ارائه مدلی جهت پیش‌بینی موضوعات مرتبط با هوشمندی کسب‌وکار

فاطمه عباسی
دانشجوی دکتری مدیریت فناوری اطلاعات، دانشکده مدیریت دانشگاه تهران، تهران، ایران[1]
بابک سهرابی
استاد گروه مدیریت فناوی اطلاعات دانشکده مدیریت دانشگاه تهران، تهران، ایران
آمنه خدیور
دانشیار گروه مدیریت دانشکده علوم اجتماعی و اقتصاد دانشگاه الزهرا(س)، تهران، ایران
امیر مانیان
استاد گروه مدیریت فناوی اطلاعات دانشکده مدیریت دانشگاه تهران، تهران، ایران

دوفصلنامه علمی- پژوهشی
مدیریت اطلاعات
دوره ۳، شماره۱ – شماره پیاپی۴، بهار و تابستان۱۳۹۶

چکیده: امروزه بسیاری از کالاها ازجمله کتاب از طریق اینترنت و به‌صورت آنلاین به فروش می‌رسد و به‌تبع آن لازم است اطلاعات دقیق‌تری در ارتباط با کالا در اختیار مشتری قرار داده شود تا تصمیم خریدش آگاهانه و هوشمندانه‌تر باشد. سایت آمازون، وب‌سایتی است که اطلاعات کتاب‌ها را برای خرید آنلاین در اختیار کاربران قرار می‌دهد. یکی از مشکلات در انتخاب کتاب از سایت آمازون مشخص نبودن زیرشاخه‌های مربوط به موضوع اصلی است. آگاهی از این زیر موضوع‌ها خریداران را در انتخاب هوشمندانه‌تر و درنتیجه خرید بهتر یاری می‌کند. در این تحقیق تلاش گردیده تا با داده‌های استخراج‌شده از سایت آمازون و با استفاده از روش‌های خوشه‌بندی و طبقه‌بندی موضوعات و زیر حوزه‌های مرتبط با هوشمندی کسب‌وکار به‌عنوان نمونه استخراج و درنهایت مدلی جهت پیش‌بینی موضوعات مرتبط با هوشمندی کسب‌وکار ارائه گردد. با استفاده از مدل پیش‌بینی، با ارائه عنوان و مقدمه کتاب‌های مرتبط با هوشمندی کسب‌وکار زیر موضوع مرتبط با کتاب پیش‌بینی می‌گردد. نتایج تحلیل نشان می‌دهد هفت خوشه مرتبط با هوشمندی کسب‌وکار به ترتیب ابزارهای هوشمندی کسب‌وکار و مصورسازی، رفتار سازمانی، مدیریت فرآیندها و دانش، سیستم‌های پشتیبانی تصمیم، رهبری، متن‌کاوی و پایگاه داده است. درنهایت درخت تصمیم و نزدیک‌ترین همسایه با دقت بالاتری مدل پیش‌بینی موضوعات مرتبط با هوشمندی کسب‌وکار را ارائه می‌نمایند.

کلیدواژه‌ها: خوشه‌بندی، داده ساختار نیافته، طبقه‌بندی، متن‌کاوی، هوشمندی کسب‌وکار.

۱. نویسنده مسئول: f_abbasi@ut.ac.ir

مقدمه

امروزه حجم داده‌ها با نرخ فزاینده‌ای در حال افزایش است. تقریباً تمام صنایع، سازمان‌ها و مؤسسات به‌صورت الکترونیکی داده‌ها خود را ذخیره می‌نمایند. حجم زیادی متن در قالب کتابخانه‌های دیجیتالی، مخازن و سایر اطلاعات متنی مانند وبلاگ‌ها، شبکه‌های اجتماعی و ایمیل‌ها از طریق اینترنت در جریان هستند (Sagayam 2012) . ازآنجاکه بالای ۸۰ درصد از اطلاعات به‌صورت متنی نگه‌داری می‌شوند باور بر آن است که متن‌کاوی ارزش تجاری بالقوه بالایی داشته باشد (Gupta and Lehal 2009) . ازآنجاکه پردازش این حجم عظیم داده به‌صورت دستی کاری طاقت‌فرسا است، روش‌های متن‌کاوی جهت تحلیل این اطلاعات ساختار نیافته موردنیاز است (پرئی و حمیدی ۱۳۹۵). متن‌کاوی با متن‌های زبان طبیعی مرتبط است که در فرمت‌های نیمه ساختاریافته و بدون ساختار ذخیره می‌شوند (Weiss et al. 2010). تکنیک‌های متن‌کاوی در صنعت، دانشگاه، برنامه‌های مبتنی بر وب و سایر حوزه‌ها کاربرد دارد (Liao, Chu and Hsiao 2012). برنامه‌های مبتنی بر وب مانند موتورهای جستجو، سیستم مدیریت روابط با مشتریان، فیلتر کردن ایمیل‌ها، تحلیل پیشنهاد محصول[1] تشخیص تقلب[2] تحلیل شبکه‌های اجتماعی[3] ازجمله حوزه‌هایی هستند که از متن‌کاوی برای عقیده کاوی[4] استخراج ویژگی‌ها[5] تحلیل احساسات[6] پیش‌بینی[7] و تحلیل روند[8] استفاده می‌نماید (Talib et al. 2016). متن‌کاوی یا تحلیل متن به کاربرد تکنیک‌های مختلفی برای استخراج اطلاعات مفید از مجموعه‌ای از مستندات اشاره دارد(Kumar 2016) . متن‌کاوی شاخه‌ای از حوزه داده‌کاوی است که سعی در دریافتن الگوهای جالب‌توجه از پایگاه داده‌های بزرگ دارد که کشف اطلاعات جدید یا استخراج خودکار اطلاعات از منابع مکتوب مختلف، به‌وسیله رایانه امکان‌پذیر می‌شود. متن‌کاوی که تحت عناوین تجزیه‌وتحلیل هوشمند متن، کاوش داده‌های متنی و کشف دانش از متون نیز شناخته می‌شود، به‌طورکلی به فرآیند استخراج اطلاعات و دانش جالب‌توجه و غیر بدیهی از متن بدون ساختار اشاره دارد (Gupta and Lehal 2009) .

آمازون یک فروشگاه آنلاین است که از طریق وب‌سایت www.amazon.com محصولات نو یا دست‌دوم را به مشتریان خود عرضه می‌نماید. وب‌گاه آمازون به‌عنوان بزرگ‌ترین کتاب‌فروشی دنیا شناخته می‌شود که در آن امکان ورق زدن و مطالعه نسخه الکترونیکی اغلب کتاب‌هایی که در این وب‌سایت برای فروش گذاشته‌شده است، وجود دارد. یکی از مسائل در انتخاب کتاب‌ها مشخص نمودن زیر حوزه‌ها یا زیر موضوعات مرتبط با موضوع اصلی کتاب است که جهت دستیابی به این نوع اطلاعات لازم است خواننده مقدمه و یا حتی فصل‌های کتاب را موردبررسی قرار دهد. به‌عنوان‌مثال در سایت آمازون با وارد نمودن

1.Product suggestion analysis
2.Fraud detection
3.Social media analytics
4.Opinion mining
5.Feature extraction
6.Sentiment
7.Predictive
8.Trend analysis

عنوان هوشمندی کسب‌وکار[1]کلیه کتاب‌هایی که در عنوان، این کلمه را دارند و در محتوا ممکن است به زیر حوزه‌های مختلفی چون داده‌کاوی، یا سیستم تصمیم‌یار[2] پرداخته باشند برای کاربر لیست شوند. درحالی‌که هدف کاربر از این جستجو یافتن کتابی در زمینه انباره داده، از زیر حوزه‌های مرتبط با هوشمندی کسب‌وکار است که استخراج این زیر حوزه جز با بررسی مقدمه و یا محتوای کتاب به دست نمی‌آید. مشکلی که امروزه اغلب کاربران با آن مواجه هستند کمبود وقت است که باعث می‌شود اغلب جزئیات اطلاعات کتاب را مطالعه نکنند و تنها بر اساس عنوان تصمیم‌گیری نمایند. عدم اشراف کامل به اطلاعات کتاب‌ها باعث می‌شود گاهی خریدها مطلوب کاربران نباشد و مطالب کتاب در حوزه‌ای باشد که مدنظر خریدار نیست. استفاده از تکنیک‌های متن‌کاوی جهت استخراج موضوعات و تم‌ها یکی از راهکارهای پیشنهادی جهت استخراج موضوعات و زیر حوزه‌های مرتبط با یک موضوع است. استخراج این حوزه‌ها و تم‌ها می‌تواند به کاربران جهت تصمیم‌گیری سریع‌تر و آگاهانه‌تر و به شرکت‌های فعال در حوزه تجارت الکترونیک برای ارائه بهتر خدمات و افزایش رضایت مشتریان کمک نماید. در این مقاله تلاش شده تا با استفاده از نرم‌افزار Rapidminer Studdio 7. 5. 003 و تکنیک‌های متن‌کاوی، مدلی جهت پیش‌بینی حوزه‌های مرتبط با هوشمندی کسب‌وکار ارائه گردد. روند انجام کار به این صورت است که در ابتدا عنوان و مقدمه کتاب‌هایی که مرتبط با هوشمندی کسب‌وکار هستند از سایت آمازون جمع‌آوری گردید. با استفاده از روش‌هایی چون ریشه‌یابی و حذف ایست واژه‌ها داده‌ها مراحل پیش‌پردازش و آماده‌سازی داده‌ها بر روی‌داده‌های گردآوری‌شده انجام شد. در مرحله مدل‌سازی ابتدا با استفاده از روش‌های خوشه‌بندی زیر حوزه‌های مرتبط با هوشمندی کسب‌وکار استخراج می‌شوند و سپس با استفاده از الگوریتم‌های دسته‌بندی کننده مدلی جهت پیش‌بینی زیر حوزه‌های مرتبط با موضوع هوشمندی کسب‌وکار ارائه می‌گردد که با وارد نمودن مقدمه و عنوان کتاب زیر حوزه‌ها و یا زیر موضوعات مرتبط با کتاب به کاربر ارائه می‌شود. ارائه این زیر حوزه‌ها به کاربر در انتخاب کتاب مرتبط با موضوع اصلی کمک می‌نماید. ساختار مقاله به‌صورت زیر سازمان‌دهی شده است:

در بخش اول مقدمه‌ای بر پژوهش صورت گرفته ارائه شد. در بخش دوم مروری بر ادبیات موضوع در این حوزه ارائه می‌شود. در بخش سوم روش‌شناسی پژوهش، در بخش چهارم یافته‌های پژوهش و در نهایت در بخش پنجم نتیجه‌گیری و پیشنهادها جهت تحقیقات آتی ارائه می‌گردد.

پیشینه نظری پژوهش

متن‌کاوی، فرآیند استفاده از کامپیوتر جهت بررسی و تحلیل داده‌های بدون ساختار (متنی) جهت استخراج اطلاعات است. فرآیند متن‌کاوی شامل سه‌گام اصلی گردآوری متن، انتقال متن و استخراج دانش است که

1.Business intelligence
2.Decision Support System (DSS)

هر یک از این گام‌ها بازیابی اطلاعات[1]، استخراج اطلاعات[2] و داده‌کاوی[3] نامیده می‌شوند (Olorisade, Brereton and Andras 2017). فرآیند متن‌کاوی شامل گام‌های زیر است:

- جمع‌آوری داده ساختار نیافته از منابع گوناگون در فرمت‌های HTML، متنی، PDF و غیره نخستین گام از فرآیند متن‌کاوی است.
- عملیات پاک‌سازی و آماده‌سازی داده‌ها جهت تشخیص و حذف ناهنجاری‌های[4] داده که پاک-سازی و آماده‌سازی داده‌ها جهت اطمینان از حذف بخش‌های غیرضروری متن و عدم وجود اطلاعات تکراری ضروری است.
- عملیات پردازش و کنترل که جهت کنترل و پاک‌سازی بیشتر متن بکار گرفته می‌شود.
- تحلیل الگو که جهت تشخیص الگوهای پنهان در متن و کشف دانش استفاده می‌شود (Talib et al. 2016).

به‌طور خاص متن‌کاوی شامل فعالیت‌های مختلفی چون دسته‌بندی متن به دو یا چندطبقه (طبقه‌بندی متن[5])، گروه‌بندی متن‌های مشابه با یکدیگر (خوشه‌بندی متن[6])، یافتن موضوع متن (استخراج موجودیت/مفاهیم)، یافتن احساس متن (تحلیل احساسات)، خلاصه‌سازی متن و یادگیری ارتباط بین موجودیت‌های متن (مدل‌سازی ارتباط موجودیت‌ها[7]) است (Truyens 2014).

متن‌کاوی مشابه داده‌کاوی است با این تفاوت که ابزارهای داده‌کاوی برای مدیریت داده‌های ساختارمند از پایگاه داده طراحی‌شده است. متن‌کاوی می‌تواند با مجموع داده‌های بدون ساختار یا نیمه ساختارمند مانند نامه‌های الکترونیکی، اسناد تمام متنی و پرونده‌های وب کار کند. درنتیجه، متن‌کاوی راه‌حل بهتری برای شرکت‌هاست. بااین‌حال تا به امروز، بیشتر چالش تحقیق و توسعه روی داده‌کاوی با استفاده از داده‌های ساختارمند متمرکز بوده‌اند (Weng and Lin 2003). در جدول زیر ارتباط میان داده‌کاوی و متن‌کاوی بیان‌شده است (پرئی و حمیدی ۱۳۹۵).

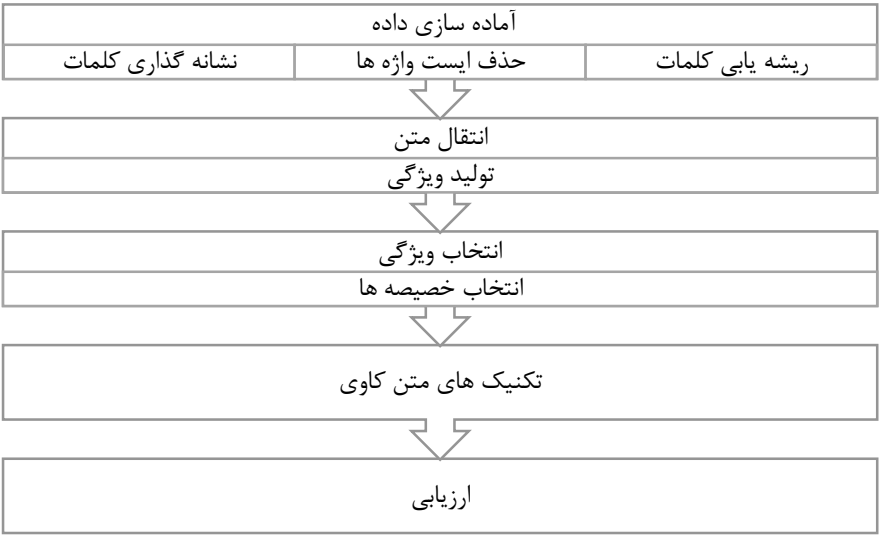
جدول ۱. ارتباط میان داده‌کاوی و متن‌کاوی

	جستجوی هدف مشخص	**کشف سازوکار مشخص**
داده‌های ساختاریافته	بازیابی داده‌ها	داده‌کاوی
داده‌های بدون ساختار (متنی)	بازیابی اطلاعات	متن‌کاوی

1.Information retrieval (IR)
2.Information Retrieval (IR)
3.Data mining
4.Anomalies
5.Text categorization
6.Text clustering
7.Entity relation modelling

برای داده‌کاوی روش‌شناسی‌های CRISP-DM[1] و SEMMA[2] جهت پیاده‌سازی داده‌کاوی پیشنهادشده است که از میان این دو، روش‌شناسی CRISP-DM محبوب‌تر است. به دلیل آنکه تمایز مهم میان داده بین داده‌کاوی و متن‌کاوی در فرآیند کشف دانش در نوع داده است روش‌شناسی CRISP-DM برای متن‌کاوی نیز موردقبول است(Miner, et al. 2012) . این روش‌شناسی شامل شش فاز درک کسب‌وکار (مشخص نمودن هدف)، درک داده (بررسی و درک داده)، آماده‌سازی داده، مدل‌سازی، ارزیابی و توسعه است (Chapman 2000) .

متن‌کاوی مشتمل بر سه مؤلفه اصلی بازیابی اطلاعات[3]، پردازش اطلاعات[4] و یکپارچگی اطلاعات[5] است. فرآیند کلی متن‌کاوی در شکل زیر نشان داده‌شده است (Kumar and Karthika 2014).



شکل ۱. فرآیند متن‌کاوی (Kumar and Karthika 2014)

گام اول- آماده‌سازی: در این گام فعالیت‌هایی چون تقسیم جملات به کلمات و حذف کاما و فاصله بین کلمات (نشانه‌گذاری کلمات)[6]، حذف کلماتی که در تحلیل ارزش‌افزوده‌ای ایجاد نمی‌نمایند مانند تک‌های html، xml، حروف ربط و اضافه (حذف ایست واژه‌ها)[7] و پیدا کردن ریشه کلمات و جایگزینی با کلمه اصلی (ریشه‌یابی کلمات)[8] انجام می‌شود.

1.Cross industry standard process for data mining
2.Sample, explore, modify, model, and assess
3.Information retrieval
4.Information processing
5.Information integration
6.Tokenization
7.Stop word removal
8.Stemming

گام دوم- انتقال متن: در این گام جهت تحلیل کاراتر، متن به بسته‌ای از کلمات[1] یا مدل فضای بردار[2] تبدیل می‌شود.

گام سوم- انتخاب ویژگی: در این مرحله ویژگی‌هایی که مرتبط باهدف تحلیل نیستند حذف می‌شوند. از مزایای پیاده‌سازی این مرحله کاهش اندازه مجموعه داده و به‌تبع آن کاهش محاسبات است.

گام چهارم- تکنیک‌های متن‌کاوی: روش‌های مختلفی چون طبقه‌بندی، خوشه‌بندی، بازیابی اطلاعات، کشف موضوع[3] استخراج موضوع[4] و خلاصه‌سازی برحسب هدف قابل پیاده‌سازی هستند.

گام پنجم- ارزیابی: در این گام نتایج حاصل از گام چهارم با معیارهایی چون accuracy، precision و recall ارزیابی می‌شوند (Korde and Mahender 2012).

پیشینه تجربی پژوهش

یکی از حوزه‌های اصلی در زمینه هوشمندی کسب‌وکار، شناسایی موضوع‌های اصلی مرتبط با آن است. نگاش هشت موضوع پردازش داده آنلاین[5] انباره داده، مصورسازی، داده‌کاوی، بازاریابی ارتباط با مشتریان[6] سیستم‌های تصمیم‌یار[7] سیستم اطلاعاتی سازمانی، مدیریت دانش، سیستم اطلاعاتی جغرافیایی[8] را به‌عنوان موضوعات اصلی هوشمندی کسب‌وکار معرفی می‌نماید(Negash 2004) . در پژوهشی دیگر به مؤلفه‌های هوشمندی کسب‌وکار اشاره می‌نماید. فرآیند تحلیل آنلاین، تحلیل پیشرفته، مدیریت عملکرد سازمان، مخزن داده و انباره داده ازجمله زیر حوزه‌های مرتبط با هوشمندی کسب‌وکار هستند (Ranjan 2009).

چن، چیانگ و استوری در مقاله‌ای با عنوان «تحلیل و هوشمندی کسب‌وکار: از کلان داده تا اثرات بزرگ» فرصت‌ها و چالش‌های حوزه هوشمندی کسب‌وکار را در تحلیل می‌نماید. وی در این مقاله پنج حوزه تحلیل داده‌های کلان[9] تحلیل وب،[10] تحلیل شبکه[11] تحلیل متن و تحلیل موبایل را موضوعات اصلی هوشمندی کسب‌وکار می‌داند (Chen, Chiang and Storey 2012).

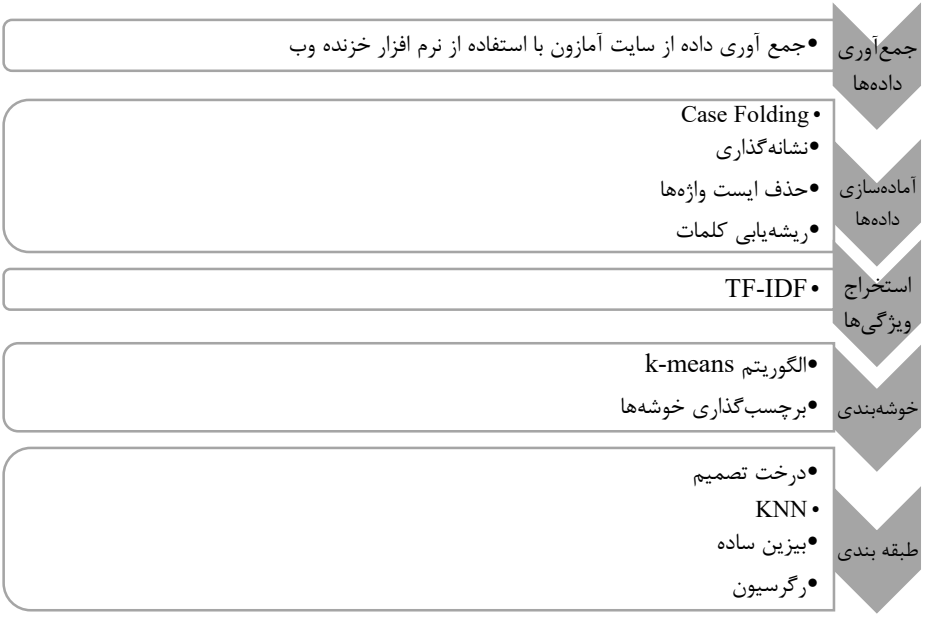
لیم، چن و چن در مقاله‌ای با عنوان «تحلیل و هوشمندی کسب‌وکار: مسیر تحقیق» سه حوزه تحلیل کلان داده‌ها، تحلیل متن و تحلیل شبکه را به‌عنوان سه موضوع اصلی هوشمندی کسب‌وکار معرفی می‌نماید (LIM, CHEN and CHEN 2013). در تحقیقی در سال ۲۰۱۴ سه روند مرتبط با هوشمندی کسب‌وکار را رایانش ابری، سلف‌سرویس، تحلیل و نمایش داده است (Obeidat et al. 2014). در پژوهشی دیگر داده‌کاوی، انباره داده و سیستم‌های پشتیبانی از تصمیم به‌عنوان زیر حوزه‌های مرتبط با هوشمندی

1.Bag of words
2.Vector space document
3.Topic discovery
4.Topic extraction
5.On-Line Data Processing (OLAP)
6.Customer Relationship Management (CRM)
7.Decision Support Systems (DSS)
8.Geographic Information Systems (GIS)
9.(Big) data analytics
1 .Web analytics 0
1 .Network analytics 1

کسب‌وکار معرفی‌شده است. در این تحقیق از روش‌های متن‌کاوی استخراج زیر حوزه‌های مرتبط با هوشمندی کسب‌وکار استفاده‌شده است (Moro, Cortez and Rita 2015).

روش‌شناسی پژوهش

در این مقاله از روش‌های متن‌کاوی[1] جهت ارائه مدلی برای پیش‌بینی حوزه‌های مرتبط با هوشمندی کسب‌وکار[2] استفاده گردیده است. در ابتدا با استفاده از الگوریتم‌های خوشه‌بندی[3] بر روی عنوان و مقدمه کتاب‌ها بر اساس شباهت محتوایی، خوشه‌های اصلی شناسایی گردید. هر یک از این خوشه‌های یک حوزه خاص مرتبط با هوشمندی کسب‌وکار هستند که بر اساس فراوانی واژگانی[4] برچسبی به هریک از آن‌ها اختصاص داده‌شده است. پس از تخصیص برچسب به هر خوشه، برچسب‌ها در الگوریتم‌های طبقه‌بندی[5] جهت پیش‌بینی حوزه‌های مرتبط با هوشمندی کسب‌وکار بکار گرفته‌شده است. مراحل انجام این پژوهش مشتمل بر چهار مرحله جمع‌آوری داده‌ها، آماده‌سازی داده‌ها، خوشه‌بندی و طبقه‌بندی است که در شکل ۲ نشان داده‌شده است.



شکل ۲. چهارچوب پژوهش

1.Text Mining
2.Business Intelligence
3.Clustering
4.Term Frequency
5.Classification

جمع‌آوری داده

برای استخراج داده از یک نرم‌افزار خزنده[1]استفاده گردیده است. این نرم‌افزار اختصاصاً برای پیمایش صفحات سایت آمازون طراحی‌شده تا بتواند بر اساس یک موضوع جستجو شده تمامی کتاب‌های مرتبط را از بین تمامی اقلام سایت جدا کرده و سپس عنوان و مقدمه را در ارتباط با هر کتاب پیدا و جدا کند و در قالب فایل JSON[2] ذخیره کند. جهت طراحی این خزنده از زبان برنامه‌نویسی جاوا استفاده‌شده است که با وارد نمودن کلمات کلیدی[3]کلیه اطلاعات مرتبط با موضوع جستجو شده استخراج و ذخیره می‌گردد. با وارد نمودن عنوان «هوشمندی کسب‌وکار» عنوان و مقدمه کلیه کتاب‌هایی که مرتبط با این کلمات هستند از سایت آمازون استخراج گردید و با فرمت JSON ذخیره گردید. با استفاده از نرم‌افزار مذکور عنوان و مقدمه ۲۲۱ کتاب مرتبط با موضوع هوشمندی کسب‌وکار استخراج گردید که داده‌های استخراج‌شده در گام‌های بعدی جهت تحلیل مورداستفاده قرارگرفته است.


```
[{"isbn10":"0133940306","bookDetailPage":"https://www.amazon.com/Information-Systems-
qid=1500441351&sr=8-1&keywords=information+system","authorName":null,"bookName":null,
.com/Management-Information-Systems-Managing-Digital/dp/0133898164/ref=sr_1_2/130-990
horName":null,"bookName":null,"comments":null},{"isbn10":"1133629628","bookDetailPage
33629628/ref=sr_1_3/130-9905790-0804621?ie=UTF8&qid=1500441351&sr=8-3&keywords=inform
1118173538","bookDetailPage":"https://www.amazon.com/Health-Care-Information-Systems-
51&sr=8-5&keywords=information+system","authorName":null,"bookName":null,"comments":n
g-Information-Systems-Binder-Version/dp/1119244285/ref=sr_1_6/130-9905790-0804621?ie=
ookName":null,"comments":null},{"isbn10":"0133428532","bookDetailPage":"https://www.a
r_1_7/130-9905790-0804621?ie=UTF8&qid=1500441351&sr=8-7&keywords=information+system",
ookDetailPage":"https://www.amazon.com/Business-Driven-Information-Systems-Management
```


شکل ۳. نمونه‌ای از داده‌های استخراج‌شده در قالب JSON

همان‌طور که اشاره شد جهت جمع‌آوری و انتقال داده‌ها از فرمت JSON استفاده‌شده است. JSON استانداردی سبک، کم‌حجم، باز و متنی جهت انتقال داده است. خصیصه‌های موجود در داده‌های خام فایل JSON در جدول (۲) شرح داده‌شده است.

جدول ۲. خصیصه‌های موجود در داده‌های خام این تحقیق، نوع داده و تعریف آن‌ها

نام خصیصه	نوع داده	تعریف
شناسه کتاب	رشته‌ای[4]	کلید اصلی یکتا برای شناسایی کتاب
آدرس کتاب	ورچر	شناسه یو-آر-الی که از آن قسمت می‌توان به کتاب دسترسی داشت.
عنوان کتاب	رشته‌ای	شامل نام کتاب که بر اساس جستجوی کلمه هوشمندی کسب‌وکار استخراج‌شده است.

1.Crawler
2.JavaScript object notation
3.Keywords
4.String

نام خصیصه	نوع داده	تعریف
مقدمه	رشته‌ای	مقدمه مستخرج از سایت آمازون که مرتبط با هر یک از عناوین جستجو شده است.

آماده‌سازی داده‌ها

آماده‌سازی داده مهم‌ترین گام در متن‌کاوی، پردازش زبان طبیعی[1] و بازیابی اطلاعات[2] است. در متن‌کاوی، پیش‌پردازش داده‌ها جهت استخراج دانش از داده‌های متنی بدون ساختار[3] استفاده می‌شود. به دلیل آنکه متن‌ها اغلب شامل فرمت‌های خاصی چون عدد، تاریخ و کلماتی هستند که به متن‌کاوی کمک نمی‌کنند و می‌توانند حذف شوند، لازم است پیش از هرگونه تحلیل بر روی داده‌ها، فرمت‌های غیرضروری حذف شوند (Gurusamy and Kannan 2014). آماده‌سازی داده‌ها شامل تکنیک‌هایی چون Case Folding، نشانه‌گذاری[4]، حذف ایست واژه‌ها[5] و ریشه‌یابی کلمات[6] است که در این پژوهش از این تکنیک‌ها جهت پیش‌پردازش داده‌ها استفاده‌شده است (Nayak et al. 2016).

نشانه‌گذاری کلمات: نشانه‌گذاری یا جداسازی متن، تقسیم‌بندی متن به واحدهای کوچک‌تری چون کلمات[7]، اصطلاحات[8]، نمادها[9] است که در اصطلاح نشانه[10] نامیده می‌شوند (Katariya and Chaudhari 2015). در این پژوهش کلیه مستندات به واحدهای کوچک‌تری تقسیم شدند و درنهایت کلمات به‌دست‌آمده از متن‌ها به‌صورت متغیرهای مجزا در نظر گرفته‌شده‌اند. از نشانه‌گذاری برای استخراج کلیه کلمات مستندات استفاده‌شده است. در ضمن در این پژوهش کلماتی که کوچک‌تر از چهار حرف و بزرگ‌تر از ۲۵ حرف هستند در نظر گرفته نشده‌اند، به دلیل آنکه کلمات زائد و بی‌تأثیر به‌عنوان متغیر در مدل در نظر گرفته نشوند.

Case Folding: در این گام کلیه کلمات ازنظر کوچک بودن یا بزرگ بودن حروف به یک‌شکل درمی‌آیند. این مرحله ازاین‌جهت انجام می‌شود که اگر کلمه‌ای چندین بار با صورت یکسان اما متفاوت در حروف بزرگ و کوچک تکرار شده باشد در مدل‌سازی یک‌بار در نظر گرفته شوند (Thelwall and Chibelushi 2009). در این پژوهش کلیه حروف بزرگ به حروف کوچک[11] تبدیل‌شده است. لازم به

1.Natural Language Processing (NLP)
2.Information Retrieval (IR)
3.Unstructured text data
4.Tokenization
5.Stop-word
6.Stemming
7.Words
8.Terms
9.Symbols
10 .Tokens
11 .Lower case

توضیح است این مرحله در مورد مستندات فارسی کاربرد ندارد و بیشتر در مورد متون به زبان انگلیسی کاربردی است.

حذف ایست واژه‌ها: ایست واژه‌ها کلماتی هستند که باوجود تکرار فراوان در متن از جهت معنایی اهمیت کمی دارند. این کلمات حاوی اطلاعات نیستند مانند ضمایر، حروف اضافه و حروف ربط (Ramasubramanian and Ramya 2013).

ریشه‌یابی کلمات: این روش جهت یافتن ریشه کلمات استفاده می‌شود. در این مرحله کلیه کلمات به فرمت ریشه اصلی خود درمی‌آیند. این مرحله به‌منظور یکسان‌سازی و ساده‌سازی پردازش در مراحل بعد اعمال می‌شود (Ramasubramanian and Ramya 2013). روش‌های ریشه‌یابی اغلب مبتنی بر زبان هستند. الگوریتم‌های Porter و Snowball ازجمله الگوریتم‌های ریشه‌یابی برای زبان انگلیسی هستند که در این مقاله از الگوریتم Snowball استفاده‌شده است. در جدول ۳ نمونه‌ای از مراحل پیش‌پردازش بر روی‌داده‌ها نشان داده‌شده است.

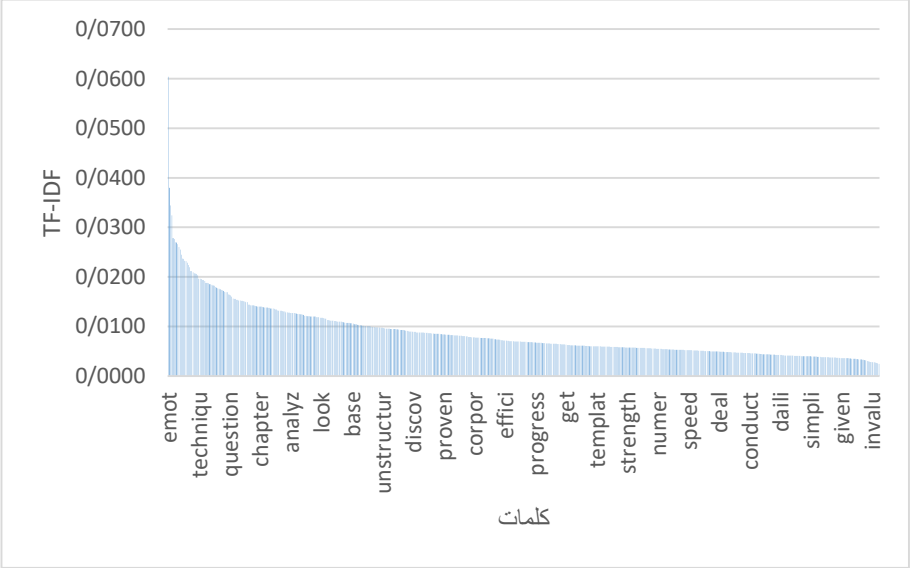
جدول ۳. مراحل پیش‌پردازش داده‌ها

پیش‌پردازش	قبل از پیش‌پردازش	پس از پیش‌پردازش
نشانه‌گذاری کلمات	Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications	“Business”, “Intelligence”, “Roadmap”, “The”, “Complete”, “Project”, “Lifecycle”, “for”, “Decision”, Support “, “Applications”
Lower case	“Business”, “Intelligence”, “Roadmap”, “The”, “Complete”, “Project”, “Lifecycle”, “for”, “Decision”, Support “, “Applications”	“business”, “intelligence”, “roadmap”, “the”, “complete”, “project”, “lifecycle”, “for”, “decision”,” support “, “applications”
حذف ایست واژه‌ها	“business” , “intelligence”, “roadmap”, “the”, “complete” , “project” , “lifecycle” , “for” , “decision”,” support “, “applications”	“business” , “intelligence”, “roadmap”, “complete” , “project” , “lifecycle” , “decision”,” support “, “applications”
ریشه‌یابی کلمات	“business” , “intelligence”, “complete”	“busin” , “intellig”, “compl”

استخراج ویژگی‌ها

تکنیک استخراج ویژگی‌ها جهت استخراج ویژگی‌های اصلی در طبقه‌بندی متن، بازیابی اطلاعات، تشخیص موضوع و خلاصه‌سازی سند مورداستفاده قرار می‌گیرد. روش‌های اصلی در این تکنیک، فراوانی کلمه در

مقابل فراوانی سند[1] (TF-IDF) اطلاعات به‌دست‌آمده[2] (IG)، آماره مربع کای[3] (CHI) است (Chakraborty 2013). هدف از این کار کم کردن تأثیر لغاتی است که ارزش‌افزوده کمتری به برای تحلیل دارند. لازم به ذکر است که الزاماً لغات پرکاربرد لغات باارزش افزوده یا معناداری برای تحلیل نیستند. قبل از آنکه قادر به اجرای الگوریتم k-means روی مجموعه‌ای از اسناد باشیم باید بتوان اسناد را به‌عنوان بردارهای دوبه‌دو مقایسه کرد که در این پژوهش از تکنیک فراوانی کلمه در مقابل فراوانی سند استفاده‌شده است که مقادیر هر متغیر نسبت به سند با استفاده از روش TF-IDF محاسبه‌شده‌اند. این روش کلمات را بر اساس اهمیتشان وزن دهی می‌کند که فراوانی کلمه نسبت تعداد تکرار یک کلمه در سند به تعداد کلمات موجود در سند است. در مقابل فراوانی سند لگاریتم نسبت تعداد سندها به تعداد اسنادی است که کلمه موردنظر را دارا است. این روش باعث می‌شود تا کلماتی که در مجموعه کمتری از اسناد هستند وزن بیشتری پیدا کنند و کلماتی که در اغلب اسناد موجودند وزن کمتری پیدا کنند (لطفی آذری داریان و جاویدان ۱۳۹۵). بدین ترتیب کلمات با وزن کمتر در آنالیز وارد نمی‌شوند که در این تحقیق پس از اعمال این گام‌ها تعداد کلمات استخراج‌شده به ۶۰۵ متغیر رسید. همچنین کلماتی که در کمتر از سه درصد و بیشتر از سی درصد از اسناد تکرار شده بودند هرس شدند تا دقت تحلیل افزایش یابد. در شکل ۴ کلمات با بیشترین TF-IDF نشان داده‌شده است.



شکل ۴. TF-IDF

1.Term Frequency Inverse Document Frequency (TFIDF)
2.Information gain
3.Chi-square statistics

خوشه‌بندی

خوشه‌بندی به معنای تقسیم نمودن داده‌ها به گروه‌های مشابه است. خوشه‌ها طوری گروه‌بندی می‌شوند که شباهت زیادی در بین اشیاء از یک خوشه و هم‌چنین عدم شباهت زیادی بین اشیاء از خوشه‌های مختلف وجود داشته باشد. هدف از خوشه‌بندی داده‌ها حداقل کردن واریانس درون‌گروهی و حداکثر کردن واریانس میان گروهی بر پایه‌یک تابع فاصله‌ای یا عدم تجانس است (JALIL et al. 2016). خوشه‌بندی یک روش گروه‌بندی بدون ناظر است. در میان الگوریتم‌های خوشه‌بندی روش k-means به دلیل زمان محاسبه کم و قدرت انطباق بالا در نمونه‌ها با سایز بزرگ و سهولت استفاده، پرکاربردترین است (Kuo.R.J. An.Y. L., Wang .H .S. and Chung .W . J. 2006). روشk-means دارای پارامترهای تعداد خوشه[1] یا نوع مقیاس[2] و واگرایی[3] است (لطفی آذری داریان و جاویدان ۱۳۹۵).

الگوریتم k-means ، پارامتر k را به‌عنوان ورودی گرفته و مجموعه n شیء را به k خوشه افراز می‌کند. به‌طوری‌که سطح شباهت داخلی خوشه‌ها بالا بوده و سطح شباهت اشیاء بیرون خوشه‌ها پائین باشد. شباهت هر خوشه نسبت به متوسط اشیاء آن خوشه سنجیده شده که متوسط مرکز خوشه نیز نامیده می‌شود .در این روش فاصله‌ها بر اساس فاصله اقلیدسی[4] تعیین می‌شود (Singh 2016).

پیاده‌سازی الگوریتم k-means شامل مراحل زیر است:

۱. مشخص نمودن تعداد خوشه‌ها یا k
۲. مشخص نمودن مراکز خوشه‌ها به‌صورت تصادفی بر اساس تعداد خوشه‌ها
۳. تخمین فاصله مراکز کلیه نمونه‌ها با مراکز اولیه مشخص‌شده
۴. تخصیص نمونه‌ها به مراکز نزدیک‌تر
۵. محاسبه مرکز هندسی خوشه‌ها
۶. تکرار گام‌های ۳ تا ۵ تا جایی که مراکز خوشه‌ها ثابت بماند (Hofmann and Chisholm 2016)

پس از مشخص شدن خوشه‌ها لازم است بر روی خوشه‌ها اعتبارسنجی صورت گیرد. هدف از اعتبارسنجی خوشه‌ها، یافتن خوشه‌هایی است که بهترین تناسب را با داده‌های موردنظر داشته باشند. اساس تکنیک‌های ارزیابی روش‌های خوشه‌بندی، بر مبنای دو معیار شباهت درون خوشه‌ای و تفاوت بین خوشه‌ای است. مفهوم کلی این دو معیار این است که اگر مشاهدات به‌گونه‌ای خوشه‌بندی شوند که هر خوشه در حداکثر تراکم بوده و خوشه‌های مختلف در حداکثر دوری از یکدیگر باشند، خوشه‌بندی خوبی انجام‌شده است (Chowdary, Prasanna and Sudhakar 2014). جهت ارزیابی خوشه‌ها، شاخص‌های متفاوتی پیشنهادشده که در این پژوهش از شاخص ارزیابی دیویس بولدین[5] استفاده‌شده است.

1.k
2.Type measure
3.Divergence
4.Squared Euclidean Distance
5.Davies bouldin index

شاخص دیویس بولدین از معیار شباهت بین دو خوشه (R_{ij}) استفاده می‌کند که بر اساس پراکندگی یک خوشه (s_i) و عدم شباهت بین دو خوشه (d_{ij}) تعریف می‌شود. شباهت بین دو خوشه را می‌توان به صورت‌های مختلفی تعریف کرد ولی بایستی شرایط زیر را دارا باشد:

- ✓ $R_{ij} \geq 0$
- ✓ $R_{ij} = R_{ji}$
- ✓ اگر s_i و s_j هردو برابر صفر باشند آنگاه R_{ij} نیز برابر صفر باشد.
- ✓ اگر $s_k < s_j$ و $d_{ik} = d_{ij}$ آنگاه $R_{ij} > R_{ik}$
- ✓ *اگر $s_k = s_j$ و $d_{ik} > d_{ij}$ آنگاه $R_{ij} > R_{ik}$*

معمولاً شباهت بین دو خوشه به‌صورت زیر تعریف می‌شود:

فرمول ۱:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

که در آن d_{ij} و s_i بر اساس فرمول‌های زیر محاسبه می‌شوند:

فرمول ۲:

$$d_{ij} = d(v_i, v_j)$$

فرمول ۳:

$$s_i = \frac{1}{|c_i|} \sum_{x \in c_i} d(x, v_i)$$

به‌این‌ترتیب شاخص دیویس بولدین به‌صورت زیر تعریف می‌شود:

فرمول ۴:

$$DB = \frac{1}{nc} \sum_{i=1}^{nc} R_i$$

که در آن R_i به‌صورت زیر محاسبه می‌شود:

فرمول ۵:

$$R_i = \max(R_{ij}), i = 1 \dots n_c$$

این شاخص درواقع میانگین شباهت بین هر خوشه با شبیه‌ترین خوشه به آن را محاسبه می‌کند. می‌توان دریافت که هرچه مقدار این شاخص کمتر باشد، خوشه‌های بهتری تولیدشده است(Bouldin 1979) .

طبقه‌بندی

در این پژوهش پس از برچسب‌گذاری خوشه‌ها از الگوریتم‌های طبقه‌بندی جهت ارائه مدلی برای پیش‌بینی حوزه‌های مرتبط با هوشمندی کسب‌وکار استفاده‌شده است. دسته‌بندی متون، فرآیند طبقه‌بندی متون به دسته‌های از پیش تعریف‌شده بر اساس محتوای آن‌هاست (Korde and Mahender 2012). طیف

گسترده‌ای از تکنیک‌های طبقه‌بندی در مورد داده‌های کمی وجود دارد. به دلیل آنکه در فرآیند پیش‌پردازش داده‌ها، متون به داده‌های کمی تبدیل می‌گردند، امکان به‌کارگیری روش‌های طبقه‌بندی برای متون نیز امکان‌پذیر است (Aggarwal 2012) . پس از برچسب‌گذاری خوشه‌ها الگوریتم‌های چهار الگوریتم طبقه‌بندی رگرسیون[1] درخت تصمیم[2] KNN[3]، بیزین ساده[4] جهت طبقه‌بندی داده‌ها مورداستفاده قرارگرفته است.

جهت اندازه‌گیری کارایی مدل طبقه‌بندی، یک مجموعه تست که مستقل از مجموعه آموزش است در نظر گرفته می‌شود و برچسب‌هایی که برای این اسناد توسط مدل تخمین زده می‌شود با برچسب واقعی اسناد مقایسه می‌شود. در این پژوهش نسبت اسناد آموزش به تست ۷۰ به ۳۰ است. نسبت اسنادی که به‌درستی طبقه‌بندی‌شده‌اند به تعداد کل اسناد accuracy نامیده می‌شود. دو معیار دیگری که برای مقایسه الگوریتم-های طبقه‌بندی استفاده می‌شود، Precision و Recall هستند. Precision نشان‌دهنده کسری از اسناد بازیابی شده‌ای که مربوط هستند و Recall نشان‌دهنده کسری از اسناد مربوط بازیابی شده است (پرئی و حمیدی ۱۳۹۵).

فرمول ۶:

$$\text{precission} = \frac{(\text{relevant} \cap \text{retrieved})}{\text{retried}}$$

فرمول ۷:

$$recall = \frac{(relevant \cap retrieved)}{relevant}$$

یافته‌های پژوهش

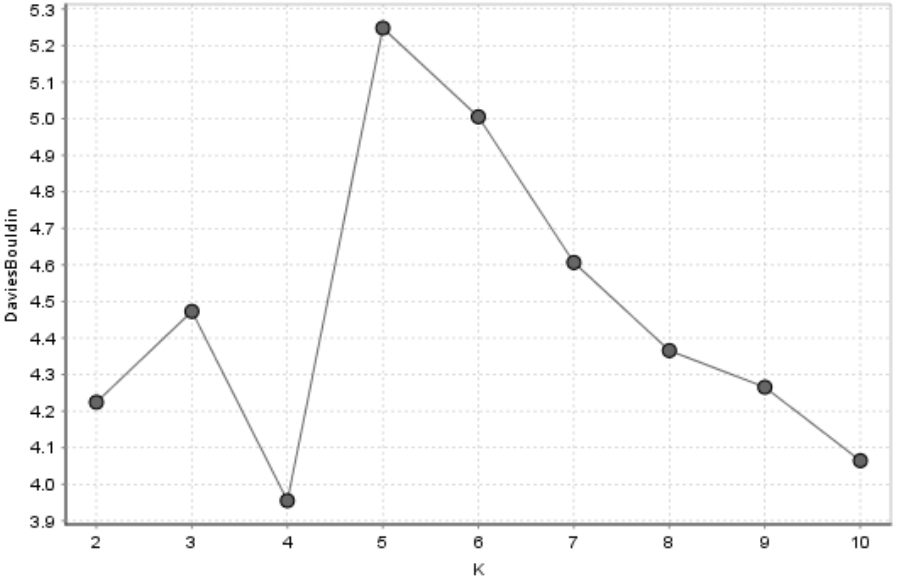
پس از تبدیل داده‌های متنی که ساختار نیافته هستند به داده‌های ساختاریافته با استفاده از روش منتخب در استخراج ویژگی‌ها یعنی TF-IDF، کلمات و واژگان جهت استفاده در مدل‌سازی استخراج گردیدند. پس‌ازاین مرحله واژگان و کلمات کلیدی استخراج‌شده برای مدل‌سازی با الگوریتم k-means مورداستفاده قرارگرفته‌اند. مدل‌سازی حاصل از خوشه‌بندی با تعداد خوشه‌های مختلف و با استفاده از شاخص دیویس بولدین مورد ارزیابی قرارگرفته‌اند. در جدول ۴ مقادیر شاخص دیویس بولدین بر اساس تعداد خوشه‌های متفاوت نمایش داده‌شده است.

1.Regression
2.Decision tree
3.K-nearest neighbors
4.Naïve bayes

جدول ۴. مقادیر دیویس بولدین

تعداد خوشه (k)	دیویس بولدین (Davies Bouldin)
۲	۴/۲۲۴
۳	۴/۴۷۳
۴	۳/۹۵۵
۵	۵/۲۴۸
۶	۵/۰۰۵
۷	۴/۶۰۷
۸	۴/۳۶۶
۹	۴/۲۶۶
۱۰	۴/۰۶۴

همانطوری‌که از شکل ۵ مشخص است تعداد خوشه‌ای بهینه است که شاخص دیویس بولدین کمتری داشته باشد. تعداد شش تا ده خوشه که شاخص دیویس بولدین آن‌ها رو به کاهش است نتایج بهتری دارند. در این تحقیق، مدل با تعداد خوشه هفت و نوع مقیاس واگرایی برگمن و نوع واگرایی فاصله اقلیدسی به‌عنوان مدل بهینه انتخاب شد و مدل نهایی بر اساس تعداد هفت خوشه تهیه گردیده است.



شکل ۵. نمودار مقایسه شاخص دیویس بولدین با تعداد k متفاوت

پس از مدل‌سازی و مشخص نمودن خوشه‌های حاصل از الگوریتم، هر یک از خوشه‌ها مورد تحلیل قرارگرفته است. در این بخش کلمات کلیدی و مفاهیم هر خوشه موردبررسی قرارگرفته و بر اساس کلمات هر خوشه، برچسبی به هر یک از خوشه‌ها تخصیص داده‌شده است. بر اساس برچسب‌های تخصیص داده‌شده هفت حوزه مرتبط با هوشمندی کسب‌وکار به ترتیب ابزارهای هوشمندی کسب‌وکار و مصورسازی[۱] رفتار سازمانی[۲] مدیریت فرآیندها و دانش[۳] سیستم‌های پشتیبانی تصمیم[۴] رهبری[۵] متن‌کاوی[۶] و پایگاه داده و نمایش[۷] است.

جدول ۵. کلمات کلیدی هر خوشه و تعداد کتاب‌های متعلق به هر خوشه

شماره خوشه	کلمات کلیدی خوشه	برچسب خوشه	تعداد کتب متعلق به خوشه
۰	Microsoft, server, sharepoint, excel, report, servic, analysi,offic, model, powerpivot,dashboard,visual	BI tools and Visualization	۲۸
۱	Culture, creativ, people, team. Situat, behavior, idea, mind, interact, social ,organi,educ,compa	organizational behavior	۱۳
۲	Process,wareh,company, project,model,proft,step,applic,knowledge,enerpris,technology ,analysis,valu, success,system,plan,	Process Management	۱۰۰
۳	Support,predict,system, deci, artifice, future, industry,technology,change	Decision Support System	۱۶
۴	Emot, leadership, skill, success, leader,people,person, relashionship, mind	Leadership	۴۰
۵	Manageri,solid,founda,future,hand,perspect,system,mine,text	Text Mining	۵
۶	Oracle,report,creat,design,platform	DataBase	۱۹

1.Business intelligence tools and visualization
2.Organizational behavior
3.Process management
4.Decision support system
5.Leadership
6.Text mining
7.Database

پس از تحلیل خوشه‌ها و استخراج برچسب هر خوشه، نتایج چهار الگوریتم طبقه‌بندی موردبررسی قرار گرفت. جهت مدل‌سازی از چهار الگوریتم طبقه‌بندی رگرسیون[1]، درخت تصمیم[2]، KNN[3]، بیزین ساده[4] جهت طبقه‌بندی داده‌ها مورداستفاده قرارگرفته است. جهت ارزیابی هر یک از مدل‌های طبقه‌بندی از سه معیارAccuracy، Recall و Precision استفاده‌شده است. همان‌طور که از نتایج ارزیابی مشخص است الگوریتم‌های درخت تصمیم و KNN با دقت بالاتری مجموعه داده‌ها را پیش‌بینی نموده‌اند.

جدول ۶. ارزیابی الگوریتم‌های طبقه‌بندی

الگوریتم	Accuracy	Recall	Precision
رگرسیون	۶۶/۶۷	۴۷/۷۰	۵۵/۱۶
درخت تصمیم	۷۸/۷۹	۶۹/۳۵	۸۵/۰۴
KNN	۷۸/۷۹	۷۹/۳۱	۷۲/۸۵
بیزین ساده	۶۰/۶۱	۷۰/۷۹	۷۶/۸۵

نتیجه‌گیری و پیشنهادها

با گسترش استفاده روزافزون از اینترنت جهت خرید آنلاین لازم است اطلاعات بیشتر جهت خرید هوشمندانه در اختیار خریداران قرار داده شود. یکی از وب‌سایت‌های مشهور در خرید آنلاین، آمازون است که امکان خرید آنلاین کتاب را برای خریداران فراهم می‌نماید. یکی از چالش‌های خرید از آمازون عدم دسترسی به اطلاعات زیر بخش‌های مرتبط با یک موضوع است یعنی با جستجو در موضوع اصلی اطلاعات دقیق‌تر و جزئی‌تری از کتاب در اختیار کاربر گذاشته نمی‌شود. ازاین‌جهت تحلیل بر روی داده‌های آمازون و ارائه اطلاعات جزئی‌تر می‌تواند به خریداران در انتخاب هوشمندانه‌تر کمک کند. در این مقاله تلاش شد تا با استفاده از تکنیک‌های بر روی عنوان و مقدمه کتاب‌های مرتبط با هوشمندی کسب‌وکار زیر بخش‌های موضوعی مرتبط با این موضوع اصلی استخراج گردد. با به‌کارگیری الگوریتم‌های خوشه‌بندی هفت خوشه اصلی استخراج گردید. با بررسی و تحلیل خوشه‌های استخراج‌شده برچسب‌هایی به هر خوشه اختصاص داده شد. هفت موضوع مرتبط با هوشمندی کسب‌وکار به ترتیب ابزارهای هوشمندی کسب‌وکار و مصورسازی، رفتار سازمانی، مدیریت فرآیندها و دانش، سیستم‌های پشتیبانی تصمیم، رهبری، متن‌کاوی و پایگاه داده و نمایش است.

نتایج نشان می‌دهند بیشترین تعداد کتاب به ترتیب متعلق به خوشه‌های ۳، ۵ و ۱ است که به ترتیب مرتبط با موضوعات مدیریت فرآیندها، رهبری و ابزارهای هوشمندی کسب‌وکار و مصورسازی هستند که نشان می‌دهد روند پژوهش در حوزه هوشمندی کسب‌وکار به سمت فرآیندها و رویه‌های پیاده‌سازی و نیز

1.Regression
2.Decision tree
3.K-nearest neighbors
4.Naïve bayes

ابزارهای پیاده‌سازی هوشمندی کسب‌وکار می‌رود. در ضمن در سال‌های اخیر موضوعات مرتبط با رفتار و فرهنگ‌سازمانی نیز موردتوجه قرارگرفته است که بیانگر این موضوع است که در سال‌های اخیر توجه به سمت مدیریت و بحث‌های مرتبط با منابع انسانی و رفتار سازمانی در زمینه هوشمندی کسب‌وکار دارد. این موضوع می‌تواند بیانگر این امر باشد که در سال‌های اخیر هوشمندی کسب‌وکار از بحث‌های صرف آکادمیک خارج‌شده است و پیاده‌سازی و عکس‌العمل منابع انسانی در مقابل تغییرات ناشی از این پیاده‌سازی از موضوعات موردتوجه بوده است. در ضمن ابزارهای هوشمندی کسب‌وکار و مصورسازی نتایج حاصل از تحلیل نیز موردتوجه بوده است که نشان می‌دهد امروزه استفاده از نتایج حاصل از تحلیل هوشمندی کسب‌وکار جهت بررسی وضعیت و پیش‌بینی آینده موردتوجه مدیریت سازمان‌ها قرارگرفته است که تأکید بر این موضوع است که عملیاتی نمودن و نمایش نتایج حاصل از تحلیل‌ها از موضوعات موردتوجه است. نتایج این پژوهش می‌تواند در سیستم‌های پیشنهاددهنده جهت ارائه پیشنهاد به کاربران و کمک به تصمیم-گیری آگاهانه‌تر و هوشمندانه‌تر کمک نماید. در ضمن به شرکت‌های فعال در حوزه تجارت الکترونیک در ارائه خدمات کاراتر و اثربخش‌تر به مشتریان کمک نماید. در این تحقیق از رویکرد خوشه‌بندی جهت تشخیص موضوع اصلی و تم کتاب‌ها استفاده‌شده است که در تحقیقات آتی می‌توان از سایر روش‌های تشخیص موضوع استفاده نمود. پیشنهاد می‌گردد با جستجو بر کلیدواژه کلان‌تر و با استفاده از روش خوشه‌بندی تجزیه‌ای، تحلیلی بر روی موضوعات مرتبط با این حوزه انجام شود.

فهرست منابع

پرئی، اعظم السادات و حجت اله حمیدی. ۱۳۹۵. "ارائه رویکردی برای مدیریت و سازماندهی اسناد متنی با استفاده از تجزیه وتحلیل هوشمند متن." *پردازش و مدیریت اطلاعات*.

لطفی آذری داریان، سولماز و رضا جاویدان. ۱۳۹۵. "استفاده از روشهای داده‌کاوی به منظور تسهیل جستجو در موتورهای جستجوگر متنی." بیست و چهارمین کنفرانس برق ایران. شیراز. ۲۸۰۹-۲۸۱۷.

Aggarwal, Charu C and Zhai, ChengXiang .۲۰۱۲ .*Mining text data* .Springer.

Bouldin, Davies D.L“ .۱۹۷۹ .A cluster separation measure ”.*IEEE Transactions on Pattern Analysis and Machine Intelligence*.۲۲۴-۲۲۷

Chakraborty , Rakhi“ .۲۰۱۳ . Domain Keyword Extraction Technique: A New Weighting Method Based on Frequency Analysis ”.*Computer Science & Information Technology*.۱۰۹-۱۱۸

Chapman, Pete .۲۰۰۰ . *CRISP-DM ۱٫۰Step-by-step data mining guide* .USA and Denmark: CRISP consortium.

Chen, Hsinchun , Roger H. L .Chiang و,Veda C .Storey“ .۲۰۱۲ .Business Intelligence and Analysis:from Big Data to Big Impact ”.*MIS Quarterly*.۱۱۶۵-۱۱۸۸

Chibelushi, Caroline و , Mike Thelwall“ .۲۰۰۹ . Text Mining for Meeting Transcript Analysis to Extract Key Decision Elements ”.*International MultiConference of Engineers and Computer Scientists*.۷۱۰-۷۱۵

Chowdary, N Sunil , D Sri Lakshmi Prasanna و,P Sudhakar“ .۲۰۱۴ .Evaluating and Analyzing Clusters in DataMining using Different Algorithms ”.*International Journal of Computer Science and Mobile Computing*.۸۶-۹۹

Gupta , V و,G.S Lehal“ .۲۰۰۹ . A Survey of Text Mining Techniques and Applications Emerging Technologies in Web Intelligence ”.,*Academy Publisher*.۶۰-۷۶

Gurusamy, Vairaprakash و,Subbu Kannan“ .۲۰۱۴ .Preprocessing Techniques for Text Mining ”.*RTRICS* .Podi.

Hofmann, Markus و , Andrew Chisholm .۲۰۱۶ .*Text Mining and Visualization Case Studies Using Open-Source Tools* .Broken Sound Parkway NewYork: Taylor &Francis Group.

JALIL, Abdennour Mohamed , Imad HAFIDI ,Lamiae ALAMI و,ENSA Khouribga . “ .۲۰۱۶Comparative Study of Clustering Algorithms in Text Mining Context ”.*International Journal of Interactive Multimedia and Artificial Intelligence*.۴۲-۴۵

Katariya, Nikita P و,M S Chaudhari“ .۲۰۱۵ .Text Preprocessing for Text Mining Using Side Information ”.*International Journal of Computer Science and Mobile Applications*.۱-۵

Korde, Vandana و , C Namrata Mahender“ .۲۰۱۲ .Text Classification and Classifiers:A Survey ”.*International Journal of Artificial Intelligence & Applications (IJAIA)*.۸۵-۹۹

Kumar B, Sathees و,Karthika R“ .۲۰۱۴ .A Survey on Text Mining Process and Techniques ”.*International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*.۲۲۷۹-۲۲۸۴

Kumar, B Shravan and Ravi, Vadlamani“ .۲۰۱۶ .A survey of the applications of text mining in financial domain ”.*Knowledge-Based Systems*.۱۲۸-۱۴۷

Kuo.R.J. An.Y. L., Wang .H .S و,.Chung .W . J“ .۲۰۰۶ .Integration of self-organizing feature maps neural network and genetic K-means algorithm for market segmentation ”.*Expert Systems with Application*.۳۳۱-۳۲۴

Liao, S H, P H Chu و,P Y Hsiao“ .۲۰۱۲ .Data mining techniques and applications–a decade review from ۲۰۰۰to ”.۲۰۱۱*Expert Systems with Applications*–۳۰۳۱۱ .۳۱۱ ۱۱

LIM, EE-PENG , HSINCHUN CHEN و,GUOQING CHEN“ .۲۰۱۳ .Business Intelligence and Analytics: Research Directions ”.*ACM Transactions on Management Information Systems*.۱۷:۱-۱۷:۱۰

Miner, Gary, Dursun Delen, John Elder, Andrew Fast, Thomas Hill و,Robert A . Nisbet .۲۰۱۲ .*Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications* .Waltham,USA: Academic Press.

Moro, Sérgio , Paulo Cortez و,Paulo Rita“ .۲۰۱۵ .Business intelligence in banking: A literature analysis from ۲۰۰۲to ۲۰۱۳using text mining and latent Dirichlet allocation ”.*Expert Systems with Applications*.۱۳۲۴–۱۳۱۴

Nayak, Arjun Srinivas, Ananthu P Kanive, Naveen Chandavekar و,Balasubramani R . “ .۲۰۱۶Survey on Pre-Processing Techniques for Text Mining ”.*International Journal Of Engineering And Computer Science*.۱۶۸۷۵-۱۶۸۷۹ :(۶) ۵

Negash, Solomon“ .۲۰۰۴ . Business Intelligence ”.*Communications of the Association for Information Systems*.۱۷۶-۱۹۶

Obeidat, Muhammad , Sarah North, Max North و,Vebol Rattanak“ .۲۰۱۴ .Business Intelligence Domain and Beyond ”.*Universal Journal of Industrial and Business Management*.۱۲۷-۱۳۴

Olorisade, Babatunde Kazeem , Pearl Brereton و,Peter Andras“ .۲۰۱۷ .Reproducibility of studies on text mining for citation screening in systematic reviews: Evaluation and checklist ”.*Journal of Biomedical Informatics*.۱-۱۳

Ramasubramanian, C و,R Ramya“ .۲۰۱۳ .Effective Pre-Processing Activities in Text Mining using Improved Porter’s Stemming Algorithm ”.*International Journal of Advanced Research in Computer and Communication Engineering*۴۵۳۶-.۴۵۳۸

Ranjan, Jayanthi“ .۲۰۰۹ .Business Intelligence: Concepts, Components, Techniques and Benefits ”.*Journal of Theoretical and Applied Information Technology* .۶۰-۷۰

Sagayam, R“ .۲۰۱۲ .survey of text mining: Retrieval, extraction and indexing techniques ”.*International Journal of Computational Engineering Research* .۱۴۴۳-۱۴۴۶

Singh, Hardeep“ .۲۰۱۶ .Clustering of text documents by implementation of K-means algorithms ”.*Streamed Info-Ocean*.۵۳-۶۳

Talib, Ramzan, Muhammad Kashif Hanif, Shaeela Ayesha و,Fakeeha Fatima .۲۰۱۶ . “Text Mining: Techniques, Applications and Issues ”.*International Journal of Advanced Computer Science and Applications*.۴۱۴-۴۱۸

Truyens, Maarten and Van Eecke, Patrick“ .۲۰۱۴ .Legal aspects of text mining ”. *Computer law &| security review*.۱۷۰--۱۵۳

Weiss, S M, N Indurkhya, T Zhang و ,F Damerau“ .۲۰۱۰ .Text mining:predictive methods for analyzing unstructured information ”.*Springer Science and Business Media .*

Weng ,S و ,Y Lin“ .۲۰۰۳ . A study on searching for similar documents based on multiple concepts and distribution of concepts ”.*Expert Systems with Applications*.۳۵۵-۳۶۸ ,

Presenting a Model to Predict Business Intelligence Domain

Fatemeh Abbasi
PhD Candidate in IT, Faculty of Management, University of Tehran, Tehran, Iran[1]
Babak Sohrabi
Prof. in IT, Faculty of Management, University of Tehran, Tehran, Iran
Ameneh Khadivar
Associate Prof., Dep. of Social and Economic, AL-Zahra University, Tehran, Iran
Amir Manian
Prof. in IT, Faculty of Management, University of Tehran, Tehran, Iran

Abstract: In recent years, the volume of data and information exchanged over web pages, social networks, emails, and blogs is increasing. Most of the information that is exchanged and stored is in text format, which is very valuable given the huge amount of text data that is analyzed and discovered from these data. Text mining is one of the most important methods to extract knowledge from structured data which help organizations to achieve their goals. These days most of products like books purchased via internet or online. So that customers need dedicated information to make the decision to buy products more intelligently and smartly. Amazon is one of trading website which give related information to their customer for online buying. One of problem for these method for buying books is subcategory of books is not cleared which customer may buy unwanted books. In this research we tried to present predicting model to predict topic related to Business Intelligence by using text mining methods. Results show that Business Intelligence tools and visualization, behavioral organization, process management, decision support system, leadership, text mining and database are seven clusters which related to Business Intelligence.

Keywords: Business Intelligence, Text Mining, Clustering, Classification, Structured Data

1.Corresponding Author: f_abbasi@ut.ac.ir