بررسی خودکار تطابق بین منابع درسی و آیین‌نامه‌های اجرایی درزمینه طراحی و اجرای ساختمان‌های بتنی: رویکرد متن‌کاوی

فاطمه حافظی مقدس[1]، مجتبی مغربی[2]*

۱- دانشجوی کارشناسی ارشد، دانشکده مهندسی، گروه عمران، گرایش مهندسی و مدیریت ساخت، دانشگاه فردوسی، مشهد، ایران،
Fatemeh.h.m@gmail.com

۲- دانشیار، دانشکده مهندسی، گروه عمران، گرایش مهندسی و مدیریت ساخت، دانشگاه فردوسی مشهد، ایران،
mojtabamaghrebi@ferdowsi.um.ac.ir

چکیده

یکی از چالش‌های آموزش عالی در تدوین متون آموزشی دانشگاهی، تطابق و تناسب حداکثری محتوای منابع آموزشی با دستورالعمل‌ها است. به همین خاطر برای رسیدن به سیستم آموزشی کارا و هم‌راستا با نیاز صنعت، باید میزان تناسب محتوای منابع آموزشی با آیین‌نامه‌ها ارزیابی شده و در صورت لزوم بازبینی شود. ضرورت بازبینی منابع آموزشی در حوزه فنی و مهندسی از آن جهت اهمیت دارد که این رشته‌ها در کاربرد صنعت و خدمات کشور مورد نیاز است و در واقع تربیت نیروهای متخصص و فنی که بتواند نیازهای فنی کشور را در سطوح مختلف مرتفع کند، مهم‌ترین کارکرد برنامه‌های درسی دانشگاه‌ها است. این مسئله اهمیت توجه به منابع درسی این رشته‌ها را دوچندان می‌کند. این پژوهش به دنبال استخراج کلمات کلیدی از آیین‌نامه بتن ایران و کتاب «سازه‌های بتن‌آرمه» است که با استفاده از سه رویکرد آماری، دانش زبانی و ساخت شبکه یک روش ترکیبی را پیشنهاد می‌دهد. در این مطالعه کلمات کلیدی هر سند در یک شبکه‌ی خوشه‌بندی شده ترسیم می‌شود و مورد تحلیل و بررسی قرار می‌گیرد. مقایسه نتایج نشان می‌دهد که محتوای این دو سند کاملاً مشابه نیست و موضوعاتی که بر آن‌ها تأکید شده با هم تفاوت‌هایی دارند. در حقیقت می‌توان گفت که آیین‌نامه یک سند دستورالعملی است که در آن به تمام جزئیات پرداخته شده و در مورد همه مسائل پیرامون سازه‌های بتنی صحبت کرده است. اما کتاب سازه‌های بتن‌آرمه به دستورالعمل‌های طراحی بسنده نکرده و مفاهیم مرتبط با طراحی سازه‌های بتنی را به صورت جزئی و عمیق بررسی کرده است.

کلمات کلیدی

متن‌کاوی، استخراج کلمات کلیدی، تحلیل شبکه، شاخص مرکزیت، آیین‌نامه بتن ایران (آبا).

۱- مقدمه

با توجه به حجم بسیار بالای متن‌ها و مقالات الکترونیکی، آرشیوهای بزرگ دیجیتالی و گسترش شبکه‌های اجتماعی که روزانه در حال افزایش است، استفاده از فنون داده‌کاوی بیش‌ازپیش ضرورت می‌یابد. حجم اطلاعات در سال ۲۰۲۰ با توجه به گزارشی از EMC از ۴۰ زتابایت ($۴۰*۱۰^{۲۱}$ بایت) عبور می‌کند [۱] و از طرفی ۸۰ درصد این اطلاعات به صورت متنی است. متن‌کاوی یا کشف دانش از متن اولین بار توسط فلدمن و داگان[1] [۲] در سال ۱۹۹۵ به عنوان زیرشاخه‌ای از علم داده‌کاوی مطرح شده است [۳].

از آنجایی که پردازش دستی تعداد زیاد سند متنی زمان‌بر و مستعد خطاست، متن‌کاوی به استخراج خودکار اطلاعات و شناسایی الگوهای پنهان ارزشمند از داده‌های متنی بدون ساختار می‌پردازد [۴]. متن‌کاوی شامل ۷ حوزه‌ی مختلف دسته بندی متون[2]، خوشه‌بندی متون[3]، بازیابی اطلاعات[4]، استخراج اطلاعات[5]، وب کاوی[6]، پردازش زبان طبیعی[7] و استخراج مفاهیم[8] است. بسیاری از فنون متن‌کاوی از جمله پردازش زبان طبیعی، استخراج اطلاعات و استخراج مفاهیم نیاز به کلمات کلیدی دارند [۵]. کلمات کلیدی مجموعه‌ای از یک یا چند کلمه هستند که نمایش فشرده‌ای از محتوای یک متن را ارائه می‌دهند. در حقیقت کلیدواژه همانند یک کوه یخ نشانگر محتوای هر سند است [۶].

برای یافتن کلمات کلیدی متن دو رویکرد کلی وجود دارد: ۱- تخصیص کلمات کلیدی ۲- استخراج کلمات کلیدی [۷]

در روش تخصیص کلمات کلیدی، واژه‌های مهم از مجموعه‌ای از کلمات از پیش تعریف شده تحت عنوان واژه‌نامه انتخاب شده و به متن اختصاص داده می‌شوند. سادگی ساختار الگوریتم، نشان دادن اسناد مشابه با یک مجموعه کلیدواژه و کنترل کردن خروجی از جمله مزایای این روش محسوب می‌شوند. در مقابل زمان‌بر و هزینه‌بر بودن ایجاد و نگهداری واژه نامه و احتمال در دسترس نبودن واژه‌نامه مناسب و کامل از جمله معایب این روش به شمار می‌آیند.

شناسایی مرتبط‌ترین عبارات و کلمات در مورد محتوای یک سند متنی را می‌توان به عنوان هدف روش استخراج کلمات کلیدی در نظر گرفت. در این رویکرد نیازی به واژه‌نامه‌ی از قبل تهیه شده نیست و از کلمات موجود در خود متن استفاده می‌شود. با این حال معمولاً فرآیند رسیدن به کلمات کلیدی سخت و پر چالش است.

رویکردهای استخراج کلمه کلیدی شامل موارد زیر است [۷]:

- رویکردهای مبتنی بر دانش زبانی: این روش‌ها عموماً مبتنی بر قوانین حاکم بر دانش زبانی بوده و از مواردی مانند قوانین دستور زبان، تحلیل واژگان، تحلیل گفتمان و تحلیل نحوی استفاده می‌کنند.
- رویکردهای آماری: این رویکردها از ویژگی‌های آماری پیکره متن استفاده می‌کنند. سرعت و سهولت استفاده و مستقل بودن از زبان را می‌توان به عنوان مزیت‌های این روش‌ها به شمار آورد.
- رویکردهای یادگیری ماشین: این روش‌ها عموماً جزء روش‌های یادگیری با نظارت هستند. در قدم اول برای آموزش مدل از مجموعه داده‌های برچسب‌گذاری شده استفاده می‌شود، سپس از مدل آموزش دیده برای استخراج کلمات کلیدی سایر اسناد استفاده می‌کنند.

---

[1] Feldman & Dagan
[2] Document Classification
[3] Document Clustering
[4] Information Retrieval-IR
[5] Information Extraction-IE
[6] Web Mining
[7] Natural Language Processing - NLP
[8] Concept Extraction

• رویکرد مبتنی بر شبکه: این روش‌ها عموماً بدون نظارت بوده و به کمک شبکه ساخته شده از کلمات یا عبارات متن و با به کارگیری معیارهایی که ساختار شبکه را کمی‌سازی می‌کنند (از قبیل معیار درجه[۱]، معیار بینابینی[۲]، معیار نزدیکی[۳]، شباهت کسینوسی[۴]، معیار بردار ویژه[۵]، معیار رتبه‌بندی صفحه[۶] [۸]) رئوس را رتبه‌بندی می‌کنند [۶].

در حوزه استخراج کلمات کلیدی تحقیقات بسیاری انجام شده است که تعدادی از آن‌ها مورد بررسی قرار می‌گیرد. متونی مورد بررسی در این تحقیقات به زبان‌های فارسی، انگلیسی و چینی هستند.

پژوهش گزنی [۹] را می‌توان از نخستین تلاش‌ها برای استخراج کلمات کلیدی از متون فارسی دانست. گزنی برای استخراج کلمات کلیدی از سه معیار فراوانی کلمات، مجاورت مکانی واژگان با یکدیگر و موقعیت مکانی کلمات استفاده کرده است. در روش گزنی کلماتی با تکرار بیشتر یا کمتر از یک حد مشخص در نظر گرفته نمی‌شوند. در مرحله بعد کلماتی که دارای ۴ حرف اول یکسان هستند با هم ترکیب می‌شوند تا به نوعی ریشه‌یابی انجام شده باشد. نویسنده میزان خطای این روش را کمتر از ۰/۵ درصد بیان کرده است. گزنی در نهایت کلمات کلیدی را از میان کلماتی که فراوانی بالایی دارند و در متن در مجاورت یکدیگرند انتخاب می‌کند.

حاجی‌پور و سدیدپور [۱۰] روشی بدون نظارت و خودکار برای استخراج کلمات کلیدی در زبان فارسی پیشنهاد دادند. آن‌ها روشی را پیشنهاد کردند که با آموزش مدل word۲vec روی متون کوتاه فارسی، معنا و مفهوم متن را درک کرده، سپس با ترکیب روش آماری و یادگیری ماشین، کلمات کلیدی را انتخاب می‌کند. آن‌ها در ادامه برای ارزیابی روش پیشنهادی خود، نتایج به دست آمده از مدل پیشنهادی را با نتایج روش‌های رتبه‌بندی صفحه و TF-IDF[۷] و روش نسبتاً جدید Yake [۱۱] مقایسه کردند.

حجازی و نصیری [۱۲] یک روش ترکیبی با استفاده از ۳ ویژگی آماری (TF-IDF، TF-ISF[۸] و بیشترین فراوانی کلمات) و دسته‌بند بیز برای استخراج کلمات کلیدی ارائه دادند. در روش پیشنهادی حجازی و نصیری، پس از استخراج کلمات کلیدی برای کاهش تعداد کلمات، پس‌پردازش انجام می‌شود. آن‌ها ۱۴۱ پایان‌نامه در رشته‌های علوم انسانی، فنی و مهندسی، هنر، علوم پایه و پزشکی را به عنوان داده‌های مورد بررسی انتخاب کردند.

در پژوهشی دیگر آذرافزا و همکاران [۱۳] به استخراج کلمات کلیدی از متن‌های کوتاه فارسی منتشر شده در شبکه های اجتماعی پرداختند. آن‌ها پس از پیش‌پردازش داده‌ها، نقش دستوری کلمات را مشخص کرده و اسامی خاص و کلماتی که نقش دستوری اسم و صفت دارند را به عنوان کلمات منتخب برگزیدند. سپس این کلمات، نرمال شده و مواردی که در عین پرتکرار بودن غیر مؤثر هستند به صورت دستی حذف می‌شوند. در نهایت شبکه‌ای بر اساس رابطه هم‌رخدادی بین کلمات ترسیم شده (طول پنجره در این تحقیق ۳ است) و کلمات بر اساس معیار رتبه‌بندی صفحه امتیاز داده می‌شوند. در نهایت نتایج این روش با نتایج حاصل از ۳ روش استخراج کلیدواژه TF-IDF، Gensim و LDA[۹] مقایسه می‌شود. نتایج آذرافزا و همکاران نشان می‌دهد که دقت روش پیشنهادی از سه روش دیگر بیشتر است.

ژو[۱۰] و همکاران [۱۴] روشی برای استخراج کلمات کلیدی از متون چینی پیشنهاد دادند. آن‌ها ابتدا کلماتی را که به تعدد کنار هم دیده شده‌اند با هم ترکیب کردند. سپس کلماتی که نقش دستوری اسم، فعل و کلمات ناشناخته (مثلاً کلماتی که به زبان انگلیسی هستند، اختصارات و کلمات طولانی) دارند انتخاب‌شده و با آن‌ها شبکه‌ای از کلمات تشکیل

---

[1] Degree centrality
[2] Betweenness centrality
[3] Closeness centrality
[4] Cosine similarity
[5] eigenvector centrality
[6] Page-rank centrality
[7] Term Frequency - Inverse Document Frequency
[8] Term Frequency - Inverse Sentence Frequency
[9] Linear Discriminant Analysis
[10] Zhi Zhou

می‌دهند. درروش ژو و همکاران ارتباط بین واژه‌ها بر اساس رابطه هم‌رخدادی با پنجره‌ای به طول ۳ تعریف می‌شود. آن‌ها وزن یال‌ها را از ضریب ژاکارد تعیین می‌کنند به این صورت که تعداد تکرار دو کلمه در کنار هم محاسبه شده و بر اساس فراوانی دو کلمه نرمال می‌شود. هرچه این ضریب بزرگ‌تر باشد، وزن یال بیشتر است. آن‌ها در نهایت امتیاز گره‌ها را از ترکیب مقدار I DF، طول کلمات و معیار نزدیکی محاسبه کرده و کلماتی که بیشترین امتیاز را دارند به عنوان کلیدواژه انتخاب می‌کنند.

دیدیر[1] و همکاران [۱۵] شبکه هم‌رخدادی بین کلمات را ترسیم می‌کنند. آن‌ها برای تشکیل این شبکه طول پنجره هم رخدادی را ۳ در نظر می‌گیرند. سپس معیارهای مرکزیت را برای شبکه محاسبه کرده و دقت آن‌ها را می‌سنجند. دیدیر و همکاران ۹ معیار مرکزیت را مورد بررسی قرار دادند و در نهایت یک معیار ترکیبی پیشنهاد کردند. آن‌ها بیان کردند که ۹ معیار مرکزیت ذکر شده به تنهایی نتایج و دقت تقریبا مشابهی را ارائه می‌دهند اما معیار ترکیبی پیشنهادی، دقت را بهبود می‌بخشد. دیدیر و همکاران از ۳ پایگاه داده‌ی مختلف شامل چکیده مقالات در حوزه مهندسی و فیزیک، متن مقالات علمی که از کتابخانه دیجیتال AMC[2] گرفته شده و اخبار وب در ۱۰ حوزه متفاوت از جمله فرهنگ، تجارت، ورزش و فناوری استفاده کردند.

بیسواس[3] [۱۶] در پژوهش خود برای استخراج کلمات کلیدی متن‌های کوتاه انگلیسی از شبکه هم‌رخدادی استفاده کرده است. او شبکه‌ای وزن‌دار ترسیم کرده که در آن یال‌ها با توجه فراوانی تکرار دو کلمه‌ای که یال واصل بین آن دو است، وزن‌دهی می‌شوند. بیسواس از معیار مرکزیت گره‌یال[4] [۱۷] و نزدیکی برای رتبه‌بندی رئوس شبکه استفاده کرده است.

با وجود اینکه استخراج کلمات کلیدی از متن‌های فارسی در سال‌های اخیر مورد توجه قرار گرفته است، اما پژوهشگران این حوزه به دلیل ماهیت عناصر زبان‌شناختی و نگارشی آن با چالش‌هایی روبرو هستند. در این تحقیق سعی شده است تا با استفاده از سه رویکرد آماری (مولفه TF-I DF)، دانش زبانی (نقش دستوری کلمه‌ها) و ساخت شبکه (معیار مرکزیت نزدیکی)، کلمات کلیدی از متن‌های طولانی و ساختارنیافته فارسی استخراج شوند. با گسترش مدلی بر اساس این ۳ مولفه سعی می‌شود که کلمات از جنبه‌های مختلف بررسی شده و بهترین نتیجه حاصل شود. مقدار TF-I DF فقط از جنبه آماری و میزان تکرار، اهمیت کلمات را بررسی می‌کند. معیار مرکزیت نزدیکی نیز بر اساس شبکه‌ی هم‌رخدادی کلمات محاسبه شده و ماهیت آن‌ها را در نظر نمی‌گیرد. به همین خاطر در این تحقیق از ۳ مولفه گفته شده به صورت هم‌زمان برای استخراج کلمات کلیدی استفاده می‌شود.

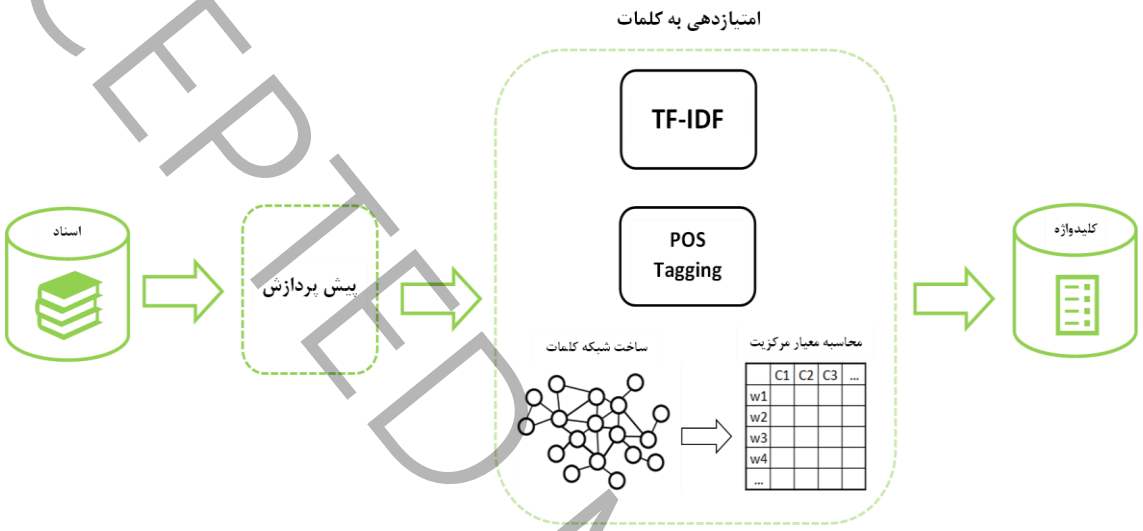
الگوریتم پیشنهادی به منظور بررسی تطابق بین منابع آموزشی و آیین‌نامه‌های اجرایی استفاده می‌شود. آیین‌نامه بتن ایران (آبا) و کتاب سازه‌های بتن‌آرمه (تألیف دکتر مستوفی‌نژاد) به عنوان نمونه جهت بررسی انتخاب شده‌اند. در این تحقیق به کمک فنون متن‌کاوی، کلمات مهم و کلیدی دو سند استخراج شده و مقایسه‌ای فنی بین این دو ارائه می‌شود. از روش پیشنهادی در این تحقیق می‌توان به عنوان ابزاری برای کمک به سیاست‌گذاران آموزش عالی کشور در بازبینی منابع آموزشی با هدف متناسب‌سازی محتوای آموزشی با نیاز صنعت استفاده کرد.

در بخش دوم این مقاله الگوریتمی برای استخراج کلمات کلیدی پیشنهاد شده است. الگوریتم پیشنهادی در بخش سوم بر روی نمونه‌های موردی پیاده‌سازی می‌شود. در بخش چهارم نتایج به دست آمده تحلیل و بررسی شده و مقایسه‌ای بین خروجی آیین‌نامه با یکی از کتاب‌های مرجع بتن در دانشگاه‌های ایران انجام می‌شود. در بخش نهایی نتیجه‌گیری آورده شده است.

[1] Didier A. Vega-Oliveros
[2] https://dl.acm.org/
[3] Saroj Kumar Biswas
[4] Node Edge Rank

۲- روش کار

در این بخش یک الگوریتم برای استخراج کلمات کلیدی از متون فارسی پیشنهاد می‌شود. ابتدا به زبان پایتون کدی توسعه داده شده و فایل متنی سند مورد بررسی را دریافت می‌کند. در قدم بعدی متن ورودی در ۷ مرحله پیش‌پردازش می‌شود. سپس کلمات با سه معیار TF-I DF (آماری)، نقش دستوری کلمات (زبانی) و معیار مرکزیت نزدیکی (تحلیل شبکه) امتیازدهی می‌شوند. امتیاز نهایی از ضرب این سه مقدار به دست آمده و کلمات کلیدی انتخاب می‌شوند. مراحل استخراج کلمات کلیدی به طور خلاصه در شکل ۱ آمده است.



شکل ۱ : مراحل استخراج کلمات کلیدی

Figure 1: Steps for extracting keywords

۲-۱- فراخوانی فایل متنی: الگوریتم ابتدا فایل متنی مورد نظر را فراخوانی می‌کند.

۲-۲- پیش‌پردازش: همان‌طور که قبلا اشاره شد داده‌های متنی غالبا ساختارنیافته هستند و قبل از پردازش توسط رایانه نیاز به آماده‌سازی دارند. این امر با انجام پیش‌پردازش میسر می‌شود. دقت خروجی تا حد زیادی به پیش‌پردازش بستگی دارد [۱۸]. در این تحقیق گام‌های زیر برای آماده‌سازی متن انجام شده است.

- شکستن جملات به کلمات مجزا
- حذف اعداد و علائم نگارشی
- نرمال‌سازی: هدف این ابزار، تمیز و مرتب کردن متن است. همچنین با جایگزین کردن نویسه‌های استاندارد، یکسان‌سازی متن انجام می‌شود.
- ریشه‌یابی: در این مرحله کلمات به شکل پایه‌ای خود درمی‌آیند تا از تنوع آن‌ها کاسته شود.
- حذف کلمات توقف: کلمات توقف یا ایست‌واژه‌ها کلماتی هستند که به تنهایی معنی خاصی را نمی‌رسانند و اطلاعات مفیدی دربر ندارند. حذف این کلمات موجب بهبود در سرعت و عملکرد پردازش می‌شود.
- حذف کلمات با طول کمتر از ۲ نویسه
- هرس کردن: حذف کلماتی که تکرار آن‌ها در پیکره اسناد کمتر از ۳ است.

پس از انجام پیش‌پردازش متن یکپارچه‌تر شده و آماده پردازش اصلی می‌شود. سپس، الگوریتم به منظور انتخاب کلمات کلیدی به کلمات بر اساس سه معیار TF-I DF (آماری)، نقش دستوری کلمات (زبانی) و معیار مرکزیت نزدیکی (تحلیل شبکه) امتیاز می‌دهد. امتیاز نهایی کلمات از ضرب این سه مقدار محاسبه می‌شود.

۳-۲- محاسبه مقدار TF-I DF: روش TF-I DF که توسط سالتون و بالکی [۱۹] معرفی شد، اهمیت هر کلمه در یک سند را با توجه به تکرار آن در مجموعه‌ای از اسناد محاسبه می‌کند. تعیین شاخص بودن هر کلمه در یک سند مزیت این روش است به این صورت که هرچه مقدار عددی TF-I DF کلمه‌ای بزرگ‌تر باشد، آن کلمه مهم‌تر بوده و می‌تواند به عنوان شاخصه‌ی آن سند تلقی شود. از آنجایی که معیار فراوانی (TF) قادر به شناسایی کلمات کلیدی با تکرار پایین نیست، سالتون و بالکی این روش را با معرفی مولفه I DF بهبود بخشیدند. توضیحات بیشتر در مورد سازوکار این روش در [۲۰] آمده است.

۴-۲- تعیین نقش دستوری کلمات: بر اساس تحقیقات آماری انجام شده (به عنوان نمونه [۱۴] و [۲۱]) کلماتی با نقش دستوری اسم یا فعل نسبت به قید یا ضمیر شانس بیشتری برای انتخاب به عنوان کلیدواژه دارند. در جدول ۱ امتیاز کلمات بر اساس نقش دستوری آن در جمله تعیین شده است.

Table 1: Steps for extracting keywords

جدول ۱ : ضرایب اهمیت کلمات بر اساس نقش دستوری

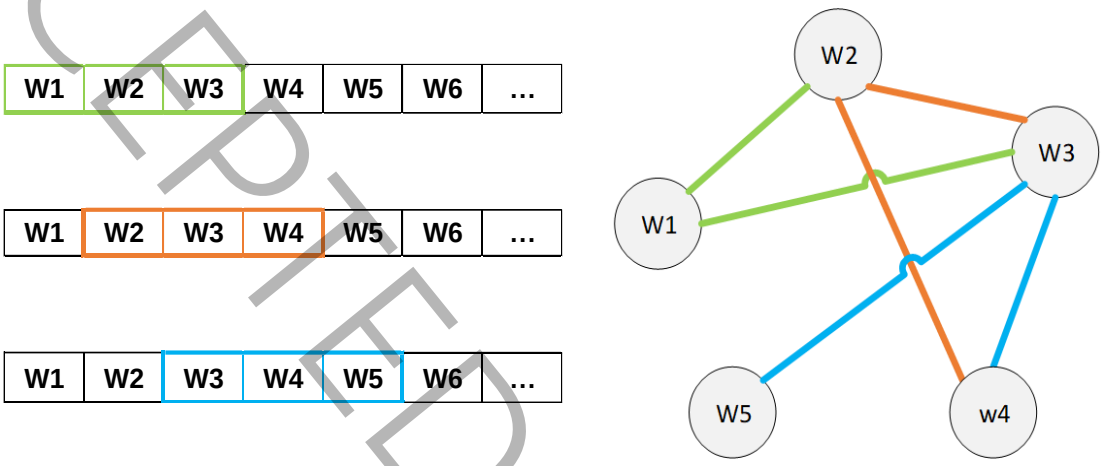
نقش دستوری کلمه	ضریب
اسم	۱/۲
فعل، صفت	۱
اسامی ناشناخته	۱
سایر	۰/۲

۵-۲- محاسبه معیار مرکزیت نزدیکی: برای محاسبه معیار مرکزیت نزدیکی ابتدا باید شبکه کلمات را توسعه داد. شبکه یک مدل ریاضی است که امکان کاوش در روابط و اطلاعات ساختاری را فراهم می‌کند. در این روش سند به صورت یک شبکه مدل‌سازی می‌شود به طوری که عبارات (کلمات) با رأس‌ها و روابط بین آن‌ها با یال‌ها نمایش داده می‌شود. رابطه بین کلمات را می‌توان با توجه به نیاز تحقیق به شکل‌های مختلفی تعریف کرد (به عنوان نمونه هم‌رخدادی، رابطه معنایی و رابطه نحوی). روش‌های مبتنی بر شبکه برای استخراج کلمات کلیدی از بسیاری از جهات ساده و قوی هستند. از مزایای این روش می‌توان به عدم نیاز به دانش زبانی پیشرفته، مستقل بودن از زبان و حفظ رابطه بین کلمات با ترسیم یال بین آن ها و جزء روش‌های بدون نظارت بودن اشاره کرد [۶].

در نظریه گراف معیارهای مرکزیت به شاخص‌هایی اطلاق می‌شوند که مهم‌ترین رئوس را در یک نمودار مشخص کرده و آن‌ها را رتبه‌بندی می‌کنند. در حوزه استخراج کلمات کلیدی معیارهای مرکزیت مختلفی پیشنهاد شده است که بسته به روش و هدف تحقیق مورد استفاده قرار می‌گیرند [۲۲].

در این مقاله برای ساخت شبکه، کلمات به عنوان رئوس در نظر گرفته می‌شوند. یال‌های واصل این رأس‌ها بر اساس رابطه هم‌رخدادی ترسیم می‌شوند. منظور از هم‌رخدادی حداکثر فاصله بین کلمات در متن است به طوری که آن دو کلمه از لحاظ معنایی به یکدیگر مرتبط باشند. اگر طول این پنجره (فاصله) خیلی کوچک باشد، برخی از ارتباط‌های دوربرد بین

کلمات نادیده گرفته می‌شود و در صورتی که طول این پنجره خیلی بزرگ باشد اطلاعات اضافی زیادی به وجود آمده و شبکه‌ی درهم‌تنیده‌ای تشکیل می‌شود [۱۴]. با توجه به بررسی‌های انجام شده در تحقیقات قبلی ([۱۳] و [۱۳]) بهترین انتخاب برای طول پنجره هم‌رخدادی عدد ۳ است. به همین دلیل در این تحقیق طول ۳ برای پنجره هم‌رخدادی انتخاب شده است. در شکل ۲ نحوه گسترش شبکه در یک متن که حاوی تعدادی کلمات پشت سر هم است، نمایش داده شده است. همان‌طور که در شکل ۲ مشخص است بین کلماتی که در یک پنجره قرار می‌گیرند یالی ترسیم شده و به همین ترتیب شبکه گسترش می‌یابد.



شکل ۲ : نحوه گسترش شبکه با پنجره‌ای به طول ۳

Figure 2: How to expand the network with a window length of 3

معیار مرکزیت نزدیکی با استفاده از شبکه رسم شده برای تمامی گره‌ها محاسبه می‌شود. معیار مرکزیت نزدیکی هر گره برابر با معکوس مجموع فاصله‌ی آن گره تا سایر گره‌های شبکه است. رابطه (۱)، معیار مرکزیت نزدیکی نرمال شده را محاسبه می‌کند. هر مقدار که معیار مرکزیت نزدیکی گره i بزرگ‌تر باشد به معنی اهمیت بیشتر آن است (مجموع فواصل گره i از سایر گره‌های شبکه کمتر بوده و به مرکز شبکه نزدیک‌تر است).

$$C_c(i) = \frac{n-1}{\sum_{u=1}^{n-1} d(u,i)} \qquad (۱)$$

در رابطه (۱)، $C_c(i)$ معیار مرکزیت نزدیکی برای گره i است. n تعداد کل گره‌های شبکه است و $n-۱$ تعداد گره‌هایی است که از گره i می‌توان به آن‌ها رسید. $d(u,i)$ طول کوتاه‌ترین مسیر موجود بین دو گره i و u را مشخص می‌کند.

۶-۲- تعیین کلمات کلیدی: امتیاز نهایی کلمات (S_W) طبق رابطه (۲)، از ضرب سه مقدار $tf\text{-}idf$ ($S_{tf\text{-}idf}$)، اهمیت نقش دستوری ($S_{POS\text{-}tagging}$) و معیار نزدیکی (S_C) محاسبه می‌شود.

$$S_W = S_{tf-idf} * S_{POS-tagging} * S_C \qquad (۲)$$

در هر فصل کلماتی که بیشترین امتیاز را گرفتند به عنوان کلیدواژه‌های آن فصل انتخاب می‌شوند. مجموعه‌ی این کلمات، کلمات کلیدی سند اصلی را تشکیل می‌دهند.

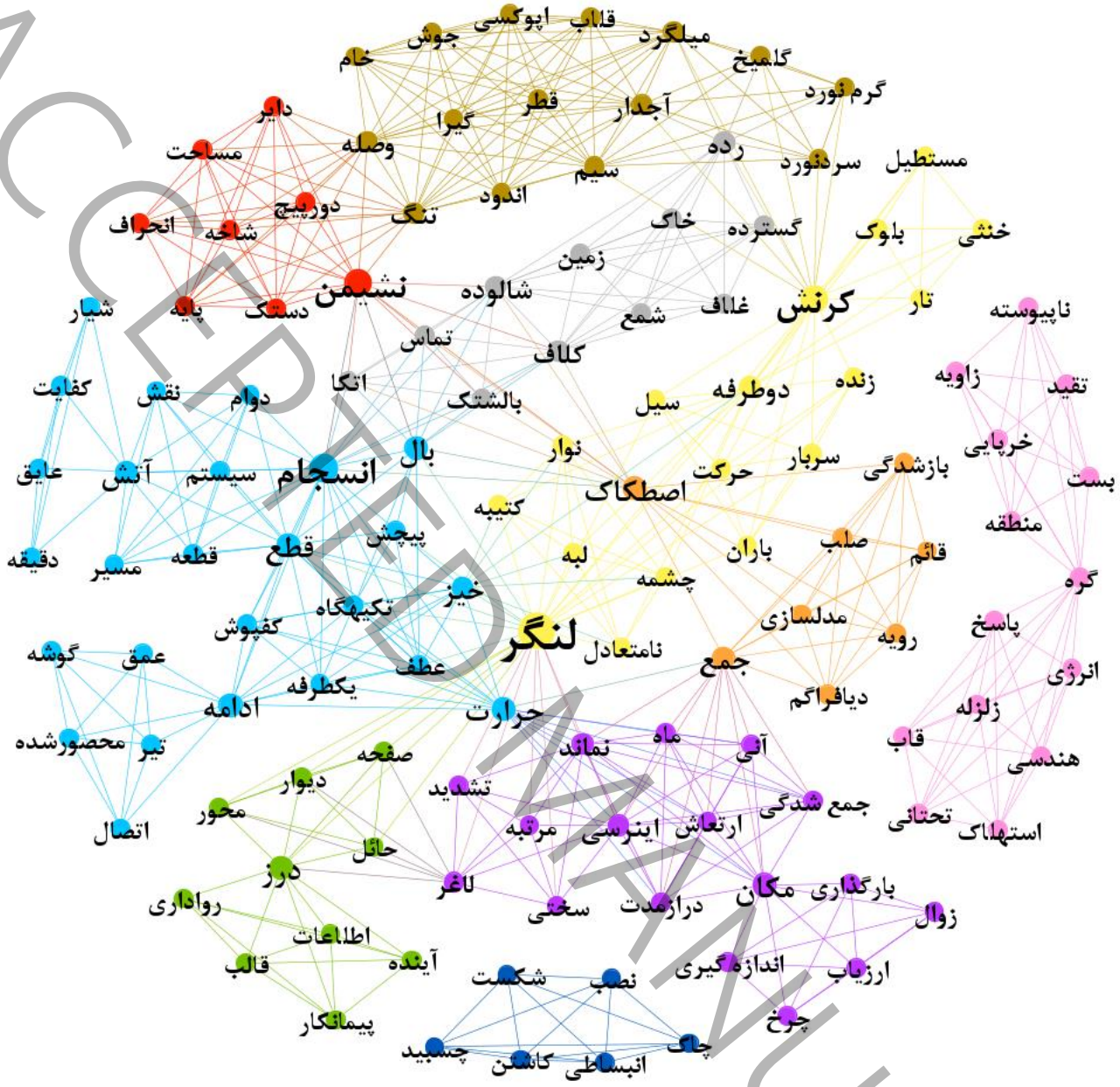
۳- پیاده‌سازی

در روش پیشنهادی ابتدا کدی بر اساس پایتون نسخه ۳,۱۰,۶ پیاده‌سازی و اجرا شده است. داده‌ی مورد بررسی، آیین نامه بتن ایران (آبا) جلد اول (تحلیل و طراحی) ویرایش ۱۴۰۱/۰۱/۰۱ است. ابتدا فایل pdf به کمک نرم‌افزارهای آنلاین به فایل متنی تبدیل می‌شود. پس از طی مراحل پیش‌پردازش تعداد کلمات متمایز به ۱۰۵۵ کلمه کاهش پیدا می‌کند. سپس مقدار TF-IDF این کلمات محاسبه می‌شود. بدین‌صورت که هر فصل به عنوان یک سند مجزا در نظر گرفته شده و اهمیت کلمات در هر فصل محاسبه می‌شود. پس از آن، کلمات برچسب‌گذاری شده و ضریب اهمیت آن‌ها در مقدار TF-IDF ضرب می‌شود. در مرحله آخر شبکه‌ای از کلمات با طول پنجره ۳ ترسیم می‌شود که این شبکه دارای ۱۰۵۵ گره (به تعداد کلمات متمایز) و ۳۰۷۲۱ یال است. معیار مرکزیت نزدیکی را برای تمام گره‌های شبکه محاسبه کرده و در دو مقدار قبلی ضرب می‌کنیم. درنهایت برای استخراج لیست کلیدواژه‌ها، کلمات با توجه به امتیاز نهایی رتبه‌بندی شده و سپس واژه‌هایی که بیشترین امتیاز را دارند به عنوان کلمات کلیدی انتخاب می‌شوند.

از آنجایی که ارائه اطلاعات به صورت یک تصویر بسیار کارآمدتر است (زیرا انسان‌ها به تصاویر حساس‌تر هستند و انتقال اطلاعات از طریق تصاویر شهودی‌تر است [۲۱])، برای ارائه تفسیر و نتیجه‌گیری ابتدا شبکه خوشه‌بندی شده کلمات به کمک نرم‌افزار متن‌باز[1] Gephi [۲۳] نسخه‌ی ۰,۱۰ ترسیم شده و خوشه‌بندی می‌شوند. تصویر شبکه خوشه‌بندی شده کلیدواژه‌های آیین‌نامه آبا در شکل ۳ آورده شده است. این شبکه شامل ۱۳۲ گره و ۵۶۹ یال است که کلمات در ۱۰ خوشه دسته‌بندی شده است.

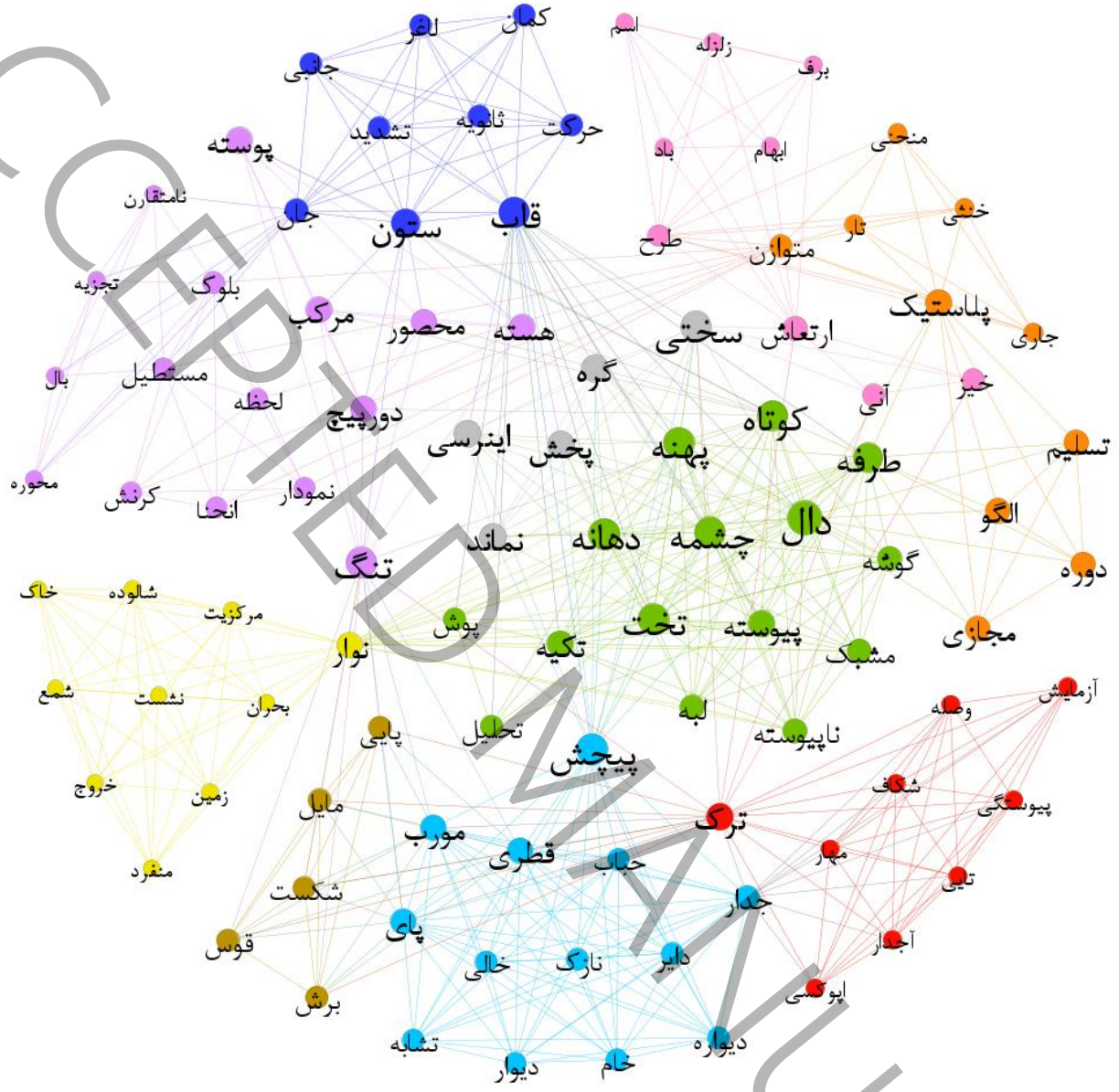
تمامی مراحلی که در بخش ۲ بیان شد، بر روی کتاب «سازه‌های بتن‌آرمه» تالیف دکتر داود مستوفی‌نژاد (این کتاب شامل ۱۸ فصل است و در دو جلد چاپ شده که در کل حدود ۱۴۰۰ صفحه است) نیز انجام می‌شود. شبکه خوشه‌بندی این کتاب در شکل ۴ آمده است. این شبکه شامل ۱۰۱ گره و ۵۴۸ یال است که در ۱۰ خوشه مختلف طبقه‌بندی شده است.

[1] Open source



شکل ۳ : شبکه خوشه‌بندی شده کلیدواژه‌های آیین‌نامه بتن ایران

Figure 3: Clustered network of keywords from the Iran's Concrete Regulations



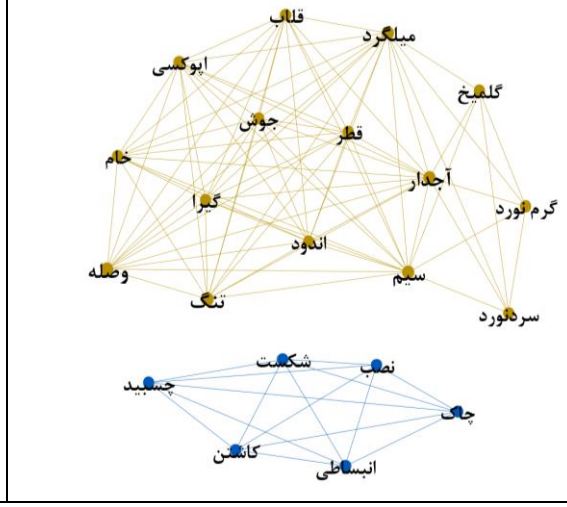
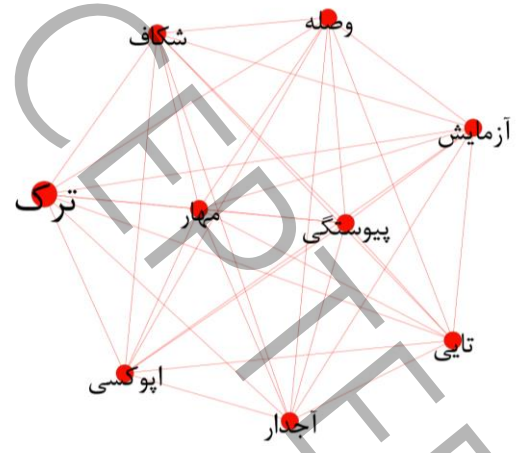
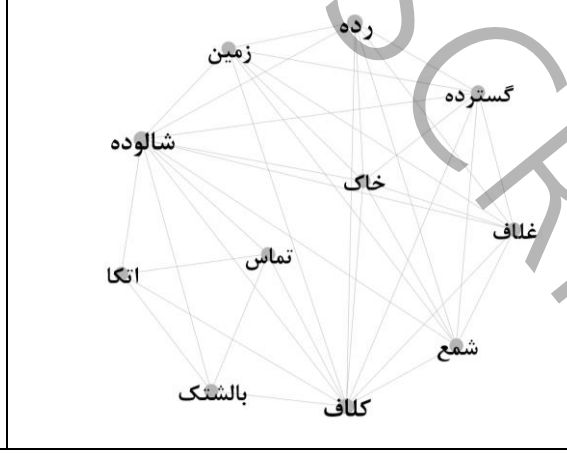
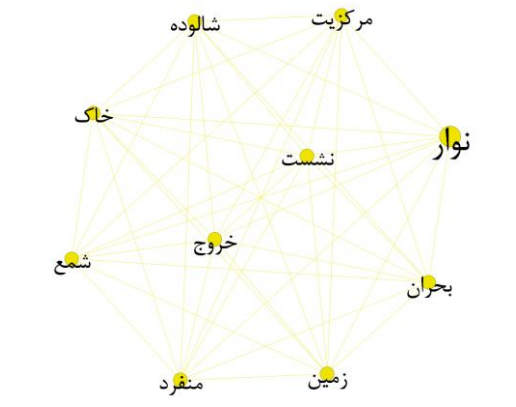
شکل ۴ : شبکه خوشه‌بندی شده کلیدواژه‌های کتاب سازه‌های بتن‌آرمه

Figure 4: Clustered network of keywords from the book "Reinforced Concrete Structures"

۴- تحلیل و بررسی

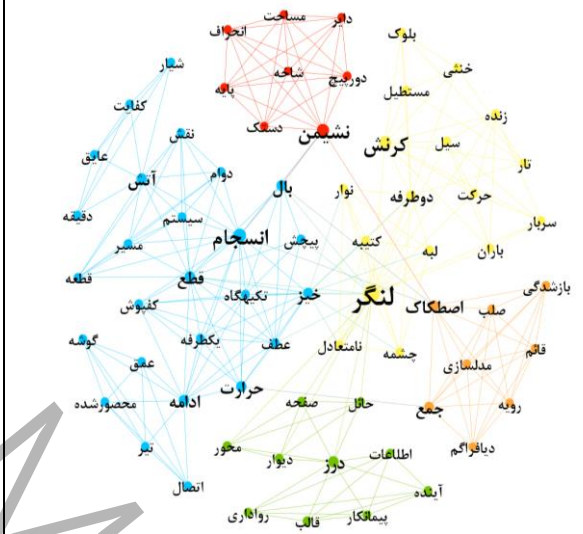
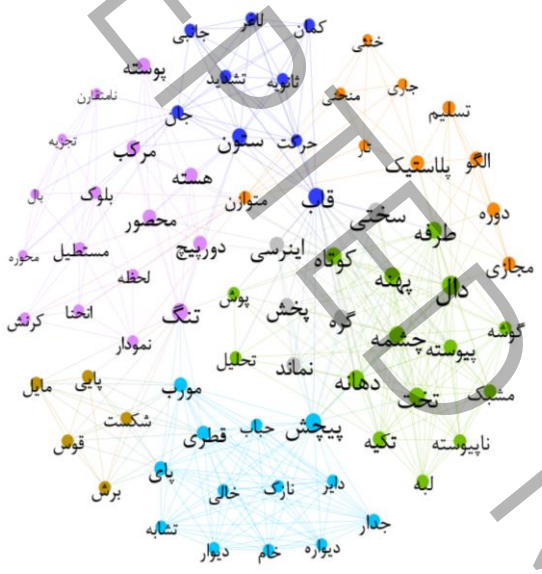
Table 2: Information on clustered keyword networks

جدول ۲- اطلاعات شبکه‌های خوشه‌بندی شده کلیدواژه‌ها

آیین‌نامه بتن ایران	کتاب سازه‌های بتن‌آرمه
	
خوشه قهوه‌ای رنگ از کتاب آبا به مسلح‌سازی بتن و مسائل پیرامون آن می‌پردازد. میلگرد و شبکه سیمی جوش‌شده از رایج‌ترین مصالح برای مسلح کردن بتن هستند که در این قسمت به آن‌ها اشاره شده است. همچنین روش‌های مختلف تولید میلگرد، انواع مصالح برای مسلح سازی بتن (میلگرد آجدار، میلگرد با اندود حفاظتی اپوکسی، خاموت، گلمیخ، شبکه سیمی جوش‌شده) و موارد مرتبط با مهار فولاد به بتن مثل طول گیرایی و قلاب، تنگ و وصله نیز آمده است. در صورتی که مهار فولاد به بتن به درستی صورت نگیرد، عضو بتن مسلح کارایی خود را از دست خواهد داد. با اشاره به کلمه قطر در مورد مقدار مناسب فولاد در عضو بتنی نیز صحبت شده است. همچنین خوشه سورمه‌ای رنگ در مورد مهار اعضای بتنی به یکدیگر و اتصال قطعات الحاقی است. در این خوشه به مهار کاشتنی چسبی، انبساطی و زیرچاکی اشاره شده است. همچنین در مورد تاثیر شکست بتن بر روی مهارها صحبت شده است. خوشه قرمز از کتاب سازه‌های بتن‌آرمه نیز در مورد مسلح‌سازی بتن و مهار فولاد به بتن است که تقریبا منطبق با آیین‌نامه است البته آیین‌نامه به جزئیات بیشتری از مصالح مورد استفاده در مسلح‌سازی و مهار فولاد به بتن اشاره کرده است. اما کتاب سازه‌های بتن‌آرمه بیشتر از منظر اجرایی روی این موضوع تاکید دارد. مثلا اشاره به کلماتی مثل ترک و شکاف، گویای این موضوع است.	
آیین‌نامه بتن ایران	کتاب سازه‌های بتن‌آرمه
	

موضوع کلمات خوشه خاکستری رنگ آبا، شالوده‌ها است. در این قسمت انواع شالوده (گسترده و شمع)، وجود یا عدم وجود غلاف در شمع و نقاط اتکایی آن‌ها آمده است. همچنین کلمات «تماس» و «خاک» به اهمیت نقطه تماس شالوده با خاک اشاره دارد که در طراحی باید به آن توجه شود. علاوه بر این در مورد کلاف‌بندی فونداسیون نیز صحبت شده است.

خوشه زرد رنگ کتاب سازه‌های بتن‌آرمه نیز به شالوده‌ها پرداخته است. کلمات این خوشه تقریبا منطبق بر آیین‌نامه است اما وجود کلماتی مثل «خروج»، «بحران» و «نشست» در این خوشه نشانگر تاکید این کتاب بر جزئیات طراحی و بارگذاری در شالوده‌هاست.

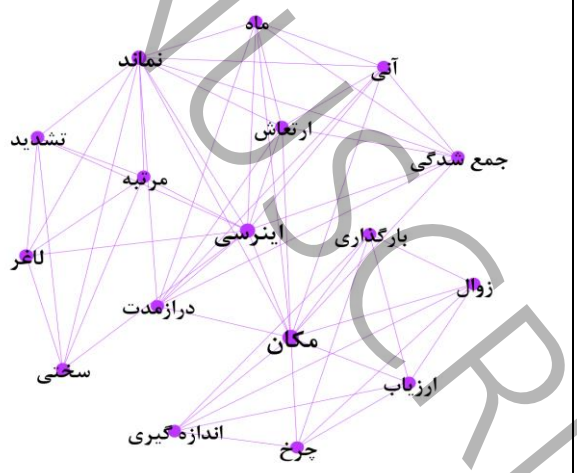
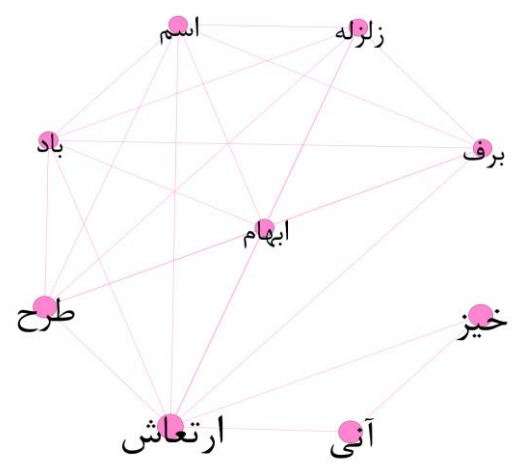
آیین‌نامه بتن ایران	کتاب سازه‌های بتن‌آرمه
	

پنج خوشه‌ی مشخص شده از آبا در شکل بالا، در مورد طراحی عضوهای بتنی (ستون، دال، دیافراگم، دیوار) و بارهای وارد بر سازه است. هفت خوشه مشخص شده در کتاب سازه‌های بتن‌آرمه نیز به موارد مشابه اشاره کرده است.

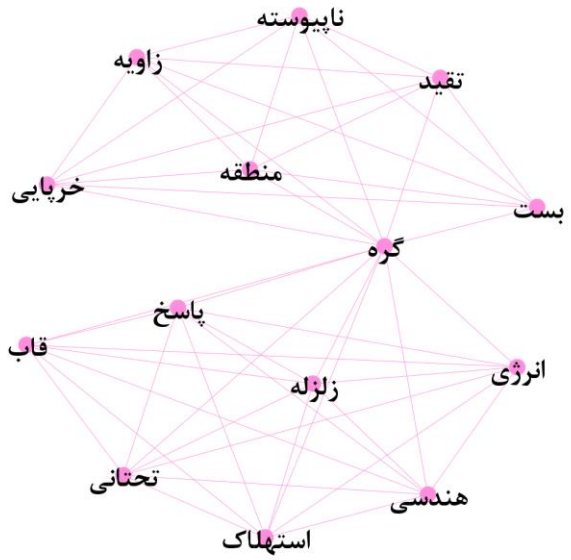
در آبا خوشه قرمز رنگ در مورد ستون و موارد مهم مرتبط با طراحی و اجرای آن است. در این بخش علاوه بر کفایت سطح مقطع ستون و مسلح‌سازی آن در مورد اتصالات تیر به ستون (نشیمن) و طراحی لرزه‌ای ستون‌ها (بالشتک) نیز صحبت شده است. علت تاکید بر کفایت سطح مقطع ستون حصول اطمینان از تامین طول مهاری میلگردهای تیر متصل به آن، ایجاد فضای کافی برای آرماتورگذاری، جلوگیری از افزایش لاغری و کمانش ستون است. خوشه نارنجی رنگ در مورد دیافراگم‌ها و انواع آن است. دیافراگم‌ها (متشکل از اجزای لبه و جمع‌کننده) مجموعه‌ای از اجزای افقی سازه هستند که نیروهای جانبی وارد بر سازه را به اعضای قائم باربر منتقل می‌کنند. این انتقال معمولا از طریق برش اصطکاکی انجام می‌شود که در این خوشه به آن اشاره شده است. همچنین از عوامل مهمی که بر رفتار دیافراگم تاثیر می‌گذارد، وجود بازشو در آن است که در این خوشه با آوردن کلمه بازشو توجه مهندسین را به این موضوع جلب کرده است. خوشه آبی رنگ در مورد دال‌های یک‌طرفه، تیرها و اتصالات است. در مورد دال‌ها، به تکیه‌گاه‌ها، پیچش در دال‌ها، قطع آرماتور، آرماتور انسجام در دال‌ها، کفپوش بتنی یکپارچه با دال، نقاط عطف در دال‌ها اشاره شده است. در بخشی دیگر از این خوشه در مورد مقاومت در برابر آتش است. علت ادغام این دو موضوع در یک خوشه اشتراک در کلمه «مسیر» است که در دال‌ها مسیر جریان برش و در آتش‌سوزی مربوط به مسیر گسترش حرارت است. خوشه زرد رنگ در مورد ترکیبات بارگذاری، مسائل مربوط به ارزیابی تنش ایجاد شده در مقاطع بتنی تحت بارهای مختلف و دال‌های دوطرفه است. در طراحی دال‌ها به مواردی مثل چشمه، لبه، نوار، لنگر، برش پانچینگ در دال‌ها (کتیبه) و نیروی نامتعادل کننده اشاره شده است. علت ادغام این موضوعات در یک خوشه اشتراک همه آن‌ها در کلمه لنگر است. خوشه سبز رنگ در مورد دیوار و الزامات اجرایی است. در این خوشه به بارهای داخل و خارج از صفحه دیوار، دیوار حائل، درز دیوار، بارهای محوری وارد بر

دیوارها اشاره شده است.

در کتاب سازه‌های بتن‌آرمه موضوع خوشه سورمه‌ای رنگ در مورد طراحی قاب‌ها و ستون‌های لاغر است. در این بخش به کمانش ستون‌های لاغر اشاره شده چراکه این ستون‌ها تحت تاثیر کمانش زیاد آسیب دیده و می‌شکنند. همچنین به مسئله تشدید لنگرها اشاره شده که باید در طراحی مدنظر قرار گیرد که اگر در محاسبه لنگر، رفتار غیرخطی منظور نشده باشد، باید مقدار لنگر در ضریب تشدید ضرب شود. همچنین به حرکت جانبی قاب و ستون اشاره شده که وجود یا عدم وجود آن در طراحی تاثیرگذار است. خوشه بنفش رنگ در مورد طراحی تیرهای بالدار و نامتقارن تحت خمش و ستون‌های کوتاه است. میلگرد عرضی در ستون‌ها علاوه بر نقش اجرایی و مهار میلگردهای طولی، در مقابل برش نیز مقاومت می‌کند. میلگرد عرضی در ستون‌ها به دو صورت تنگ بسته و دورپیچ اجرا می‌شود. در صورت اجرای دورپیچ، هسته بتنی ستون محصور شده و مقاومت فشاری آن افزایش می‌یابد. اما رفتار ستون با تنگ بسته متفاوت است و در صورت رسیدن به بار فشاری حداکثر پوسته بتن از باربری خارج شده و تا حدودی جدا می‌شود. در این بخش با اشاره به کلماتی مثل هسته، محصور، دورپیچ، تنگ و پوسته توجه مهندسین را به اهمیت این موارد در ستون‌ها جلب می‌کند. نوعی دیگر از ستون‌ها، ستون مرکب است که در اینجا به آن اشاره شده است. بخش دیگر این خوشه به طراحی تیرهای بالدار و نامتقارن و خمش دومحوره اشاره دارد که حالات مختلف مقاطع بالدار و نامتقارن را تحت خمش بررسی می‌کند. خوشه آبی و خوشه قهوه‌ای رنگ طراحی تیرهای بتنی تحت برش و پیچش را مورد بررسی قرار می‌دهند. علت نزدیک بودن این دو خوشه، تشابه روند طراحی آن‌ها است به این صورت که در مقاطع تحت بار پیچشی، تنش برش نیز ایجاد می‌شود. در این دو خوشه در مورد پیچش در مقاطع دایروی و مقاطع توخالی جدار نازک صحبت شده است. همچنین به روش تشابه حباب صابون اشاره شده است. استفاده از خاموت برشی مایل، انواع شکست تیرها در اثر برش، استفاده از تشابه خرپایی برای بررسی عضو بتن‌آرمه و عملکرد قوس تیر بتنی تحت بار خارجی از موارد دیگری است که در این دو خوشه آمده است. خوشه سبز و خوشه خاکستری و خوشه نارنجی رنگ در مورد دال‌ها و تحلیل آن‌ها است. در خوشه سبز به انواع دال‌های دوطرفه (تخت، دال تخت با پهنه، دال مشبک و دال با تیر) اشاره شده است. در مورد تکیه‌گاه‌های دال (لبه پیوسته یا ناپیوسته، لبه کوتاه یا بلند) و سختی آن‌ها صحبت شده است. همچنین به فولاد گذاری ویژه در گوشه دال، دهانه‌های دال، چشمه‌های باربر در دال‌ها، تحلیل دال‌ها، پوش نیروهای داخلی، الگوی خط تسلیم، رفتار پلاستیک و روش کار مجازی اشاره شده است.

آیین‌نامه بتن ایران	کتاب سازه‌های بتن‌آرمه
	

در آبا خوشه بنفش رنگ در مورد تحلیل سیستم‌های سازه‌ای و بهره‌برداری و مسائل پیرامون آن است. نقطه مشترک این دو موضوع بحث خیز (تغییر مکان)، ترک و ارتعاش است. خیز، ترک و ارتعاش تا حدی مجاز است که مانع بهره‌برداری ساختمان نشود و طراح باید علاوه بر ایمنی این موضوع را از دیدگاه بهره‌برداری و آرامش خاطر ساکنین نیز بررسی کند.

خوشه صورتی رنگ در کتاب سازه‌های بتن‌آرمه هم در مورد بهره‌برداری است اما به جزئیات کمتری اشاره شده است.	
آیین‌نامه بتن ایران	کتاب سازه‌های بتن‌آرمه
	
در آبا خوشه صورتی رنگ خوشه صورتی رنگ در مورد طراحی لرزه‌ای سازه‌ها صحبت می‌کند. کلمه کلیدی هندسی اشاره به نقش مهم شکل هندسی سازه‌ها در پاسخ به زلزله و مستهلک کردن انرژی دارد. از آنجایی که سازه‌های خرپایی یکی از انواع مصالح ضد زلزله بوده و برای پوشش ناپیوستگی بین بتن و فولاد به کار می‌روند، مدل‌های خرپایی نیز در این خوشه آمده است. کتاب سازه‌های بتن‌آرمه به طور خاص در مورد این موضوع صحبت نکرده است.	

با مقایسه موضوعات و کلمات کلیدی مطرح شده در دو منبع مورد بررسی می‌توان گفت که آیین‌نامه بتن ایران به صورت جامع تمامی موارد و موضوعات مطرح در طراحی، آنالیز، اجرا و مشخصات مواد تشکیل دهنده سازه‌های بتنی را مورد بررسی قرار داده است. درواقع این آیین‌نامه یک معیار و ملاک مناسب برای تشخیص درست و غلط در حوزه ساخت سازه‌های بتنی به شمار می‌رود. آبا کمک می‌کند تا سازه‌های بتنی از حداقل استانداردهای روز دنیا و ایمنی کافی برخوردار باشند. همچنین ایستایی، مقاومت و پایایی سازه را در شرایط مختلف تضمین می‌کند و همان‌طور که پیش‌تر اشاره شد، هر مهندس باید بتواند بر اساس اصول کلی طراحی و رعایت قوانین یک آیین‌نامه، طراحی صحیح و مطمئنی را ارائه دهد. اما با توجه به آنچه در جدول ۲ بیان شد، بخش عمده‌ای از کتاب سازه‌های بتن‌آرمه به بررسی روش‌های مختلف طراحی اعضای بتنی تحت بارهای مختلف پرداخته است و موارد مرتبط با طراحی این اعضا را با جزئیات اجرایی زیادی آورده است که این جزئیات اجرایی در آیین‌نامه وجود ندارد. در حقیقت می‌توان گفت که آیین‌نامه یک سند دستورالعملی است که در آن به تمام جزئیات پرداخته شده و در مورد همه مسائل پیرامون سازه‌های بتنی صحبت کرده است. اما کتاب سازه‌های بتن‌آرمه به دستورالعمل‌های طراحی بسنده نکرده و مفاهیم مرتبط با طراحی سازه‌های بتنی را به صورت جزئی و عمیق بررسی کرده که از آن می‌توان به عنوان یک منبع و راهنمای گام‌به‌گام برای آموزش طراحی عضوهای بتنی استفاده کرد.

۵- نتیجه‌گیری

در این پژوهش مدلی خودکار برای استخراج کلمات کلیدی از متن‌های طولانی فارسی پیشنهاد شده است. مدل پیشنهادی در مدت زمان کوتاهی محتوای یک سند متنی طولانی را به صورت مجموعه‌ای از کلمات خلاصه می‌کند که می‌تواند ابزار مناسبی برای بررسی محتوای متن‌های فارسی باشد. همچنین متخصصین و کارگروهان ارزیاب می‌توانند از روش پیشنهادی این پژوهش برای تهیه و تدوین پیش‌نویس آیین‌نامه‌ها استفاده کنند. اما به دلیل وجود محدودیت‌ها، اجرای این فرآیند روی متن‌های فارسی نسبت به زبان انگلیسی سخت‌تر است. به عنوان مثال در هر دو سند کلمه «نماند» به عنوان یک کلمه کلیدی آورده شده است. این کلمه همان کلمه مُمان (تکانه) است که در فرآیند ریشه‌یابی به این صورت تبدیل شده است. در حقیقت این کلمه به صورت مَمان (تغییر یافته فعل امر نَمان) خوانده شده و سپس به ریشه خود یعنی فعل ماضی «نماند» تبدیل شده است. مسئله‌ی خوانش‌های متفاوت یک کلمه یکی از چالش‌های زبان فارسی است که پیش‌تر به آن اشاره شد. نمونه‌ای دیگر از تشخیص اشتباه ریشه، انتخاب کلمه «خام» به عنوان یک کلمه کلیدی در آیین‌نامه آبا است که این واژه ریشه‌یابی شده کلمه «خاموت» می‌باشد. شاید این ایرادات را بتوان به دلیل تخصصی بودن اسناد مورد بررسی دانست، اما ابزارهای پردازش زبان فارسی هنوز به اندازه زبان انگلیسی جامع و قدرتمند نیستند. با وجود تمامی این چالش‌ها همچنان متن‌کاوی فرآیند مفید و موثری برای صرفه‌جویی در زمان و کشف دانش از متن است.

مدل پیشنهادی بر روی آیین‌نامه بتن ایران و کتاب طراحی سازه‌های بتن‌آرمه تالیف دکتر مستوفی‌نژاد (به عنوان یکی از کتب مرجع در دانشگاه‌های سراسر کشور) پیاده‌سازی شده است. مقایسه نتایج نشان می‌دهد که محتوای این دو سند کاملا مشابه نیست و موضوعاتی که بر آن‌ها تاکید شده با هم تفاوت‌هایی دارند. دلیل عدم تطابق کامل در محتوای این دو سند را می‌توان متفاوت بودن هدف نگارش آن‌ها در نظر گرفت. آیین‌نامه‌ها مشابه چراغ راهنمایی و رانندگی هستند که گاهی مهندسین را از انجام برخی از امور منع می‌کنند اما گاهی با نشان دادن رنگ زرد یادآور می‌شوند که باید محتاط بود و در موارد دیگر اجازه پیشروی می‌دهند. اما رویکرد کتاب دکتر مستوفی نژاد آموزش طراحی اعضای بتنی در شرایط مختلف است. تنوع و گستردگی مسائل مطرح شده در این کتاب به نحوی است که دانشجویان پس از تسلط بر انواع این مسائل، آمادگی نسبی جهت مواجهه با طرح‌های عملی خواهند داشت. ولی در صورت برخورد با مورد جدید باید حتما به آیین‌نامه رجوع کنند.

مدل پیشنهادی در این پژوهش بر روی دو متن طولانی فارسی پیاده‌سازی شد. به طور قطع هر چه بازه اطلاعات ورودی بیشتر در نظر گرفته شود می‌تواند به جامع‌تر بودن نتایج کمک کند، اما به دلیل محدودیت‌های اجرایی و زمانی امکان بررسی اسناد بیشتر میسر نشد. این مورد می‌تواند در تحقیقات آتی مورد توجه قرار گیرد. همچنین در این پژوهش طول پنجره هم‌رخدادی با استناد به تحقیقات گذشته انتخاب شده است که پیشنهاد می‌شود در مطالعات آینده این موضوع برای زبان فارسی به صورت خاص مورد بررسی قرار گیرد.

۶- منابع

[1] J. Gantz, D. Reinsel, The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east, IDC iView: IDC Analyze the future, 2007(2012) (2012) 1-16.

[2] R. Feldman, I. Dagan, Knowledge Discovery in Textual Databases (KDT), in: KDD, 1995, pp. 112-117.

[3] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E.D. Trippe, J.B. Gutierrez, K. Kochut, A brief survey of text mining: Classification, clustering and extraction techniques, arXiv preprint arXiv:1707.02919, (2017).

[4] H. Alrasheed, Word synonym relationships for text analysis: A graph-based approach, Plos one, 16(7) (2021) e0255127.

[5] G. Miner, Practical text mining and statistical analysis for non-structured text data applications, Academic Press, 2012.

[6] S. Beliga, A. Meštrović, S. Martinčić-Ipšić, An overview of graph-based keyword extraction methods and approaches, Journal of information and organizational sciences, 39(1) (2015) 1-20.
[7] S. Siddiqi, A. Sharan, Keyword and keyphrase extraction techniques: a literature review, International Journal of Computer Applications, 109(2) (2015).
[8] L. Page, S. Brin, R. Motwani, T. Winograd, The pagerank citation ranking: Bring order to the web, Technical report, stanford University, 1998.
[9] A. Gazni, Automatic extraction of key phrases from Persian article texts, Librarianship and Information Science, 9(3) (2006) 95-106. (in persian)
[10] O. Hajipoor, s. sadidpour, Automatic keyword extraction of short Persian texts using word2vec, Electronic and Cyber Defense, 8(2) (2020) 105-114. (in persian)
[11] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, A. Jatowt, YAKE! Keyword extraction from single documents using multiple local features, Information Sciences, 509 (2020) 257-289.
[12] B. Hejazi, J.A. Nasiri, Keywords Extraction from Persian Thesis Using Statistical Features and Bayesian Classification, Language Related Research, 12(6) (2022) 339-367.
[13] M. Azarafza, M.-R. Feizi-Derakhshi, M.B. Shendi, Textrank-based microblogs keyword extraction method for Persian language, in: Conference: 3rd International Congress on Science and Engineering, Hamburg, Germany, 2020.
[14] Z. Zhou, X. Zou, X. Lv, J. Hu, Research on weighted complex network based keywords extraction, in: Chinese Lexical Semantics: 14th Workshop, CLSW 2013, Zhengzhou, China, May 10-12, 2013. Revised Selected Papers 14, Springer, 2013, pp. 442-452.
[15] D.A. Vega-Oliveros, P.S. Gomes, E.E. Milios, L. Berton, A multi-centrality index for graph-based keyword extraction, Information Processing & Management, 56(6) (2019) 102063.
[16] S.K. Biswas, Keyword extraction from tweets using weighted graph, in: Cognitive Informatics and Soft Computing: Proceeding of CISC 2017, Springer, 2019, pp. 475-483.
[17] A. Bellaachia, M. Al-Dhelaan, Ne-rank: A novel graph-based keyphrase extraction in twitter, in: 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, IEEE, 2012, pp. 372-379.
[18] V. Kalra, R. Aggarwal, Importance of Text Data Preprocessing & Implementation in RapidMiner, ICITKM, 14 (2017) 71-75.
[19] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information processing & management, 24(5) (1988) 513-523.
[20] f. hafezi moghaddas, m. maghrebi, Using text mining techniques to analyze technical texts: A case study, content analysis of the American Concrete Code (ACI 318-08), in: 13th National Congress on Civil Engineering, 2022. (in persian)
[21] J. Sun, K. Lei, L. Cao, B. Zhong, Y. Wei, J. Li, Z. Yang, Text visualization for construction document information management, Automation in construction, 111 (2020) 103048.
[22] Z. Xie, Centrality measures in text mining: prediction of noun phrases that appear in abstracts, in: Proceedings of the ACL student research workshop, 2005, pp. 103-108.
[23] M. Bastian, S. Heymann, M. Jacomy, Gephi: an open source software for exploring and manipulating networks, in: Proceedings of the international AAAI conference on web and social media, 2009, pp. 361-362.

Automatic verification of correspondence between teaching resources and executive regulations in the field of design and implementation of concrete buildings: a text mining approach

Fatemeh Hafezi Moghaddas[a], Mojtaba Maghrebi[b1]

[a] Master's student, Faculty of Engineering, Department of Civil Engineering and Construction Management, Ferdowsi University, Mashhad, Iran.

[b] Associate Professor, Faculty of Engineering, Department of Civil Engineering and Construction Management, Ferdowsi University, Mashhad, Iran.

ABSTRACT

One of the challenges of higher education in editing university educational texts is to achieve the maximum compatibility of the content of educational resources with the instructions. therefore, to reach an efficient educational system and in line with industry needs, the appropriateness of the content of the educational resources with the regulations should be evaluated and revised if necessary. The need to review educational resources in the field of engineering and technology is important because these disciplines are needed in the application of industry and services in the country, and in fact, the training of expert and technical forces that can meet the technical needs of the country at different levels is the most important task of curricula in universities. This issue doubles the importance of paying attention to the teaching resources of these disciplines. this research seeks to extract keywords of "Iranian Concrete Regulations" and "reinforced concrete structures" which using three statistical approaches, linguistic knowledge and graph-based approaches proposes a hybrid method. then, the keywords of each document are visualized in a clustered network and analyzed. Comparing the results shows that the contents of these two documents are not completely similar. In fact, it can be said that the Regulations is an instructional document in which all the details have been addressed and all the issues surrounding concrete structures have been discussed. But the book is not satisfied with the design instructions and has examined the concepts related to the design of concrete structures in detail.

KEYWORDS

Text mining, keyword extraction, network analysis, centrality index, Iran's Concrete Regulations (ABA)

[1] mojtabamaghrebi@ferdowsi.um.ac.ir