

# مقدمه‌ای بر بازیابی اطلاعات

## تألیف:

کریستوفر دی. مینینگ

پرباکر رِگِن

هنریک شوتزه

## ترجمه:

دکتر هدیه ساجدی (عضو هیأت علمی دانشگاه تهران)

مهندس زهرا سادات تقوی

فرناز سادات تقوی

نیاز دانش

## مقدمه مترجمین

بازیابی اطلاعات به فن آوری و دانش پیچیده جستجو و استخراج اطلاعات، داده‌ها و فراداده‌ها در انواع گوناگون منابع اطلاعاتی اطلاق می‌شود. با افزایش روزافزون حجم اطلاعات ذخیره‌شده در منابع مختلف، فرآیند بازیابی و استخراج اطلاعات اهمیت ویژه‌ای پیدا کرده است. برخلاف پایگاه داده‌ها، اطلاعات ذخیره شده در منابع اطلاعاتی بزرگ مانند وب و زیرمجموعه‌های آن مانند شبکه‌های اجتماعی، از ساختار مشخصی پیروی نمی‌کنند و عموماً دارای معانی تعریف شده و مشخصی نیستند. هدف بازیابی اطلاعات در چنین شرایطی، کمک به کاربر برای یافتن اطلاعات مورد نظر در انبوهی از اطلاعات بدون ساختار است.

کتاب مقدمه‌ای بر بازیابی اطلاعات اولین کتاب با درک منسجمی از بازیابی اطلاعات سنتی و وب است که شامل جستجوی وب و حوزه‌های مرتبط با دسته‌بندی و خوشه‌بندی متن است. یکی از دلایل ترجمه این کتاب، انسجام خوب آن در این حیطه بود که توسط اساتید متخصص دانشگاهی نوشته شده است و به خوبی می‌تواند نیاز خوانندگان را مرتفع کند. از طرفی تلاش ما بر این بود که کتاب درسی منسجمی به زبان فارسی ارائه شود که به تحولات نسبتاً زیاد حیطه بازیابی اطلاعات پرداخته باشد.

در بخش پیشگفتار، مؤلفان کتاب، ویژگی‌های کتاب را به خوبی مطرح کرده‌اند و نیازی به تکرار آنها نیست. شایان توجه است که این کتاب در مجموع، با قلمی روان طیف قابل قبولی از مطالب را در زمینه بازیابی اطلاعات مطرح کرده است. به طوریکه طبق مصوبات شورای عالی برنامه‌ریزی وزارت علوم، تحقیقات و فن آوری تقریباً کلیه سرفصل‌های درس کارشناسی و کارشناسی ارشد بازیابی اطلاعات پیشرفته را پوشش می‌دهد.

در ترجمه این کتاب بر امانت‌داری، صحت و روانی ترجمه تاکید شده است. از ابداع واژه‌های جدید و مترادف واژه‌های موجود پرهیز و سعی شده از منابع موجود استفاده شود. مجموعه واژگان (فارسی به انگلیسی و انگلیسی به فارسی) در انتهای کتاب آورده شده است.

در اینجا لازم می‌دانیم تشکر ویژه‌ای از اعضای خانواده داشته باشیم. بدیهی است انجام اینکار بدون حمایت و همراهی آنها عملی نبود.

به یقین ترجمه حاضر خالی از اشکال و خطا نیست. بنابراین از تمامی عزیزانی که این کتاب را مطالعه می‌کنند، تقاضا داریم که ما را از نظرات و انتقادات خود از طریق ناشر مطلع نمایند. در انتها امیدواریم که مطالب این کتاب بتواند پاسخگوی نیاز اساتید، دانشجویان و افراد علاقه‌مند به حیطه بازیابی اطلاعات بوده و موجبات تعالی دانش آنها را فراهم آورد.

مترجمین

زمستان ۱۳۹۳

## پیشگفتار

مطالعات دهه ۱۹۹۰ نشان دادند که اکثر افراد ترجیح می‌دادند که اطلاعات را از یکدیگر دریافت کنند تا اینکه از سیستم‌های بازیابی اطلاعات بگیرند. البته، در آن دوره زمانی، اکثر افراد از دفاتر آژانس‌های مسافرتی نیز برای رزرو سفر خود استفاده می‌کردند. هر چند، در طول دهه گذشته، بهینه‌سازی‌ها در راستای اثربخشی بازیابی اطلاعات، موتورهای جستجوی وب را ارتقاء داده، طوری که اکثر افراد بیشتر اوقات از آن راضی بوده و جستجوی وب به یک استاندارد تبدیل شده و اغلب مرجع یافتن اطلاعات گشته است. برای مثال، Pew Internet Survey در سال ۲۰۰۴ (Fallows (2004)) بیان کرد که «۹۲٪ کاربران اینترنت بیان می‌کنند که اینترنت جای خوبی برای گرفتن اطلاعات روزانه است». حوزه بازیابی اطلاعات برای بسیاری از افراد جالب توجه است و از اطلاعات سیستم دانشگاهی تا مبانی اساسی دستیابی به اطلاعات را تحت تأثیر قرار داده است. این کتاب پشتیبانی علمی در این زمینه را در سطحی که برای دانشجویان و مقاطع بالاتر قابل استفاده باشد، فراهم می‌آورد.

بازیابی اطلاعات همزمان با وب آغاز نشد: در پاسخ به چالش‌های متعدد در فراهم آوردن دسترسی به اطلاعات، حوزه بازیابی اطلاعات برای ارائه روش‌های جستجو برای صورت‌های مختلف محتوا مطرح شد. این حوزه با انتشارات علمی و رکوردهای کتابخانه‌ای آغاز گشت، اما خیلی زود به صورت‌های دیگر محتوا، به ویژه آن صورت‌هایی که مورد استفاده متخصصان است مانند، روزنامه-نگاران، موکلان، و پزشکان است بسط پیدا کرد. اکثر تحقیقات علمی در بازیابی اطلاعات روی این محتواها رخ داده است و اکثر کاربردهای ادامه‌دار از بازیابی اطلاعات، به فراهم آوردن دسترسی به اطلاعات بدون ساختار در شرکت‌ها و دامنه‌های دولتی مختلف سروکار داشته و این کاربردها و مباحث بخش اساسی این کتاب را تشکیل می‌دهد.

با این حال، در سال‌های اخیر، محرک اصلی بازیابی اطلاعات، شبکه جهانی وب بوده است که در مقیاس ده‌ها میلیون تولیدکننده محتوا عمل کرده است. اگر اطلاعات بازیابی و تفسیر و تحلیل نمی‌شد طوریکه کاربر بتواند به سرعت اطلاعاتی مرتبط به نیازش را بیابد این انفجار اطلاعات منتشر شده مطرح نمی‌شد. در اواخر دهه ۱۹۹۰، افراد زیادی دریافتند که ادامه شاخص‌گذاری کل وب، به دلیل رشد نمایی وب از نظر اندازه، به سرعت غیرممکن خواهد شد. اما بخش اعظم نوآوری‌های علمی و مهندسی، کاهش سریع قیمت سخت افزار کامپیوتر و افزایش تجارت تحت جستجوی وب، همگی به

قدرت امروزی اکثر موتورهای جستجو افزودند. این موتورها قادر هستند، نتایج با کیفیت بالا را در زمان پاسخ کمتر از یک ثانیه برای هزاران میلیون جستجو در روز، روی میلیاردها صفحه وب فراهم کنند.

### سازماندهی کتاب و توسعه دوره

این کتاب نتیجه یک سری دوره‌هایی است که در دانشگاه استنفورد و دانشگاه اشتوتگارت در بازه زمانی یک نیم ترم، یک ترم کامل و دو نیم ترم تدریس کرده‌ایم. این دوره‌ها بیشتر برای دانشجویان در اوایل دوره کارشناسی ارشد علوم کامپیوتر برگزار شده است. اما دانشجویان وکالت، پزشکی، آمار، زبان‌شناسی و گرایش‌های مختلف مهندسی را نیز شامل شده است. از اینرو، اصل کلیدی طراحی این کتاب، پوشش مطالبی بود که اعتقاد داریم در دوره یک ترم تحصیلی در بازیابی اطلاعات مهم است. اصل دیگر ایجاد هر فصل براساس مطالبی است که می‌توان آن را در ارائه درس ۷۵ تا ۹۰ دقیقه‌ای پوشش داد.

هشت فصل اول کتاب، به مبانی بازیابی اطلاعات و به ویژه قلب موتورهای جستجو تخصیص داده شده، که ما این مطالب را اساس هر دوره‌ای در بازیابی اطلاعات می‌دانیم. فصل ۱، شاخص‌های وارونه را معرفی می‌کند و نشان می‌دهد که چگونه پرس و جوهای بولی ساده می‌تواند با استفاده از چنین شاخص‌هایی پردازش شود. فصل ۲، این مقدمه را با ارائه جزئیات روشی که در آن اسناد قبل از شاخص‌گذاری پیش پردازش شده و با بحث راجع به بکارگیری شاخص وارونه به شیوه‌های مختلف برای افزایش سرعت و عملیاتی بودن، بسط می‌دهد. فصل ۳، ساختارهای جستجو برای لغت‌نامه‌ها و چگونگی پردازش پرس و جوهایی که خطاهای املائی داشته و دیگر تطبیق‌های غیردقیق به مجموعه واژگان در مجموعه اسنادی که در حال جستجو است را مورد بحث قرار می‌دهد. فصل ۴، شماری از الگوریتم‌ها را برای ساخت شاخص وارونه از مجموعه متن با توجه ویژه به الگوریتم‌های توزیع شده و مقیاس‌پذیر توصیف می‌کند، که می‌تواند در مجموعه‌های بسیار بزرگ به کار گرفته شوند. فصل ۵، روش‌های فشرده‌سازی لغت‌نامه و شاخص‌های وارونه را پوشش می‌دهد. این روش‌ها، برای دستیابی به زمان‌های پاسخ کمتر از یک ثانیه برای پرس و جوهای کاربر در موتورهای جستجوی بزرگ، ضروری هستند. شاخص‌ها و پرس و جوهایی که در فصول ۱ تا ۵ مورد توجه قرار گرفتند، تنها با بازیابی بولی سروکار دارند که در آن یک سند یا با یک پرس و جو تطبیق دارد یا تطبیق ندارد. تمایل به اندازه‌گیری میزانی که یک سند با یک پرس و جو تطبیق دارد، یا نمره یک سند برای یک پرس و جو، موجب طرح وزن‌دهی عبارت و محاسبات نمره‌ها در فصول ۶ و ۷ می‌شود و به ایده‌ی لیستی از اسناد منتهی می‌شود که برای پرس و جو رتبه بندی شده‌اند. فصل ۸، بر ارزیابی سیستم بازیابی اطلاعات مبتنی بر ربط سندی که آن را بازیابی می‌کند، تمرکز کرده و به ما اجازه می‌دهد تا عملکردهای نسبی سیستم‌های مختلف را براساس مجموعه‌های سند و پرس و جو که برای محک و ارزیابی سیستم‌های بازیابی اطلاعات ارائه شده‌اند مقایسه کنیم.

فصول ۹ تا ۲۱ بر مبنای هشت فصل اول، انواع موضوعات پیشرفته‌تر را پوشش می‌دهند. فصل ۹ روش‌هایی را بررسی می‌کند که توسط آن بازیابی می‌تواند از طریق روش‌هایی مانند، بازخورد ربط و توسعه پرس‌وجو، که برای افزایش درست‌نمایی بازیابی اسناد مرتبط تلاش می‌کنند، بهبود یابد. فصل ۱۰، بازیابی اطلاعات از اسنادی را در نظر می‌گیرد که با زبان‌های نشانه‌گذاری مانند XML و HTML ساختار یافته‌اند. بازیابی ساخت یافته را به وسیله کاهش آن به روش‌های نمره‌گذاری فضای بردار که در فصل ۶ مطرح شده ارائه می‌دهیم. فصول ۱۱ و ۱۲، به نظریه احتمالاتی برای محاسبه نمره‌های سند در پرس و جو استناد می‌کند. فصل ۱۱، بازیابی اطلاعات احتمالاتی سنتی را توسعه می‌دهد که چارچوبی را برای محاسبه احتمال ربط یک سند با معلوم بودن مجموعه عبارات پرس و جو فراهم می‌آورد. از اینرو، این احتمال می‌تواند به عنوان نمره‌ای در رتبه بندی استفاده شود. فصل ۱۲، راه حل دیگری را شرح می‌دهد که در آن برای هر سندی در مجموعه، یک مدل زبانی می‌سازیم که از آن می‌توان احتمالی که مدل زبانی یک پرس و جوی مشخص را ایجاد کند، تخمین بزنیم. این احتمال کمیت دیگری است که با آن می‌توان اسناد را رتبه بندی کرد.

فصول ۱۳ تا ۱۷ صورت‌های مختلف یادگیری ماشین و روش‌های عددی در بازیابی اطلاعات را بیان می‌کنند. فصول ۱۳ تا ۱۵ مسئله دسته بندی اسناد را با دانستن مجموعه اسناد و دسته‌هایی که به آن تعلق دارند، بیان می‌کند. فصل ۱۳، دسته بندی آماری را به عنوان یکی از تکنولوژی‌های کلیدی برای موتور جستجو مطرح کرده و Naive Bayes را که روش ساده و کارآمد دسته بندی متن است، معرفی می‌کند و روش استاندارد را برای ارزیابی دسته بندی متن طرح ریزی می‌کند. فصل ۱۴، مدل فضای بردار را از فصل ۶ اتخاذ کرده و دو روش دسته بندی Rocchio و k-نزدیکترین همسایه (kNN) را معرفی می‌کند. همچنین، مصالحه بایاس-واریانس را به عنوان ویژگی مهمی از مسائل یادگیری مطرح می‌کند که ضابطه‌ای را برای انتخاب یک روش مناسب برای مسئله دسته بندی متن فراهم می‌آورد. فصل ۱۵، ماشین‌های بردار پشتیبان را معرفی می‌کند که بسیاری از محققان آن را در حال حاضر به عنوان کارآمدترین روش دسته بندی معرفی می‌کنند. همچنین، روابطی را بین مسئله دسته‌بندی و موضوعات به ظاهر مجزا، مانند استنتاج توابع نمره‌گذاری از مجموعه مثال‌های آموزشی، ارائه می‌دهیم.

فصول ۱۶ تا ۱۸ مسئله ایجاد خوشه‌ها را از اسناد مرتبط در یک مجموعه در نظر می‌گیرند. در فصل ۱۶، ابتدا مروری بر شماری از کاربردهای مهم خوشه‌بندی در بازیابی اطلاعات خواهیم داشت. سپس دو الگوریتم خوشه بندی را معرفی می‌کنیم: الگوریتم K-means که به طور گسترده و مؤثر برای خوشه‌بندی اسناد به کار می‌رود؛ و الگوریتم بیشینه‌سازی امیدریاضی<sup>۱</sup> که از نظر ریاضی پرهزینه‌تر بوده اما منعطف‌تر می‌باشد. فصل ۱۷، خوشه‌بندی سلسله مراتبی ساخت یافته را در بسیاری از کاربردهای بازیابی اطلاعات مطرح کرده و شماری از الگوریتم‌های خوشه‌بندی که سلسله مراتب خوشه‌ها را تولید می‌کنند، معرفی می‌کند. این فصل، همچنین مسئله دشوار محاسبه خودکار برچسب‌ها

<sup>۱</sup> Expectation Maximization

را برای خوشه‌ها مطرح می‌کند. فصل ۱۸، به طرح روش‌های جبرخطی که در خوشه‌بندی می‌توانند مورد استفاده قرار گیرند پرداخته و همچنین کاربردهای روش‌های جبری در بازیابی اطلاعات را مطرح می‌کند که در روش شاخص‌گذاری معنایی نهان بکار گرفته شده‌اند.

فصول ۱۹ تا ۲۱، مسئله جستجوی وب را در نظر می‌گیرند. در فصل ۱۹، بطور خلاصه به چالش‌های پایه در جستجوی وب به همراه مجموعه‌ای از روش‌ها در بازیابی اطلاعات وب پرداخته می‌شود. سپس، فصل ۲۰، معماری و الزامات پیمایشگر وب را توصیف می‌کند. در نهایت فصل ۲۱، قدرت تحلیل پیوند<sup>۱</sup> را در جستجوی وب در نظر می‌گیرد که در این روال، چندین روش را از جبر خطی گرفته تا نظریه احتمالات پیشرفته به کار می‌گیرد.

این کتاب، تمامی عناوین مرتبط با بازیابی اطلاعات را بطور جامع پوشش نمی‌دهد. ما شماری از موضوعات را کنار گذاشتیم زیرا خارج از محدوده مطالبی است که تمایل به پوشش آنها در مقدمه‌ای بر بازیابی اطلاعات داشتیم. با این حال، برای افرادی که علاقه‌مند به این موضوعات هستند، چندین کتاب معرفی می‌کنیم.

**بازیابی اطلاعات بین‌زبانی؛** Grossman and Frieder (2004)، فصل ۴؛ و Oard and Dorr (1996).  
**بازیابی اطلاعات تصویر و چندرسانه‌ای؛** Grossman and Frieder (2004)، فصل ۴؛  
 Baeza-Yates and Ribeiro-Neto (1999)، فصل ۶؛ Baeza-Yates and Ribeiro-Neto (1999)،  
 فصل ۱۱؛ Baeza-Yates and Ribeiro-Neto (1999)، فصل ۱۲؛ del Bimbo (1999)؛ Lew (2001)؛  
 و Smeulders et al. (2000).

**بازیابی صوت؛** Coden et al. (2002).

**بازیابی موسیقی؛** Downie (2006) و <http://www.ismir.net/>.

**واسط‌های کاربری برای بازیابی اطلاعات؛** Baeza-Yates and Ribeiro-Neto (1999)، فصل ۱۰.  
**بازیابی اطلاعات موزی و نظیر به نظیر؛** Grossman and Frieder (2004)، فصل ۷؛  
 Baeza-Yates and Ribeiro-Neto (1999)، فصل ۹؛ و Aberer (2001).

**کتابخانه‌های دیجیتال؛** Baeza-Yates and Ribeiro-Neto (1999)، فصل ۱۵؛ و Lesk (2004).  
**چشم‌انداز علوم اطلاعات؛** Korfhage (1997)، al. Meadowet (1999)، و  
 Ingwersen and Järvelin (2005).

**روش‌های مبتنی بر منطق، برای بازیابی اطلاعات؛** van Rijsbergen (1989).

**روش‌های پردازش زبان طبیعی؛** Manning and Schütze (1999)، Jurafsky and Martin (2008)؛  
 و Lewis and Jones (1996).

### پیش‌نیازها

دوره‌های مقدماتی در ساختمان داده‌ها و الگوریتم‌ها، جبر خطی و نظریه احتمالات به عنوان پیش‌نیاز برای تمامی ۲۱ فصل کفایت می‌کند.

<sup>1</sup> Link Analysis

فصول ۱ تا ۵، دوره مبانی الگوریتم و ساختمان داده‌ها را به عنوان پیش نیاز در نظر می‌گیرد. به علاوه، فصول ۶ و ۷ به دانش جبرخطی پایه شامل بردارها و ضرب نقطه‌ای نیاز دارند. هیچ پیش نیاز دیگری تا فصل ۱۱ لازم نیست. در فصل ۱۱ دوره مبانی در نظریه احتمالات مورد نیاز می‌باشد؛ بخش ۱-۱۱ مرور کوتاهی دارد بر مفاهیم مورد نیاز فصول ۱۱ تا ۱۳. فصل ۱۵ فرض می‌کند که خواننده با مفهوم بهینه‌سازی غیرخطی آشنا است، اگر چه این فصل ممکن است بدون دانش دقیق در مورد الگوریتم‌های بهینه‌سازی غیرخطی نیز مطالعه شود. فصل ۱۸ نیازمند دوره مقدماتی در جبرخطی شامل آشنایی با مفاهیم رتبه ماتریس و بردار مشخصه است که مرور مختصری بر آن در بخش ۱۸-۱ ارائه شده است. دانش مرتبط با مقادیر ویژه و بردارهای ویژه نیز در فصل ۲۱ مورد نیاز است. سطح دشواری تمرینات با آسان (\*)، متوسط (\*\*\*)، و یا دشوار (\*\*\*\*) مشخص می‌گردد.

### تقدیر و تشکر

از انتشارات دانشگاه کمبریج تشکر می‌کنیم که به ما اجازه دادند پیش‌نویسی از کتاب را به صورت آنلاین ایجاد کنیم و بازخوردی از کتاب در حین نگارش آن به دست آوریم. همچنین از Lauren Cowles تشکر می‌کنیم. او ویراستار برجسته‌ای است و چندین دوره نظراتی در مورد هر فصل از نظر شیوه نگارش، سازماندهی و پوشش مطالب، ارائه کرد. وی در رسیدن ما به اهداف نگارش این کتاب نقش ویراستاری بسزایی داشته است.

ما مدیون افراد بسیاری هستیم که نظرات، پیشنهادات و اصلاحاتی بر مبنای نسخه پیش‌نویس این کتاب ارائه کرده‌اند.

افراد بسیاری بازخورد دقیق در مورد هر یک از فصول یا بنا به درخواست و یا براساس خواست خود ایجاد کرده‌اند و برای این امر قدردان آنها هستیم.

و در آخر از داوران که بر حسب کیفیت و کمیت نظراتی را فراهم آورده‌اند به خاطر تأثیر قابل توجه آنها بر محتوا و ساختار کتاب تشکر می‌کنیم. ما قدردانی خود را از Pavel Berkhin، Andrew Trotman، Torsten Suel، Byron Dom، Jamie Callan، Stefan Büttcher، بخش‌های پیش‌نویس اولیه فصول ۱۳، ۱۴، و ۱۵، براساس اسلایدهایی است که توسط Ray Mooney فراهم شده‌اند. با وجود اینکه مطالب اصلاحات زیادی داشته است، اما از مشارکت Ray در این سه فصل و به ویژه توصیف پیچیدگی زمانی در تمامی الگوریتم‌های دسته‌بندی متن قدردانی می‌کنیم.

مؤلفان از دانشگاه استنفورد و اشتوتگارت برای فراهم آوردن محیط دانشگاهی برای به بحث گذاشتن ایده‌ها و فرصت تدریس دوره‌هایی که از آن این کتاب نتیجه شده و در آن محتوایش اصلاح شده است، تشکر می‌کنند. کریستوفر دی. مینینگ از خانواده‌اش برای ساعات بسیاری که وی صرف کار بر روی این کتاب کرده است، تشکر کرده و امیدوار است که سال آینده وقت آزاد بیشتری را آخر هفته‌ها با آنها سپری کند. پرباکر رگون از خانواده‌اش برای پشتیبانی صبورانه‌شان در طول نگارش این کتاب قدردانی کرده و به Yahoo! برای فراهم آوردن محیط پرثمیری که در آن این کتاب کار شده

است، مدیون است. هنریک شوتزه از والدین، خانواده و دوستانش برای حمایت‌شان در طول نگارش این کتاب تشکر می‌کند.

### اطلاعات تماس و وب

این کتاب، وب سایتی با آدرس <http://informationretrieval.org> دارد. بر روی این وب سایت مجموعه‌ای از اسلایدها که برای هر فصل ایجاد شده قرار داده شده است و می‌تواند برای تدریس درس بازیابی اطلاعات بکار رود. ما از بازخوردها، اصلاحات و پیشنهادات استقبال می‌کنیم که می‌تواند از طریق [informationretrieval \(at\) yahogroups \(dot\) com](mailto:informationretrieval(at)yahogroups(dot)com) به تمامی مؤلفان فرستاده شود.

# فهرست مطالب

فصل ۱	بازیابی بولی	۲۱
۱-۱	مثالی از مسئله بازیابی اطلاعات	۲۳
۲-۱	اولین برداشت در ساخت شاخص وارونه	۲۷
۳-۱	پردازش پرس وجوهای بولی	۳۰
۴-۱	مدل بسط یافته بولی در مقابل بازیابی رتبه بندی شده	۳۵
۵-۱	مراجع و مطالعات آتی	۳۸
فصل ۲	مجموعه واژگان عبارات و لیست پست ها	۴۱
۱-۲	شرح و توصیف سند و کدگشایی دنباله ی کاراکتر	۴۲
۱-۱-۲	دستیابی به دنباله کاراکتر در یک سند	۴۲
۲-۱-۲	انتخاب واحد سند	۴۳
۲-۲	تعیین مجموعه واژگان عبارت	۴۴
۱-۲-۲	نشانه گذاری	۴۴
۲-۲-۲	حذف عبارات متعارف: کلمات توقف	۴۹
۳-۲-۲	نرمال سازی (دسته کردن هم ارزی عبارات)	۵۰
۴-۲-۲	ریشه گیری و مدخل گیری	۵۵
۳-۲	اشتراک سریعتر لیست پست ها از طریق پرش اشاره گرها	۵۹
۴-۲	پست های موقعیتی و پرس وجوهای اصطلاح	۶۲
۱-۴-۲	شاخص های دو کلمه ای	۶۲
۲-۴-۲	شاخص های موقعیتی	۶۴
۳-۴-۲	طرح های ترکیب	۶۶
۵-۲	مراجع و مطالعات آتی	۶۹

۷۳	فصل ۳ لغت‌نامه‌ها و بازیابی مقاوم
۷۴	۱-۳ ساختار جستجو برای لغت‌نامه
۷۶	۲-۳ پرس‌وجوهای جایگزین
۷۷	۱-۲-۳ پرس‌وجوهای جایگزین کلی
۷۹	۲-۲-۳ شاخص‌های $k$ -گرمی برای پرس‌وجوهای جایگزینی
۸۱	۳-۳ تصحیح املائی
۸۱	۱-۳-۳ پیاده‌سازی تصحیح املائی
۸۲	۲-۳-۳ صورت‌های تصحیح املائی
۸۲	۳-۳-۳ فاصله ویرایشی
۸۴	۴-۳-۳ شاخص‌های $k$ -گرمی برای تصحیح املائی
۸۶	۵-۳-۳ تصحیح املائی حساس به متن
۸۸	۴-۳ تصحیح آوایی
۸۹	۵-۳ مراجع و مطالعات آتی
۹۱	فصل ۴ ساخت شاخص
۹۲	۱-۴ مبانی سخت‌افزاری
۹۳	۲-۴ شاخص‌گذاری بلوکی مبتنی بر مرتب‌سازی
۹۷	۳-۴ شاخص‌گذاری درون حافظه‌ای تک‌گذره
۹۹	۴-۴ شاخص‌گذاری توزیع شده
۱۰۳	۵-۴ شاخص‌گذاری پویا
۱۰۵	۶-۴ انواع دیگر شاخص‌ها
۱۰۸	۷-۴ مراجع و مطالعات آتی
۱۱۱	فصل ۵ فشرده‌سازی شاخص
۱۱۲	۱-۵ ویژگی‌های آماری عبارات در بازیابی اطلاعات
۱۱۴	۱-۱-۵ قانون Heaps: تخمین تعداد عبارات
۱۱۵	۲-۱-۵ قانون Zipf: مدل‌سازی توزیع عبارات
۱۱۷	۲-۵ فشرده‌سازی لغت‌نامه
۱۱۷	۱-۲-۵ لغت‌نامه به صورت یک رشته

۱۱۹.....	۲-۲-۵	ذخیره‌سازی بلوکی
۱۲۲.....	۳-۵	فشرده‌سازی فایل پست‌ها
۱۲۲.....	۱-۳-۵	کدهای بایت متغیر
۱۲۴.....	۲-۳-۵	کدهای $\gamma$
۱۳۳.....	۴-۵	مراجع و مطالعات آتی

## فصل ۶ نمره‌گذاری، وزن‌دهی عبارات و مدل فضای بردار ۱۳۵.....

۱۳۶.....	۱-۶	شاخص‌های ناحیه‌ای و پارامتری
۱۳۸.....	۱-۱-۶	نمره‌گذاری وزندار ناحیه
۱۳۹.....	۲-۱-۶	یادگیری وزن‌ها
۱۴۱.....	۳-۱-۶	وزن بهینه‌ی $g$
۱۴۳.....	۲-۶	وزن‌دهی و فراوانی عبارت
۱۴۳.....	۱-۲-۶	فراوانی وارونه سند
۱۴۵.....	۲-۲-۶	وزن‌دهی $tf-idf$
۱۴۶.....	۳-۶	مدل فضای بردار برای نمره‌گذاری
۱۴۶.....	۱-۳-۶	ضرب نقطه‌ای
۱۴۹.....	۲-۳-۶	پرس‌وجو به عنوان بردار
۱۵۱.....	۳-۳-۶	محاسبه‌ی نمره‌های بردار
۱۵۳.....	۴-۶	توانع گوناگون $tf-idf$
۱۵۳.....	۱-۴-۶	مقیاس‌گذاری زیرخطی $tf$
۱۵۳.....	۲-۴-۶	نرمال‌سازی بیشینه $tf$
۱۵۴.....	۳-۴-۶	طرح‌های وزن‌دهی سند و پرس‌وجو
۱۵۵.....	۴-۴-۶	نرمال‌سازی محوری طول سند
۱۵۹.....	۵-۶	مراجع و مطالعات آتی

## فصل ۷ محاسبه‌ی نمره‌ها در یک سیستم کامل جستجو ۱۶۱.....

۱۶۱.....	۱-۷	رتبه‌بندی و نمره‌گذاری کارآمد
۱۶۳.....	۱-۱-۷	بازیابی غیردقیق $K$ سند برتر
۱۶۳.....	۲-۱-۷	حذف شاخص
۱۶۴.....	۳-۱-۷	لیست‌های قهرمان

۱۶۴.....	۴-۱-۷	مرتب‌سازی و نمره‌های کیفیت ایستا.....
۱۶۶.....	۵-۱-۷	مرتب‌سازی برخوردار.....
۱۶۷.....	۶-۱-۷	هرس خوشه‌ای.....
۱۶۹.....	۲-۷	مؤلفه‌های یک سیستم بازیابی اطلاعات.....
۱۷۰.....	۱-۲-۷	شاخص‌های لایه‌ای.....
۱۷۰.....	۲-۲-۷	تقریب عبارت - پرس‌وجو.....
۱۷۱.....	۳-۲-۷	طراحی توابع نمره‌گذاری و تجزیه.....
۱۷۳.....	۴-۲-۷	کنار هم قرار دادن همه بخش‌ها.....
۱۷۴.....	۳-۷	نمره‌گذاری فضای بردار و تعامل عملگر پرس‌وجو.....
۱۷۶.....	۴-۷	مراجع و مطالعات آتی.....

## فصل ۸ ارزیابی در بازیابی اطلاعات..... ۱۷۷

۱۷۸.....	۱-۸	ارزیابی سیستم بازیابی اطلاعات.....
۱۷۹.....	۲-۸	مجموعه‌های آزمایشی استاندارد.....
۱۸۱.....	۳-۸	ارزیابی مجموعه‌های بازیابی رتبه بندی نشده.....
۱۸۴.....	۴-۸	ارزیابی نتایج بازیابی رتبه بندی شده.....
۱۹۱.....	۵-۸	ارزیابی ربط.....
۱۹۳.....	۱-۵-۸	انتقادات و توجیهات مفهوم ربط.....
۱۹۵.....	۶-۸	یک چشم انداز وسیع‌تر: کیفیت سیستم و مطلوبیت کاربر.....
۱۹۵.....	۱-۶-۸	مسائل سیستمی.....
۱۹۶.....	۲-۶-۸	مطلوبیت کاربر.....
۱۹۷.....	۳-۶-۸	تصحیح سیستم ساخته شده.....
۱۹۸.....	۷-۸	بخش‌هایی از نتایج.....
۲۰۰.....	۸-۸	مراجع و مطالعات آتی.....

## فصل ۹ بازخورد ربط و گسترش پرس‌وجو..... ۲۰۳

۲۰۴.....	۱-۹	بازخورد ربط و شبه بازخورد ربط.....
۲۰۷.....	۱-۱-۹	الگوریتم Rocchio برای بازخورد ربط.....
۲۰۹.....	۲-۱-۹	بازخورد ربط احتمالاتی.....
۲۰۹.....	۳-۱-۹	چه زمانی بازخورد ربط کار می‌کند؟.....

۲۱۱.....	بازخورد ربط روی وب	۴-۱-۹
۲۱۲.....	ارزیابی استراتژی‌های بازخورد ربط	۵-۱-۹
۲۱۳.....	شبه بازخورد ربط	۶-۱-۹
۲۱۳.....	بازخورد ربط غیرمستقیم	۷-۱-۹
۲۱۴.....	خلاصه	۸-۱-۹
۲۱۵.....	روش‌های سراسری برای فرموله کردن پرس و جو	۲-۹
۲۱۵.....	ابزارهای مجموعه واژگان برای فرموله کردن مجدد پرس و جو	۱-۲-۹
۲۱۶.....	گسترش پرس و جو	۲-۲-۹
۲۱۸.....	تولید خودکار فرهنگ لغات جامع	۳-۲-۹
۲۱۹.....	مراجع و مطالعات آتی	۳-۹

## فصل ۱۰ بازایی XML..... ۲۲۱

۲۲۵.....	مفاهیم پایه‌ای XML	۱-۱۰
۲۲۸.....	چالش‌ها در بازایی XML	۲-۱۰
۲۳۳.....	یک مدل فضای بردار برای بازایی XML	۳-۱۰
۲۳۸.....	ارزایی بازایی XML	۴-۱۰
۲۴۲.....	بازایی XML متن-محور در مقابل داده-محور	۵-۱۰
۲۴۵.....	مراجع و مطالعات آتی	۶-۱۰

## فصل ۱۱ بازایی اطلاعات احتمالاتی..... ۲۴۷

۲۴۸.....	مروری بر نظریه احتمالات پایه	۱-۱۱
۲۴۹.....	اصل رتبه بندی احتمالاتی	۲-۱۱
۲۴۹.....	مورد اتلاف ۱/۰	۱-۲-۱۱
۲۵۰.....	اصل رتبه بندی احتمالاتی با هزینه‌های بازایی	۲-۲-۱۱
۲۵۰.....	مدل استقلال دودویی	۳-۱۱
۲۵۲.....	به دست آوردن تابع رتبه بندی برای عبارات پرس و جو	۱-۳-۱۱
۲۵۴.....	برآوردهای نظری احتمالاتی	۲-۳-۱۱
۲۵۵.....	برآورد عملی احتمالاتی	۳-۳-۱۱
۲۵۶.....	روش‌های احتمالاتی برای بازخورد ربط	۴-۳-۱۱
۲۵۹.....	یک ارزیابی و ارائه تعدادی نسخه	۴-۱۱

۲۵۹.....	۱-۴-۱۱	ارزیابی مدل‌های احتمالاتی
۲۶۰.....	۲-۴-۱۱	ساختار درختی برای وابستگی‌های بین عبارات
۲۶۰.....	۳-۴-۱۱	Okapi BM25: یک مدل غیردودویی
۲۶۳.....	۴-۴-۱۱	روش‌های شبکه بیزی برای بازیابی اطلاعات
۲۶۴.....	۵-۱۱	مراجع و مطالعات آتی

## فصل ۱۲ مدل‌های زبانی برای بازیابی اطلاعات ۲۶۵.....

۲۶۵.....	۱-۱۲	مدل‌های زبانی
۲۶۵.....	۱-۱-۱۲	آتاماتای متناهی و مدل‌های زبانی
۲۶۸.....	۲-۱-۱۲	انواع مدل‌های زبانی
۲۶۹.....	۳-۱-۱۲	توزیع چندجمله‌ای روی کلمات
۲۷۰.....	۲-۱۲	مدل درست‌نمایی پرس‌وجو
۲۷۰.....	۱-۲-۱۲	استفاده از مدل‌های زبانی درست‌نمایی پرس‌وجو در بازیابی اطلاعات
۲۷۲.....	۲-۲-۱۲	برآورد احتمال تولید پرس‌وجو
۲۷۴.....	۳-۲-۱۲	آزمایشات Ponte و Croft
۲۷۷.....	۳-۱۲	مدل‌سازی زبانی در مقابل روش‌های دیگر در بازیابی اطلاعات
۲۷۸.....	۴-۱۲	نظریه‌های مدل‌سازی زبانی توسعه یافته
۲۸۰.....	۵-۱۲	مراجع و مطالعات آتی

## فصل ۱۳ دسته‌بندی متن و Naive Bayes ۲۸۳.....

۲۸۶.....	۱-۱۳	مسئله دسته‌بندی متن
۲۸۸.....	۲-۱۳	دسته‌بندی متن Naive Bayes
۲۹۳.....	۱-۲-۱۳	رابطه با مدل زبانی یک-گرمی چندجمله‌ای
۲۹۴.....	۳-۱۳	مدل برنولی
۲۹۶.....	۴-۱۳	ویژگی‌های Naive Bayes
۳۰۱.....	۱-۴-۱۳	یک نوع از مدل چندجمله‌ای
۳۰۲.....	۵-۱۳	انتخاب ویژگی
۳۰۳.....	۱-۵-۱۳	اطلاعات متقابل
۳۰۵.....	۲-۵-۱۳	انتخاب ویژگی $\chi^2$
۳۰۸.....	۳-۵-۱۳	انتخاب ویژگی مبتنی بر فراوانی

۳۰۹.....	انتخاب ویژگی برای دسته‌بندهای چندگانه.....	۴-۵-۱۳
۳۰۹.....	مقایسه روش‌های انتخاب ویژگی.....	۵-۵-۱۳
۳۱۰.....	ارزیابی دسته‌بندی متن.....	۶-۱۳
۳۱۸.....	مراجع و مطالعات آتی.....	۷-۱۳

## فصل ۱۴ دسته‌بندی فضای بردار..... ۳۲۱

۳۲۳.....	نمایش سند و معیارهای وابستگی در فضاها بردار.....	۱-۱۴
۳۲۴.....	دسته‌بندی Rocchio.....	۲-۱۴
۳۲۹.....	k-نزدیکترین همسایه.....	۳-۱۴
۳۳۲.....	پیچیدگی زمانی و بهینگی k-نزدیکترین همسایه.....	۱-۳-۱۴
۳۳۴.....	دسته‌بندهای خطی در برابر دسته‌بندهای غیرخطی.....	۴-۱۴
۳۳۹.....	دسته‌بندی با بیشتر از دو دسته.....	۵-۱۴
۳۴۱.....	مصالحه بایاس-واریانس.....	۶-۱۴
۳۵۰.....	مراجع و مطالعات آتی.....	۷-۱۴

## فصل ۱۵ ماشین‌های بردار پشتیبان و یادگیری ماشین روی اسناد..... ۳۵۳

۳۵۴.....	ماشین‌های بردار پشتیبان: حالت قابل جداسازی به صورت خطی.....	۱-۱۵
۳۶۱.....	توسعه مدل ماشین بردار پشتیبان.....	۲-۱۵
۳۶۱.....	دسته‌بندی حاشیه‌ای نرم.....	۱-۲-۱۵
۳۶۴.....	دسته‌بندهای ماشین بردار پشتیبان چند دسته‌ای.....	۲-۲-۱۵
۳۶۵.....	دسته‌بندهای ماشین بردار پشتیبان غیرخطی.....	۳-۲-۱۵
۳۶۸.....	نتایج آزمایشگاهی.....	۴-۲-۱۵
۳۶۹.....	مسائل در دسته‌بندی اسناد متنی.....	۳-۱۵
۳۶۹.....	انتخاب نوع دسته‌بند برای استفاده.....	۱-۳-۱۵
۳۷۱.....	بهبود کارایی دسته‌بند.....	۲-۳-۱۵
۳۷۵.....	روش‌های یادگیری ماشین در بازیابی اطلاعات موردی.....	۴-۱۵
۳۷۶.....	مثال ساده‌ای از نمره‌گذاری یادگرفته شده توسط ماشین.....	۱-۴-۱۵
۳۷۸.....	رتبه بندی نتایج یادگیری ماشین.....	۲-۴-۱۵
۳۸۰.....	مراجع و مطالعات آتی.....	۵-۱۵

۳۸۵	فصل ۱۶ خوشه‌بندی مسطح
۳۸۷	۱-۱۶ خوشه‌بندی در بازیابی اطلاعات
۳۹۰	۲-۱۶ بیان مسئله
۳۹۲	۱-۲-۱۶ کاردینالیتی - تعداد خوشه‌ها
۳۹۲	۳-۱۶ ارزیابی خوشه‌بندی
۳۹۶	۴-۱۶ الگوریتم K-means
۴۰۱	۱-۴-۱۶ کاردینالیتی خوشه در K-means
۴۰۴	۵-۱۶ خوشه‌بندی مبتنی بر مدل
۴۱۱	۶-۱۶ مراجع و مطالعات آتی
۴۱۵	فصل ۱۷ خوشه‌بندی سلسله مراتبی
۴۱۶	۱-۱۷ خوشه‌بندی سلسله مراتبی تجمعی
۴۲۰	۲-۱۷ خوشه‌بندی Single-Link و Complete-Link
۴۲۳	۱-۲-۱۷ پیچیدگی زمانی
۴۲۶	۳-۱۷ خوشه‌بندی تجمعی Group-Average
۴۲۸	۴-۱۷ خوشه‌بندی Centroid
۴۳۰	۵-۱۷ بهینگی خوشه‌بندی سلسله مراتبی تجمعی
۴۳۴	۶-۱۷ خوشه‌بندی تقسیم‌کننده
۴۳۴	۷-۱۷ برچسب‌زنی خوشه
۴۳۶	۸-۱۷ نکات پیاده‌سازی
۴۳۹	۹-۱۷ مراجع و مطالعات آتی
۴۴۱	فصل ۱۸ تجزیه‌های ماتریس و شاخص‌گذاری معنایی نهان
۴۴۲	۱-۱۸ مروری بر جبرخطی
۴۴۴	۱-۱-۱۸ تجزیه‌های ماتریس
۴۴۶	۲-۱۸ ماتریس‌های عبارت-سند و تجزیه‌های مقدار منفرد
۴۴۹	۳-۱۸ تقریب‌های رتبه پایین
۴۵۱	۴-۱۸ شاخص‌گذاری معنایی نهان
۴۵۷	۵-۱۸ مراجع و مطالعات آتی

## فصل ۱۹ مبانی جستجوی وب ..... ۴۵۹

- ۴۵۹..... ۱-۱۹ پس زمینه و تاریخچه
- ۴۶۲..... ۲-۱۹ مشخصات وب
- ۴۶۳..... ۱-۲-۱۹ گراف وب
- ۴۶۵..... ۲-۲-۱۹ هرزنامه
- ۴۶۸..... ۳-۱۹ تبلیغات به عنوان مدل اقتصادی
- ۴۷۰..... ۴-۱۹ تجربه کاربر جستجو
- ۴۷۱..... ۱-۴-۱۹ نیازهای مربوط به پرس و جوی کاربر
- ۴۷۲..... ۵-۱۹ اندازه و برآورد شاخص
- ۴۷۶..... ۶-۱۹ اسناد دونه‌های و پوشاندن
- ۴۸۱..... ۷-۱۹ مراجع و مطالعات آتی

## فصل ۲۰ پیمایش وب و شاخص‌ها ..... ۴۸۳

- ۴۸۳..... ۱-۲۰ مرور کلی
- ۴۸۴..... ۱-۱-۲۰ ویژگی‌هایی که یک پیمایشگر موظف است فراهم آورد
- ۴۸۴..... ۲-۱-۲۰ ویژگی‌هایی که یک پیمایشگر باید فراهم بیاورد
- ۴۸۵..... ۲-۲۰ پیمایش کردن
- ۴۸۵..... ۱-۲-۲۰ معماری پیمایشگر
- ۴۹۰..... ۲-۲-۲۰ تحلیل DNS
- ۴۹۱..... ۳-۲-۲۰ URL آتی
- ۴۹۵..... ۳-۲۰ توزیع شاخص‌ها
- ۴۹۶..... ۴-۲۰ سرویس‌دهنده‌های اتصال
- ۴۹۹..... ۵-۲۰ مراجع و مطالعات آتی

## فصل ۲۱ تحلیل پیوند ..... ۵۰۱

- ۵۰۲..... ۱-۲۱ وب به عنوان یک گراف
- ۵۰۲..... ۱-۱-۲۱ لنگر متن و گراف وب
- ۵۰۴..... ۲-۲۱ رتبه صفحه
- ۵۰۶..... ۱-۲-۲۱ زنجیره‌های مارکوف

۵۰۸.....	۲-۲-۲۱	محاسبه رتبه صفحه.....
۵۱۱.....	۳-۲-۲۱	رتبه صفحه موضوع خاص.....
۵۱۶.....	۳-۲۱	قطبیت و نفوذ.....
۵۱۹.....	۱-۳-۲۱	انتخاب زیرمجموعه‌ای از وب.....
۵۲۲.....	۴-۲۱	مراجع و مطالعات آتی.....
۵۲۵.....		مراجع.....
۵۵۳.....		لیست اختصارات.....
۵۵۵.....		واژه‌نامه فارسی به انگلیسی.....
۵۷۰.....		واژه‌نامه انگلیسی به فارسی.....