می‌تواند کمک زیادی در حل مسایل مبتلابه این داده‌ها باشد.

اگر چه MapReduce، «موتور» مستقر بر روی خوشه‌های رایانه‌ها است، خالص‌ترین فرم آن برای نوشتن روتین‌های تحلیلی مناسب نیست. اگر چه بسیاری از عملیات روی داده‌های رسانه‌های اجتماعی را می‌توان به‌طور مستقیم برای ساده‌ترین قالب‌های MapReduce نوشت، حرکت به سمت برنامه‌ریزهای اجرای سطح بالاتر که در آنها، بیان این عملیات به‌صورت طبیعی‌تر و نزدیک‌تر به طرز تفکر روزانه مقدور است، مفید می‌باشد. یکی از این قالب‌ها، Salding است که توان استفاده از زبان برنامه‌سازی Scala برای ساخت مراحل پردازش تحلیل داده‌های فعالیت‌های رسانه‌های اجتماعی را تسهیل می‌کند. با Scalding، ایده‌ها و الگوهای طراحی نهشته برای انجام محاسبات صحیح، ولی تقریبی، بر روی مجموعه داده‌های خروجی از سرویس‌های رسانه‌های اجتماعی، ارایه خواهد شد.

الزامات سیستم اجرای مثال‌ها

مثال‌های نشان داده شده در این کتاب، بر روی سیستم عملیات لینوکس Ubuntu، روایت 18.04 LTS، اجرا شده است. اگر از سیستم عملیات دیگری، خصوصاً ویندوز، استفاده می‌شود، توصیه، پیکره‌بندی محیط توسعه در ماشین مجازی با Untubu 18.04 LTS به عنوان مهمان، یا کاوش یکی دیگر از سرویس‌های میزبانی کلاود برای تعبیه نمونه آغازگری شده با این سیستم است.

پس از به‌دست آوردن مخزن کد منبع ارایه شده با این کتاب (به «مخزن برخط برای کتاب» در انتهای این فصل رجوع شود)، آن را می‌توان به فولدر مورد نظر استخراج نموده و در آن فولدر، setup/setup.sh را اجرا نمود. پس از اجرا، سیستم مورد نیاز، بسته‌های R و پایتون نصب شده تا بتوان کد منبع ارایه شده در این کتاب را اجرا کرد.

علاوه بر این، همان‌گونه که قبلاً اشاره شد، اجرای اسکریپت data/download_all.sh نیز برای موجود شدن فایل‌های داده‌ها برای کار مثال‌ها روی آنها لازم است؛ قبل از اجرای هر کدام از مثال‌ها، این را انجام دهید. همان‌گونه که اشاره شد، دانلودها ۶۰ گیگابایت فضا روی دیسک نیاز دارند.

مرور فصول

سازمان‌دهی کتاب، حول کاوش و شناخت واحدهای ساختمانی اصلی سامانه‌های رسانه‌های اجتماعی بوده که به‌صورت *چه کسی*، *چگونه*، *چه موقع*، و *چه‌چیز* فرایندهای رسانه‌های اجتماعی ساده‌سازی شده است. با توجه به اینکه رسانه‌های اجتماعی، ذاتاً در مورد جمع شدن افراد حول سایت‌های مختلف برای گفتگو، تفریح، و به اشتراک‌گذاشتن است، به این موضوعات از دیدگاه کاربران نگریسته می‌شود. کاربران چه افرادی هستند؟ چگونه با یکدیگر ارتباط برقرار می‌کنند؟ چه موقع درگیر می‌شوند؟ و نهایتاً، محتوایی که آنها به‌صورت جمعی ایجاد و مصرف می‌کنند، چیست؟

فصل ۱: *چه‌کسی* رسانه‌های اجتماعی. در این فصل، صحبت بر روی یکی از مهم‌ترین سوالاتی است که معمولاً در مورد کاربران یک سرویس مطرح می‌شود: کاربران چقدر فعال هستند؟ جوانب فراگیر فعالیت‌های انسانی خاص این سرویس‌ها، و علت تفاوت گسترده واقع شده بین کاربران، مورد کاوش قرار گرفته، و با سنجه‌های ویکی‌پدیا و توییتر، پشتیبانی می‌گردد.

فصل ۲: شبکه‌ها: *چگونگی* رسانه‌های اجتماعی. در این فصل یکی دیگر از تسهیلات ارایه شده توسط سرویس‌های رسانه‌های اجتماعی شرح داده می‌شود: شبکه اجتماعی. در بعضی از مواقع، این اصطلاح به خودی خود برای کل سرویس استفاده می‌شود؛ با این حال، در اینجا تمرکز بر روی گراف ارتباط جهت‌مند (مشاهده شده در ویکی‌پدیا، توییتر، و LiveJournal) و نوع نظمی که در آن می‌توان کشف کرد، است.

فصل ۳: فرایندهای زمانی: *چه‌موقع* رسانه‌های اجتماعی. این فصل در مورد زمانی است که وقایع رخ می‌دهند. داده‌های زمانی از ویراست‌های توییت و ویکی‌پدیا جمع‌آوری شده، و مشاهده می‌شود مهرهای زمانی چه چیزی در مورد رفتار کاربران روی این سایت‌ها، بیان می‌کنند. هم‌چنین، نتایج، با انتظارات پایه مفروض در سیستم‌های پویا مقایسه شده و روش‌های پایه انجام پیش‌پابی‌های سری زمانی معرفی خواهد شد.

فصل ٤: محتوا: *چیستی* رسانه‌های اجتماعی. در این فصل، رویکردهای سطح پایین و سطح بالا، بررسی شده، تا با استفاده از روش‌های پردازش زبان طبیعی، مشخص شود افراد در پست‌های متنی خود راجع به چه چیزهایی صحبت می‌کنند. ورای آماده‌سازی داده‌های متنی پایه، ویژگی‌های آماری متن شرح داده شده، و با استفاده از چندین الگوریتم، سعی می‌شود موضوعات مطرح شده در پست‌های جمع‌آوری شده از سوال و جواب‌های Stack Exchange و

در مقالات منتشر شده در Cora، مورد کاوش قرار گیرد.

فصل ۵: پردازش مجموعه داده‌های بزرگ. در این فصل، چالش‌های تحلیل مجموعه داده‌های حجیم در مطالعات رسانه‌های اجتماعی مطرح خواهد شد. پس از بررسی MapReduce، مثال‌های کد با استفاده از قالب Scalding مشاهده خواهد شد تا الگوهای برنامه‌سازی عام در حین کار با مجموعه داده‌های متأثر از رفتار انسانی، نمایان گردد. در بسط موضوع، نحوه استفاده از الگوریتم‌های تقریبی مطرح می‌شود که در آن، هدف دست‌یابی سریع به نتایج، ولی در محدوده خطاهای معلوم حول نتیجه صحیح، است. روش پیکره‌بندی خوشه‌های رایانه‌ها روی سرویس کلاود برای اجرای قالب‌های موازی، نیز به‌طور اجمالی مطرح خواهد گردید.

فصل ٦: یادگیری، نگاشت، و توصیه. این فصل در مورد انجام توصیه به کاربران و نشان دادن روش‌های یادگیری ماشین برای پیش‌بینی گرایش افراد به فیلم‌ها و برای ارزیابی نتایج پیش‌بینی است. مدل نیز مورد بررسی قرار گرفته تا مشخص شود چیزی در مورد نحوه رده‌بندی آیتم‌ها (عناوین فیلم‌ها) وجود دارد یا خیر.

فصل ۷: نتیجه‌گیری‌ها. در فصل نهایی، نگاهی عمیق‌تر به الگوهای آماری کلی برآمده از مسایل مختلف در سرتاسر کتاب صورت گرفته، و نحوه استفاده از روش‌های تحلیلی مشابه برای شناخت آنها مطرح می‌گردد.

مخزن برخط

وب‌سایت کتاب در wiley.com، حاوی همه فایل‌های کد منبع برای مثال‌ها، مطالب پشتیبان، و به‌روزرسانی‌ها است. برای پیداکردن آنها، عنوان کتاب یا ISBN آن (978-1-118-82485-6) را روی wiley.com جستجو کرده، و پس از قرار گرفتن در صفحه اصلی، به محل Downloads بروید.