Journal of New Achievements in Electrical, Computer and Technology
Vol. 2, No. 3, Jun-Sep 2022,(103-121)
WWW.JNAECT.IR

مجله دستاوردهای نوین در برق،کامپیوتر و فناوری
دوره ۲، شماره ۳، تابستان ۱۴۰۱ (۱۲۱-۱۰۳)

طراحی و تولید مجموعه دادگان اخبار فارسی IHU-PersianNewsDataSet- Javadzade-et-al دانشگاه جامع امام حسین (ع)

حسین حسینی[۱]، محمد قلعه‌نوئی[۲]، محمدمهدی مختاری[۳]، محمدعلی جوادزاده[۴]

[۱]دانشجوی کارشناسی ارشد دانشگاه جامع امام حسین (ع) hosseinhosseini@ihu.ac.ir
[۲]دانشجوی کارشناسی ارشد دانشگاه جامع امام حسین (ع) mghalenoei@ihu.ac.ir
[۳]دانشجوی کارشناسی ارشد دانشگاه جامع امام حسین (ع) m_m_mokhtari@ihu.ac.ir
[۴]استادیار دانشگاه جامع امام حسین (ع) javadzade@ihu.ac.ir

چکیده

اگر چه کمبود داده برای تحقیقات در حوزه پردازش زبان طبیعی یکی از چالش‌ها مهم است لیکن این چالش در خصوص زبان فارسی حادتر جلوه می‌کند، برای همین یافتن مجموعه دادگان باکیفیت و جامع در زبان فارسی کار دشواری است. علاوه بر آن دارا بودن برخی مشکلات از قبیل قابلیت دسته‌بندی و عدم رعایت استاندارد ذخیره‌سازی از نمونه مشکلات مجموعه دادگان موجود می‌باشد که هرکدام از این موارد می‌تواند بر میزان یادگیری مدل، نتایج و میزان خطا در آزمایش‌ها تأثیر بگذارد. به همین منظور تمامی این دلایل سبب شد که به دنبال جمع‌آوری و تهیه مجموعه دادگانی باشیم که تمام این‌گونه مشکلات را پوشش و میزان خطا هنگام به‌کارگیری داده‌ها در مدل‌های مختلف را کاهش دهد. ما در این پژوهش خزشگری را در جهت جمع‌آوری دادگان متنی طراحی و استفاده نموده‌ایم که با خزش بر روی یکی از پایگاه‌های خبری توانسته است مجموعه‌ای از دادگان را در پنج ستون عنوان، خلاصه، متن، برچسب و تاریخ انتشار خبر جمع‌آوری نماید. داده‌های متنی به کمک یکی از کتابخانه‌های مخصوص زبان فارسی در زبان برنامه‌نویسی پایتون، نرمال‌سازی شده و در دو فرمت csv و xml ذخیره‌سازی شده و در اختیار پژوهشگران همکار قرار گرفته است. برچسب‌ها در این مجموعه داده شامل ۱۳ برچسب اصلی ورزشی، هنر و رسانه، فرهنگ، علم و پیشرفت، سیاسی، سیاست خارجی، زندگی، خانواده، جامعه، تعلیم و تربیت، بین‌الملل، اقتصادی و استان‌ها می‌باشد. از جمله کارهایی که بر روی این مجموعه داده قابل انجام است می‌توان به دسته‌بندی متن، استخراج متن، خلاصه‌سازی متن و تشخیص عنوان اشاره کرد. همچنین از ویژگی‌های بارز این مجموعه داده می‌توان به جامعیت، تعداد داده‌های مناسب، وجود ویژگی‌های مفید، دارا بودن ویژگی‌های منحصربه‌فرد و همچنین ذخیره‌سازی در قالب استاندارد اشاره کرد. این مجموعه داده محصول گروه پردازش زبان دانشگاه جامع امام حسین (ع) می‌باشد و از طریق لینک مذکور در پانویس صفحه بعد و با رعایت حق کپی‌رایت قابل دریافت و استفاده می‌باشد.

واژه‌های کلیدی: مجموعه داده، اخبار فارسی، پردازش زبان طبیعی، مجموعه داده اخبار فارسی، یادگیری ماشین، دسته‌بندی متن، استخراج متن، خلاصه‌سازی متن، تشخیص عنوان

.

۱. مقدمه

در پژوهشی که بر روی دسته‌بندی متون خبری داشتیم سعی بر آن شد، با استفاده از مجموعه داده‌های استاندارد کار تحقیقاتی‌مان را پیش ببریم؛ اما با وجود تحقیقات بسیاری که در حوزه دسته‌بندی متون خبری در زبان فارسی انجام شده، متأسفانه مجموعه داده فارسی مناسبی در این حوزه یافت نشد. عدم وجود مجموعه داده مناسب بخصوص در حوزه متن فارسی بسیار پررنگ‌تر است به طوری که برای متن‌کاوی و داده‌کاوی در حوزه خبر، چندین مجموعه داده موجود می‌باشد مانند مجموعه داده همشهری، توسعه اندیشه نوین، هزار خبر فارسی و سایرین اما با وجود کارهای تحقیقاتی فراوان که در رابطه با دسته‌بندی متون خبری موجود است، هیچ یک از این تحقیقات با وجود زحمات بسیاری که برای تدوین مجموعه داده خود کشیده شده است، اقدام به نشر مجموعه داده تولیدی و استفاده شده در تحقیقاتشان نکرده‌اند و سایرین هر یک مجبور به تهیه دوباره مجموعه داده شده‌اند؛ که این عدم انتشار مجموعه داده‌ها باعث ایجاد دو مشکل می‌شود.

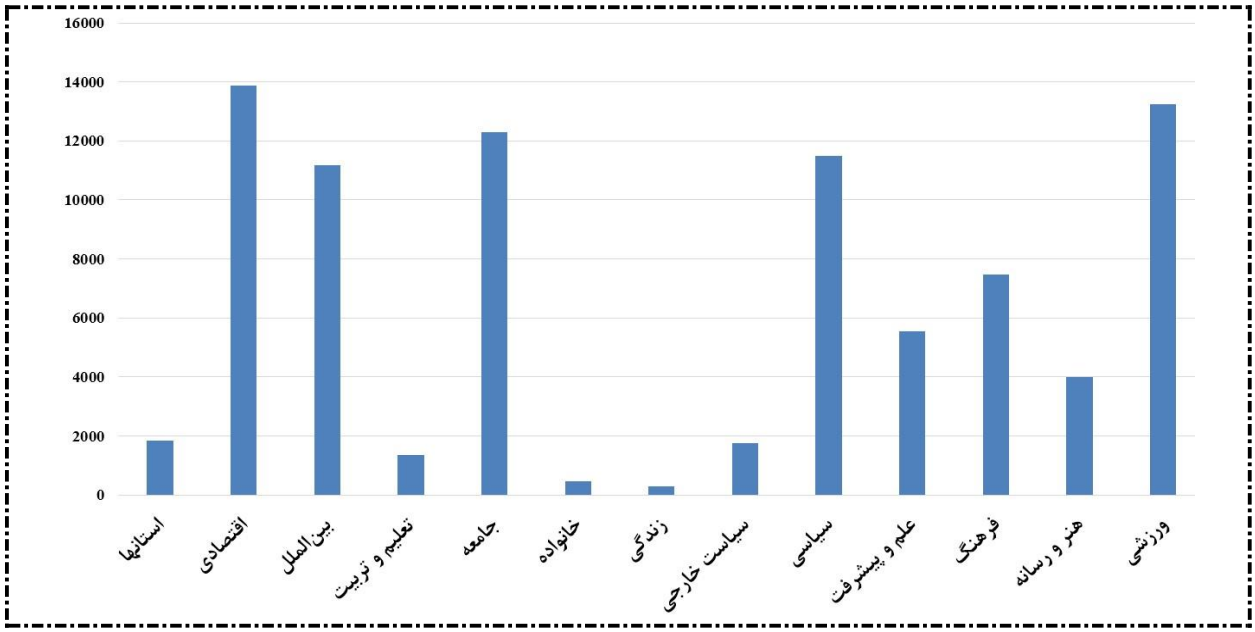
۱. اگر محققی بخواهد تحقیقات خود را در حوزه دسته‌بندی متون خبری انجام دهد، با توجه به عدم وجود مجموعه داده در این حوزه، می‌بایست خود اقدام به تهیه مجموعه داده نماید. باید در نظر داشت که این عمل پروسه‌ای وقت‌گیر و هزینه‌بر است.

۲. با توجه به عدم وجود مجموعه داده یکسان نمی‌توان پژوهش جدید را با دیگر پژوهش‌های انجام شده در این حوزه مقایسه نمود.

از این جهت در این پژوهش سعی شد، مجموعه داده استانداردی برای تحقیقات ایجاد شود، در نهایت نسبت به انتشار آن اقدام گردد که در ادامه توضیحات بیشتری نسبت به آن خواهیم داد.

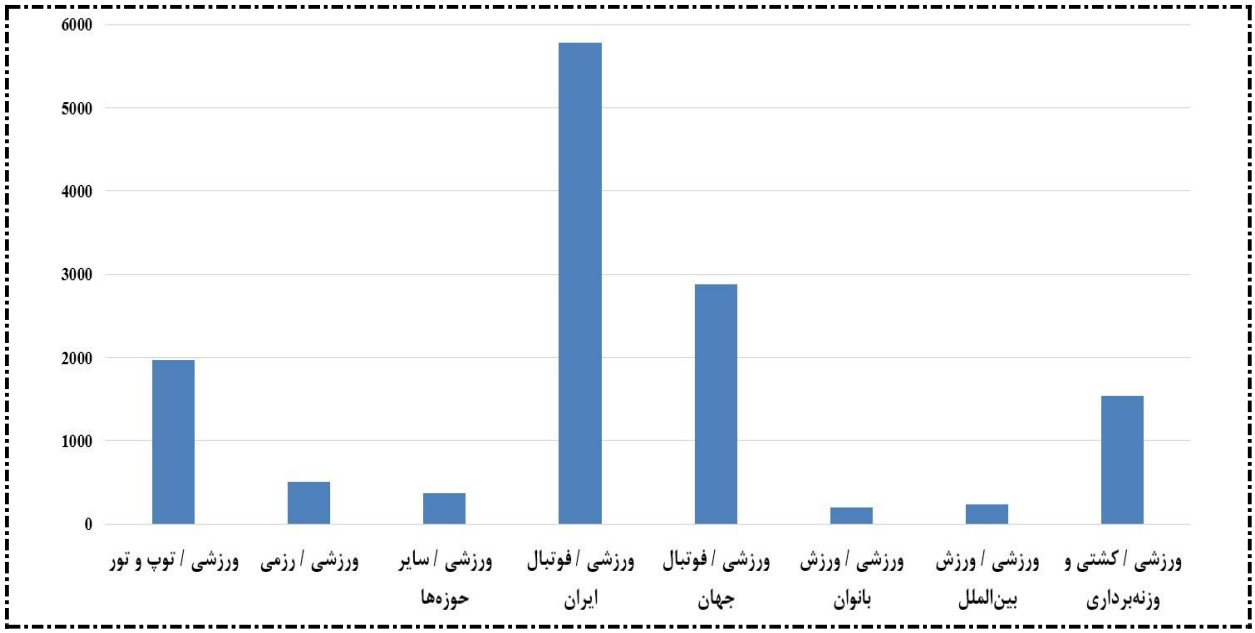
مجموعه دادگان اخبار فارسی (IHU-PersianNewsDataSet-Javadzade-et-al)[1] جمع‌آوری شده یک مجموعه دادگان ۸۶ هزارتایی است که دربرگیرنده پنج ویژگی عنوان، خلاصه، متن اصلی، برچسب موضوعی و تاریخ انتشار خبر می‌باشد. ویژگی‌های در نظر گرفته شده و تعداد داده‌ها در این مجموعه دادگان باعث جامعیت و کاربردی بودن این مجموعه گشته است. در ویژگی برچسب موضوعی دو نوع برچسب تعبیه شده است؛ (۱) برچسب‌های اصلی که شامل ۱۳ برچسب است و (۲) برچسب‌های فرعی که زیر برچسب‌های هرکدام از برچسب‌های اصلی در نظر گرفته می‌باشد. با توجه به شکل ۱ درمی‌یابیم که بیشترین فراوانی داده برچسب‌های اصلی مربوط به برچسب‌های سیاسی، جامعه و اقتصادی است و همچنین کمترین فراوانی نیز مربوط به برچسب‌های زندگی و خانواده می‌باشد.

[1] https://github.com/IHU-PersianNewsDataSet-Javadzade-et-al/dataset



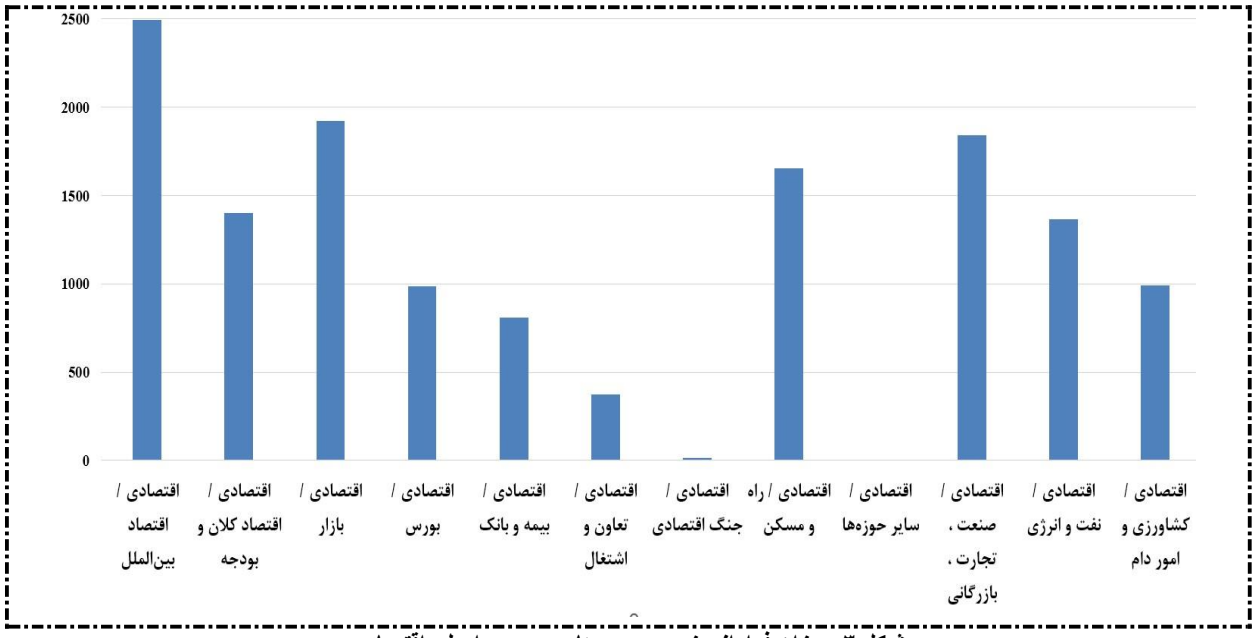
شکل ۱. میزان فراوانی برچسب‌های اصلی

در ادامه نیز به مشاهده فراوانی داده زیر برچسب‌های برچسب‌های اصلی می‌پردازیم. در شکل ۲ میزان فراوانی زیر برچسب‌های ورزشی قابل مشاهده است که شامل ۸ زیر برچسب است و بیشترین فراوانی داده در زیر برچسب‌های فوتبال ایران و فوتبال جهان وجود دارد و کمترین آن مربوط به زیر برچسب ورزش بانوان است.



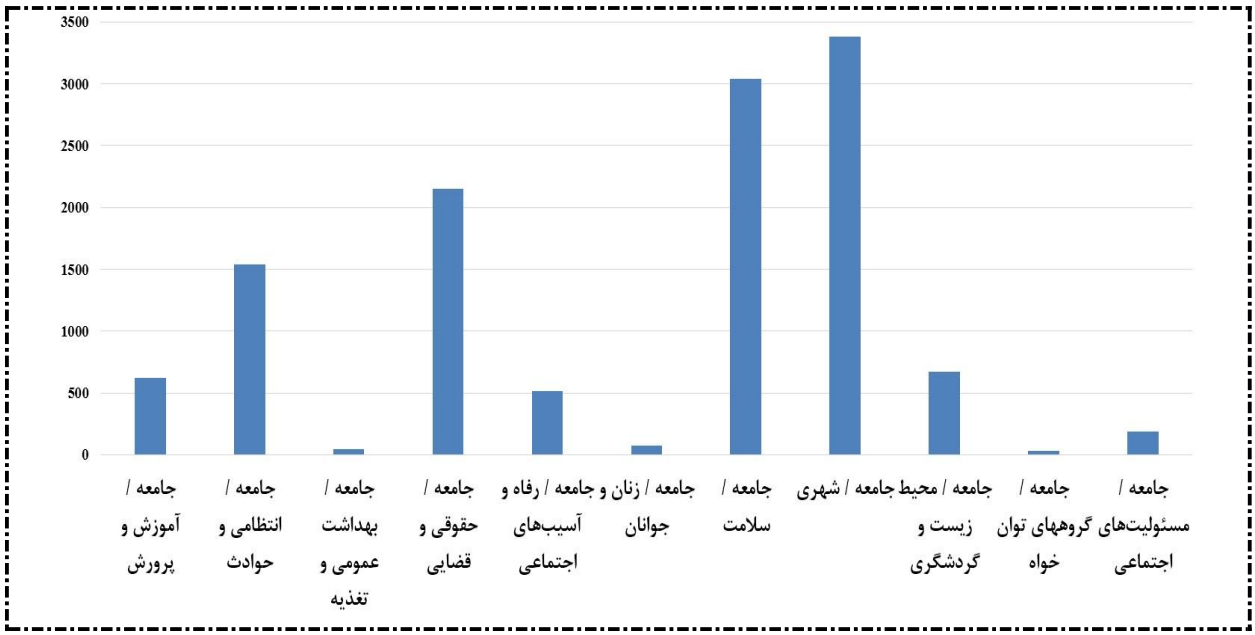
شکل ۲. میزان فراوانی زیر برچسب‌های برچسب اصلی ورزشی

در شکل ۳ فراوانی زیر برچسب‌های برچسب اصلی اقتصادی مورد بررسی قرار گرفته است که دارای ۱۱ زیر برچسب می‌باشد و بیشترین فراوانی مربوط به زیر برچسب‌های اقتصاد بین‌الملل، صنعت و تجارت و بازرگانی، بازار و راه و مسکن است و کمترین آن نیز مربوط به جنگ اقتصادی و سایر حوزه‌ها می‌باشد.



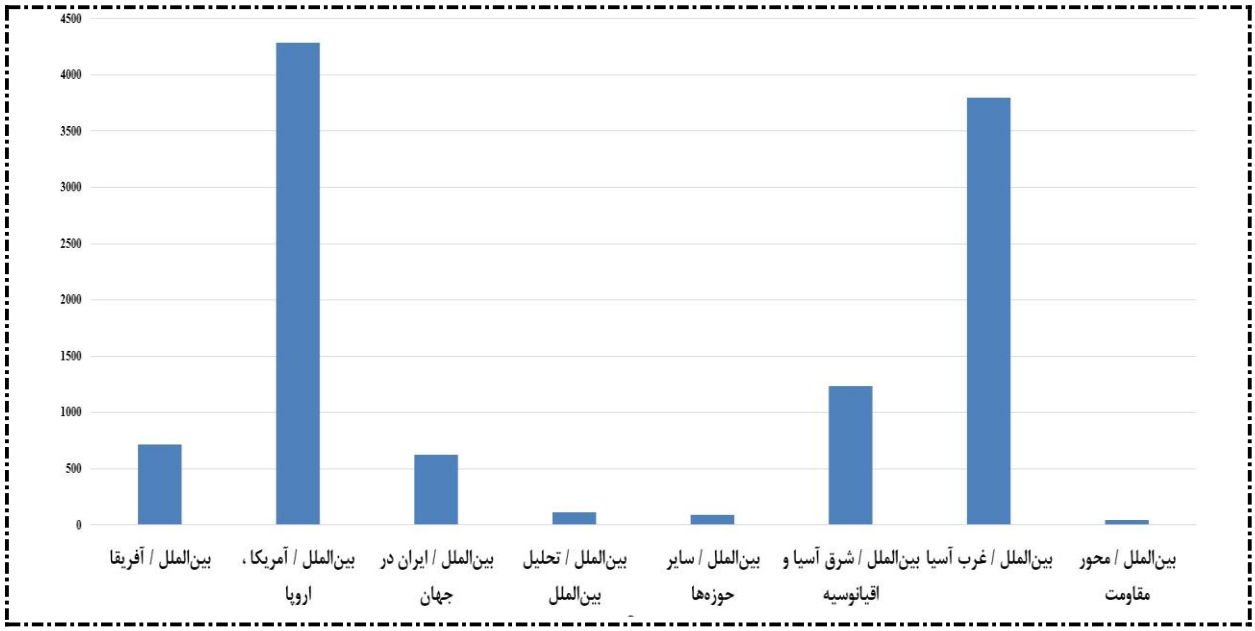
شکل ۳. میزان فراواني زير برچسب‌های برچسب اصلي اقتصادي

در شکل ۴ فراوانی زیر برچسب‌های برچسب اصلی جامعه مورد بررسی قرار گرفته است که دارای ۱۱ زیر برچسب است و بیشترین فراوانی مربوط به زیر برچسب‌های شهری و سلامت و کمترین فراوانی متعلق به زیر برچسب‌های زنان و جوانان، بهداشت عمومی و تغذیه و گروه‌های توان‌خواه می‌باشد.



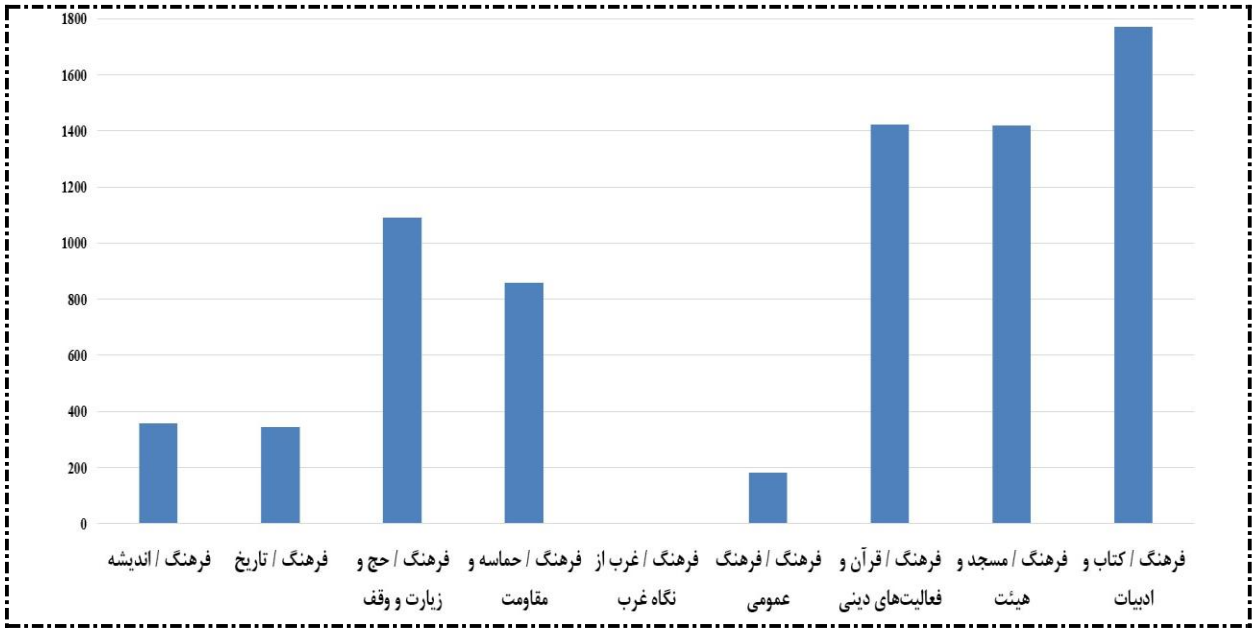
شکل ۴. میزان فراوانی زیر برچسب‌های برچسب اصلی جامعه

در شکل ۵ فراوانی زیر برچسب‌های برچسب اصلی بین‌الملل مورد بررسی قرار گرفته است که دارای ۸ زیر برچسب است و بیشترین فراوانی مربوط به زیر برچسب‌های آمریکا، اروپا و غرب آسیا است که اختلاف زیادی در فراوانی با سایر زیر برچسب-های بین‌الملل دارند و کمترین آن متعلق به زیر برچسب محور مقاومت است.



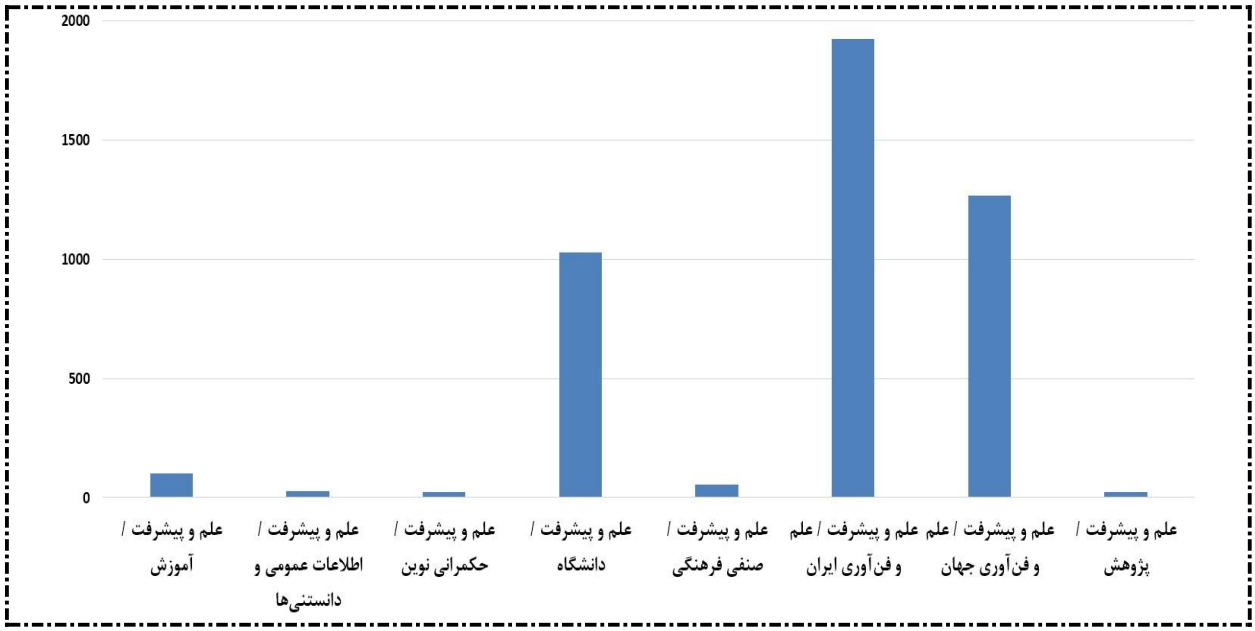
شکل ۵. میزان فراوانی زیر برچسب‌های برچسب اصلی بین‌الملل

در شکل ۶ فراوانی زیر برچسب‌های برچسب اصلی فرهنگ مورد بررسی قرار گرفته است که دارای ۹ زیر برچسب می‌باشد و بیشترین فراوانی مربوط به سه زیر برچسب کتاب و ادبیات، مسجد و هیئت و قرآن و فعالیت‌های دینی می‌باشد و کمترین آن نیز مربوط به غرب از نگاه غرب است



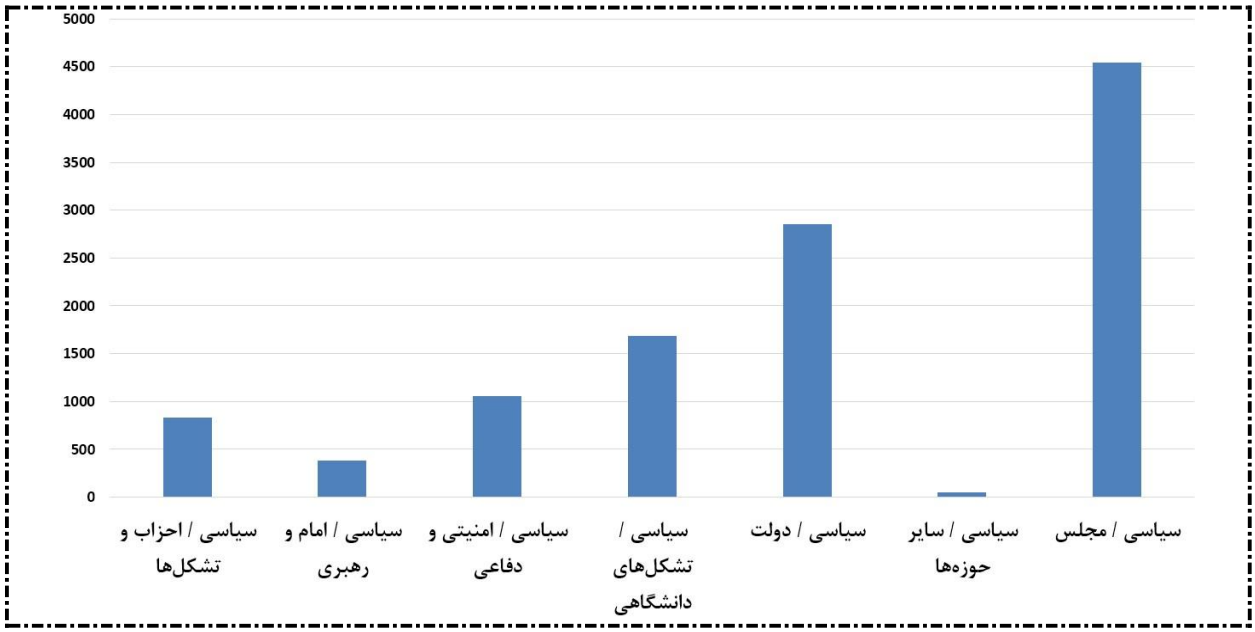
شکل ۶. میزان فراوانی زیر برچسب‌های برچسب اصلی فرهنگ

در شکل ۷ فراوانی زیر برچسب‌های برچسب اصلی علم و پیشرفت مورد بررسی قرار گرفته است که دارای ۸ زیر برچسب است و بیشترین فراوانی متعلق به سه زیر برچسب علم و فن‌آوری ایران، علم و فن‌آوری جهان و دانشگاه می‌باشد و کمترین آن نیز مربوط به زیر برچسب‌های پژوهش و حکمرانی نوین می‌باشد.



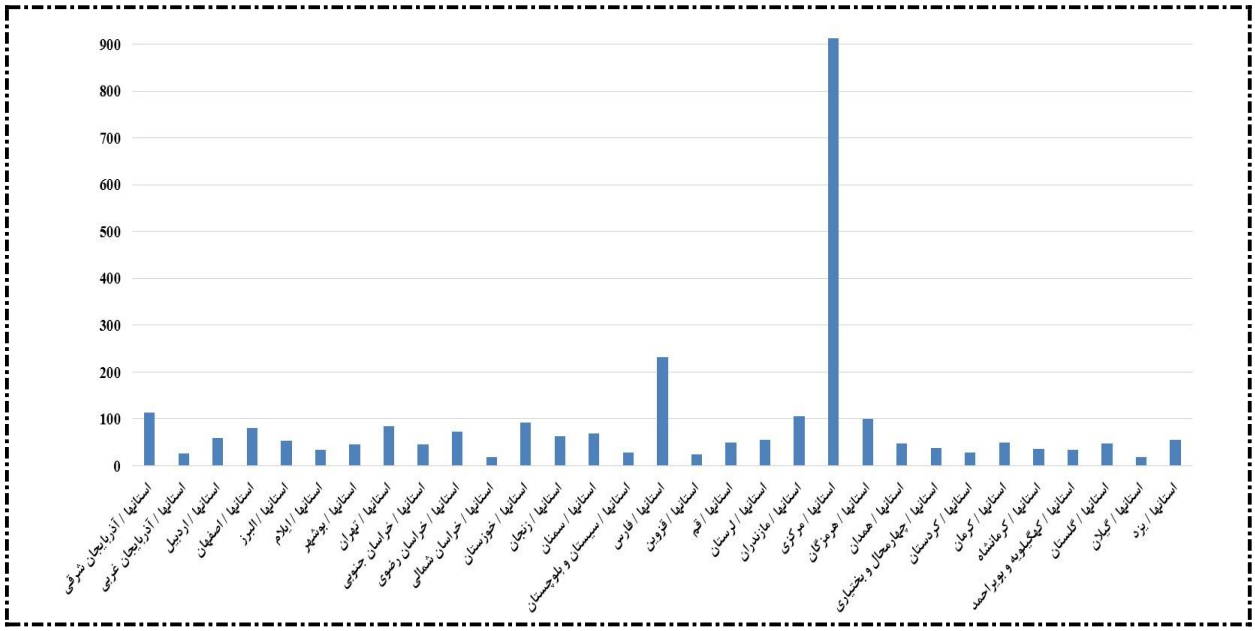
شکل ۷. میزان فراوانی زیر برچسب‌های برچسب اصلی علم و پیشرفت

در شکل ۸ فراوانی زیر برچسب‌های برچسب اصلی سیاسی مورد بررسی گرفته است که دارای ۷ زیر برچسب است و بیشترین فراوانی مربوط به زیر برچسب مجلس می‌باشد و کمترین آن نیز مربوط به سایر حوزه‌ها می‌باشد.



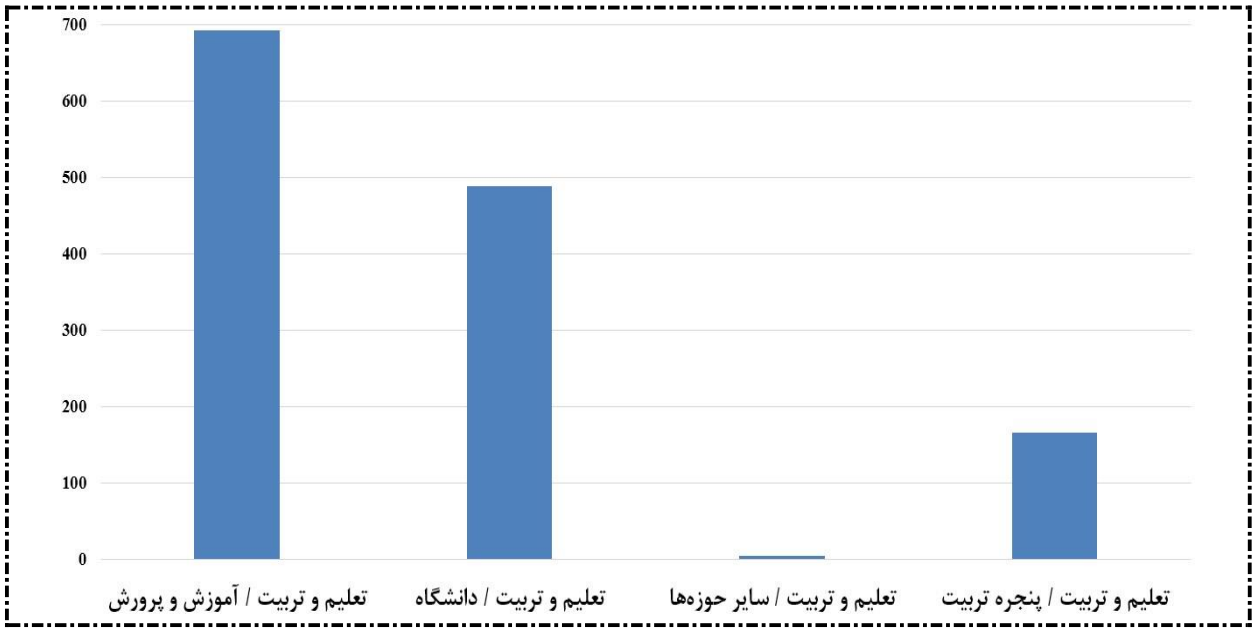
شکل ۸. میزان فراوانی زیر برچسب‌های برچسب اصلی سیاسی

در شکل ۹ فراوانی زیر برچسب‌های برچسب اصلی استان‌ها مورد بررسی قرار گرفته است که دارای ۳۱ زیر برچسب است و بیشترین فراوانی با اختلاف به زیر برچسب مرکزی تعلق دارد و کمترین آن نیز مربوط به گیلان و خراسان شمالی می‌باشد.



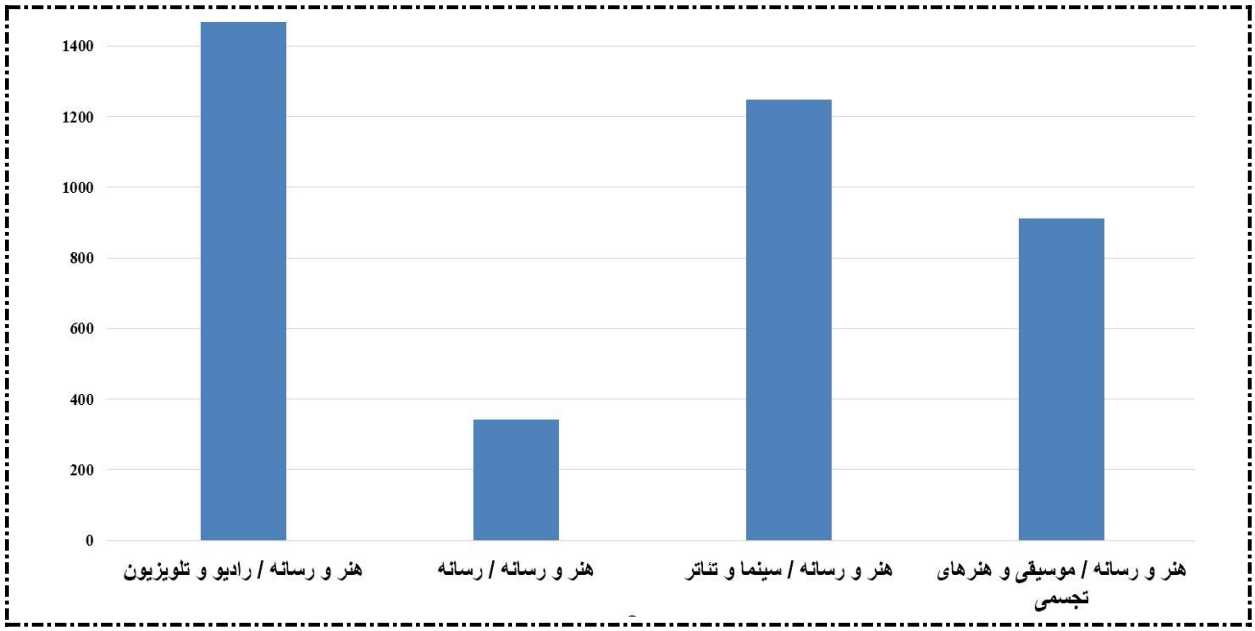
شکل ۹. میزان فراوانی زیر برچسب‌های برچسب اصلی استان‌ها

در شکل ۱۰ فراوانی زیر برچسب‌های برچسب اصلی تعلیم و تربیت مورد بررسی قرار گرفته است که دارای ۴ زیر برچسب می‌باشد و بیشترین فراوانی مربوط به زیر برچسب آموزش و پرورش است و کمترین فراوانی مربوط به سایر حوزه‌ها می‌باشد.



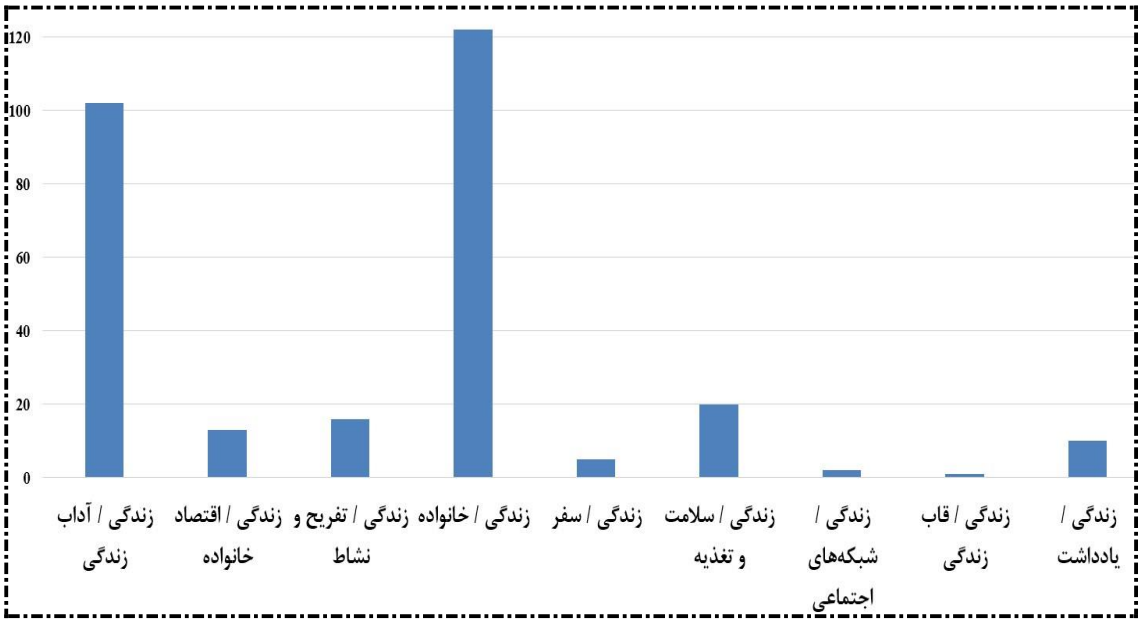
شکل 10. میزان فراوانی زیر برچسب‌های برچسب اصلی تعلیم و تربیت

در شکل ۱۱ فراوانی زیر برچسب‌های برچسب اصلی هنر و رسانه مورد بررسی قرار گرفته است که دارای ۴ زیر برچسب است و بیشترین فراوانی متعلق به زیر برچسب رادیو و تلویزیون است و کمترین آن نیز مربوط به زیر برچسب رسانه می‌باشد.



شکل 11. میزان فراوانی زیر برچسب‌های برچسب اصلی هنر و رسانه

در شکل ۱۲ فراوانی زیر برچسب‌های برچسب اصلی زندگی مورد بررسی قرار گرفته است که دارای ۹ زیر برچسب است و بیشترین فراوانی‌ها متعلق به زیر برچسب خانواده و آداب زندگی است و کمترین آن متعلق به زیر برچسب قاب زندگی می‌باشد.



شکل ۱۲. میزان فراوانی زیر برچسب‌های برچسب اصلی زندگی

در انتها باید گفت از آنجایی که تعداد برچسب‌های اصلی ما ۱۳ تا می‌باشد اما در اینجا به ۱۱ نمودار فراوانی زیر برچسب‌های برچسب اصلی پرداخته شد و دلیل آن این است که نمودار فراوانی برچسب‌های اصلی سیاست خارجی و خانواده به خودی خود یک برچسب اصلی واحد هستند و زیر برچسبی را شامل نمی‌شوند. از این رو از ترسیم نمودار فراوانی آن‌ها صرف نظر کرده‌ایم.

در ادامه، بخش دوم نسبت به معرفی چند مجموعه داده که تحقیقات بیشتری با آن‌ها در حوزه دسته‌بندی متون خبری انجام شده و به نوعی معروفیت بیشتری داشتند معرفی گردید تا بررسی و مشاهده شوند، در بخش سوم ادبیات موضوعی مربوطه ذکر

گردید، در بخش چهارم روش جمع‌آوری مجموعه داده ارائه شده معرفی گردید، در بخش پنجم ارزیابی بین مجموعه داده تولیدی و مقایسه آن با سایرین انجام شد و در نهایت به جمع‌بندی و نتیجه‌گیری پرداخته‌ایم.

۲. کارهای مرتبط

در این بخش به بررسی چند نمونه مجموعه دادگان فارسی که پیش از این منتشر شده‌اند می‌پردازیم و آن‌ها را به تفکیک تحلیل می‌نماییم.

اولین مورد مربوط به یک مجموعه دادگان با ۱۷۵ هزار داده است که از پنج خبرگزاری جمع‌آوری شده است. این مجموعه داده شامل ویژگی‌های عنوان، توضیحات، عکس، سایت خبرگزاری، آدرس (لینک) خبر و برچسب خبر است. ویژگی برچسب خبر نیز تنها دارای ۹ برچسب است [۱]. این مجموعه داده به صورت پایگاه داده‌ای در سه فایل جداگانه ذخیره شده است. فایل اول مربوط به ۹ برچسب خبری و برچسب عددی (id) آن‌ها است. فایل دوم سایر ویژگی‌ها را به همراه برچسب عددی خبر در خود جا داده است. در آخر نیز فایل سوم که جدول رابط بین دو فایل دیگر است که شامل برچسب عددی خبرها و برچسب عددی مربوط به برچسب خبری می‌باشد. این روش ذخیره‌سازی دسترسی کاربر را به داده‌های هر فایل از طریق فایل دیگر زمان‌بر می‌سازد. از آنجایی که حالت استاندارد ذخیره‌سازی به این‌گونه نیست و دسترسی به داده‌های مختلف را مشکل می‌سازد، این روش ذخیره‌سازی ایراد به حساب می‌آید. از سوی دیگر نیز ویژگی‌های غیرمفید مانند عکس، سایت خبرگزاری و آدرس خبر در پژوهش‌ها و تحقیقات پردازش زبان و یادگیری ماشین کاربردی ندارند و تنها کاربردشان در بررسی صحت یک خبر قابل استفاده می‌باشد و همچنین این ویژگی‌های غیرمفید، ایراد دوم مجموعه معرفی شده تلقی می‌گردد.

مجموعه دادگان بعدی، هزار داده جمع‌آوری شده در ۸ ویژگی است. کلمات کلیدی، عنوان، جمله اول و متن کامل خبر به همراه تعداد کلمات هرکدام از آن‌ها، ویژگی‌های این مجموعه دادگان را تشکیل می‌دهد [۲]. این مجموعه دادگان، تعداد کمی داده را در خود ذخیره دارد و جهت استفاده در پژوهش‌ها مناسب نیست. ویژگی‌های تعداد کلمات برای هر یک از ویژگی‌های اصلی نیز کاربرد خاصی ندارند و از سوی دیگر در صورت نیاز در حین پروژه با استفاده از تعداد بسیار کمی خط کد برنامه‌نویسی و یا با بهره‌گیری از ابزارهای متناسب به دست می‌آیند که این نشان از غیرمفید بودن این ویژگی‌ها می‌باشد. همچنین در این مجموعه ویژگی منحصربه‌فردی جهت انجام تشخیص و قابلیت دسته‌بندی مانند دسته‌بندی موضوعی وجود ندارد. موارد فوق‌الذکر از جمله مشکل‌ها و ایرادات مجموعه داده مورد بررسی قرار گرفته می‌باشد.

مجموعه دادگان سوم توسط گروه توسعه اندیشه نوین در ۱۰۰ هزار داده از یک خبرگزاری جمع‌آوری شده است. تعداد ویژگی‌های آن به دو ویژگی عنوان و متن خبر محدود شده است [۳]. این مجموعه تعداد دادگان بسزایی دارد اما ایرادات مهمی بر آن وارد است از جمله تعداد ویژگی کم و عدم ویژگی‌های منحصربه‌فرد و قابل دسته‌بندی مهم‌ترین مواردی هستند که این مجموعه دادگان را ناکارآمد و غیر جامع می‌سازد.

آخرین مجموعه داده مورد بررسی نیز، مجموعه دادگان ۱۶۶ هزار داده‌ای همشهری است. این مجموعه داده شامل اخبار بین سال‌های ۱۳۷۵ تا ۱۳۸۷ می‌باشد [۴]. این مجموعه داده از لحاظ تعداد داده و وجود ویژگی منحصر پوشش داده شده است؛ اما عدم وجود تعداد ویژگی بیشتر از کاربرد و جامع بودن آن می‌کاهد و این مجموعه را دچار ایراد می‌نماید.

۳. ادبیات موضوعی

در این بخش ما به شرح کلماتی که در این تحقیق به آن‌ها ذکر شده است و لازم است که مفهوم آن‌ها روشن گردد، پرداخته‌ایم.

کرولر: به دریافت خودکار اطلاعات از وب و بررسی و گزینش کردن بخش‌های مهم آن کراول کردن گفته می‌شود که این کار در موتورهای جستجو، جمع‌آوری اطلاعات و غیره کاربرد دارد. برای ساخت کراول روش‌های متفاوتی موجود هست و زبان‌های برنامه‌نویسی مختلف روش‌های مختلفی را برای آن ارائه کردند. همچنین کرولرها به دو دسته خطی و تودرتو تقسیم می‌شوند که در مدل خطی یک یا لیستی از آدرس‌ها برای تحلیل ارائه می‌گردد و برنامه کراولر باید یک‌به‌یک آن‌ها رو دریافت کرده و سپس تحلیل کند؛ اما در کراولر تودرتو هنگام باز کردن یک صفحه وب، لیستی از لینک‌های درون آن صفحه را گرفته و یک‌به‌یک آن‌ها را نیز بررسی می‌کند [۵].

URL: کلمه اختصاری URL که مخفف Uniform Resource Locator است، راهی است برای شناسایی مکان یا موضع یک فایل بر روی اینترنت. URLهمان آدرسی است که ما نه تنها برای باز کردن وب‌سایت‌ها برای مرور، بلکه برای دانلود تصاویر، ویدیو، برنامه‌های نرم‌افزاری و انواع مختلف دیگر فایل‌های استفاده می‌کنیم که بر روی یک سرور میزبانی می‌شوند [۶].

ریدایرکت: ریدایرکت به معنای انتقال کاربر به آدرس جدید است؛ به عبارت دیگر هر زمان بخواهیم بازدیدکنندگان سایت را به آدرس مشخصی منتقل کنیم از Redirect استفاده می‌کنیم. مدیران سایت به‌خصوص در ارتباط با «رعایت اصول سئو» همواره با ریدایرکت آدرس صفحات مختلف سایت سر و کار دارند. یکی از موارد رایج استفاده از ریدایرکت زمانی است که آدرس تمام یا برخی صفحات سایت تغییر یافته و ما می‌خواهیم آن‌ها را به آدرس‌های جدید هدایت کنیم و یا برای یک صفحه از سایت، چندین آدرس متفاوت وجود داشته باشد و به دلیل مشکل محتوای چندگانه یا Duplicate Content در سئو، مجبوریم آدرس‌های اضافی را بر روی آدرس اصلی ریدایرکت کنیم [۷].

کتابخانه beautifulsoup: یک کتابخانه پایتون است که به‌منظور استخراج داده از فایل‌های html و xml مورد استفاده قرار می‌گیرد. این کتابخانه صفحات مورد نظر خود را به‌صورت یک درخت تجزیه می‌کند. درخت تجزیه این امکان را برای برنامه ایجاد می‌کند که هرگونه دسترسی به عناصر صفحه html با سرعت بیشتری امکان‌پذیر گردد. با این روش شرایط مناسبی برای جستجوی اطلاعات مورد نظر فراهم می‌شود [۸].

۴. روش پیشنهادی

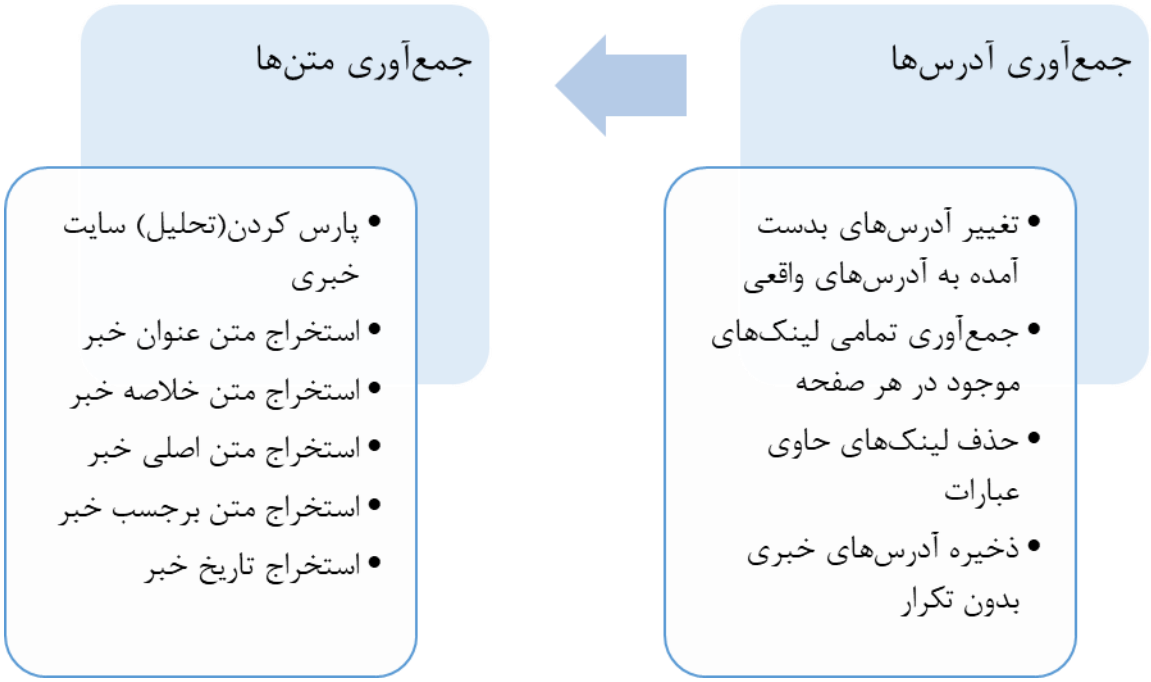
همان طور که در شکل ۱۳ روند کلی کار را مشاهده می‌کنید گفتنی است که اولین قدمی که ما برای طراحی خزنده وب مورد استفاده شده طی کردیم، مشخص کردن مسیرهایی است که خزنده ما باید پیمایش نماید. آدرس‌های اولیه ما شامل ۳ بخش، (۱) آدرس پایه که همان آدرس خبرگزاری است، (۲) عنوان دسته‌بندی و زیر دسته‌بندی خبر که همان برچسب‌ها و زیر برچسب‌های ما را شامل شده است و بخش آخر (۳) شماره صفحه می‌باشد. برای نمونه اگر بخواهیم در دسته سیاسی و زیر دسته مجلس (politics/parliament) و همچنین صفحه سوم این زیر دسته پیمایش انجام دهیم، آدرس به صورت https://www.sample.ir/politics/parliament?p=3 خواهد بود. در جمع‌آوری داده‌ها ما به ازای تمام دسته‌های خبری موجود در سایت صفحات ۰ تا ۲۵۰ زیر دسته‌های آن‌ها را پیمایش کردیم که در این پیمایش ما به دنبال جمع‌آوری آدرس خبرهای متنی هستیم. در نظر داشته باشید که در محتوای یک سایت ممکن است آدرس‌های خبری تکراری، آدرس خبرهای غیر متنی، آدرس صفحات تبلیغاتی و آدرس‌های غیر مرتبط با نیاز ما وجود داشته باشد.

در هر صفحه، زمانی که با هدف جست‌وجو جمع‌آوری آدرس‌ها، ساختار صفحه را بررسی کردیم، دریافتیم که باید به دنبال تگ‌های (tag) «a» باشیم که عبارت href در داخل تگ آن‌ها استفاده شده است. برای نمونه؛

```
<a class="d-flex flex-column h-100 justify-content-between"
href="/news/14010219000367/جمعی-از-کارگران-سراسر-کشور-با-رهبر-انقلاب-دیدار-کردند"
target="_blank">
```

در روند طراحی خزنده به زبان برنامه‌نویسی پایتون، نکته بیان شده را این‌گونه پیاده‌سازی کرده‌ایم؛

```
for element in page.find_all ('a'):
    link1 = element.get ('href')
```



شکل ۱۳. روند کلی جمع‌آوری مجموعه داده

برای حل مشکل تکراری نبودن آدرس‌ها، از یک ویژگی کاربردی زبان برنامه‌نویسی پایتون برای انواع متغیرها، به نام ست (set) استفاده کردیم. این نوع از متغیرها عضو تکراری نمی‌پذیرد. برای حل مشکل آدرس خبرهای غیر متنی، آدرس صفحات تبلیغاتی و آدرس‌های غیر مرتبط با نیاز ما، در طول روند آزمون‌وخطا برنامه، به این نتیجه رسیده بودیم، آدرس‌هایی که حاوی عباراتی مانند /، /my، /login، /news، /media و /media باشند، از نوع خبر متنی نیستند.

برای همین ما به کمک کتابخانه ریجکس (regex) در پایتون، با کمک دستورات شرطی، آدرس‌هایی که حاوی این عبارات بودند، به لیست آدرس‌های ذخیره شده اضافه نکردیم.

```
if re.match(r"^/my", link1 ,re.IGNORECASE):
    continue
if re.match(r"^/login", link1 ,re.IGNORECASE):
    continue
if re.match(r"^/media", link1 ,re.IGNORECASE):
    continue
if re.match(r"^/photo", link1 ,re.IGNORECASE):
    continue
```

یکی دیگر از چالش‌هایی که در انجام کار خود با آن مواجه شدیم قرار گرفتن بعضی از اخبار در دسته خبرهای متنی بود در صورتی که خبر مورد نظر متنی نبوده است و دلیل آن بدین خاطر بود که همه اخبار جزئی از دسته news هستند اما ممکن است که فیلم یا عکس باشند و در دسته media قرار بگیرند که همان‌طور که مشخص است این قبیل از داده‌ها به کار ما نمی‌آیند زیرا ما به دنبال داده‌های متنی هستیم. بدین منظور از تابع رکوئست (requests) پایتون و تابع هد (head) با مقدار مثبت برای پارامتر مربوطه (allow_redirect) آدرس واقعی را بدست آورده و دستورات شرطی را روی این آدرس‌ها پیاده‌سازی کردیم. برای نمونه آدرس اولیه‌ای داریم؛

https://www.farsnews.ir/**news**/14010321000239/فیلم|-سلام-فرمانده-از-زبان-کوکان-سرطانی

آدرس نهایی بدست آمده به صورت زیر خواهد بود؛

https://www.farsnews.ir/**media**/14010321000239/فیلم|-سلام-فرمانده-از-زبان-کوکان-سرطانی

که همان‌طور که مشاهده می‌شود این یک آدرس خبر غیر متنی می‌باشد و به کار ما نمی‌آید و سبب ایجاد اختلال در روند جمع‌آوری داده‌ها می‌گردد که با انجام عمل ریدایرکت سبب می‌شود که ما این دسته از لینک‌ها را شناسایی کرده و از آن‌ها عبور کنیم.

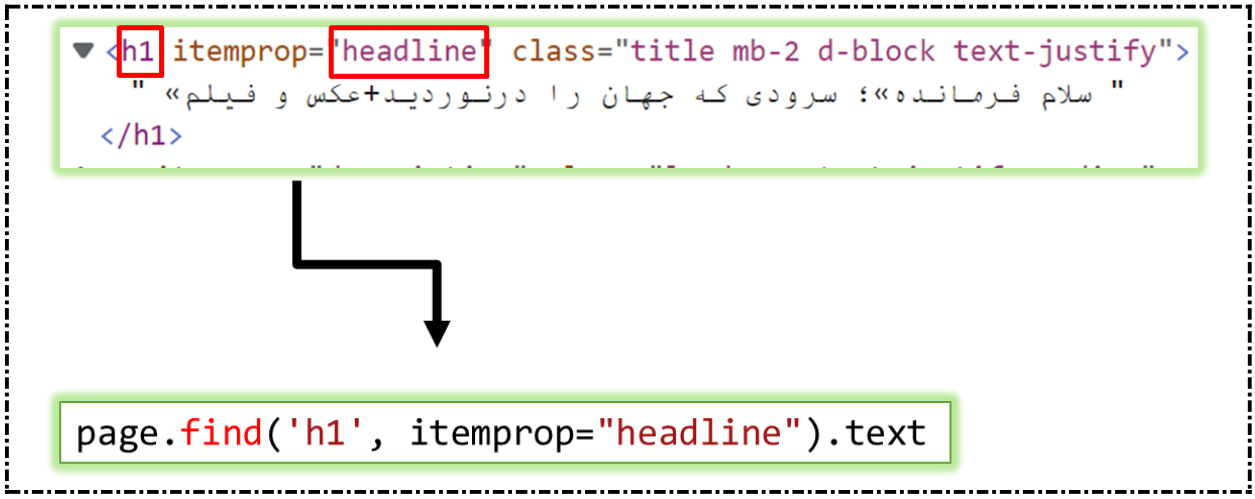
در روند پیمایش ۲۵۰ صفحه از هر زیر دسته خبری، یکی دیگر از چالش‌هایی که به آن برخورد کردیم تعداد درخواست بالا (request) به سایت خبری بود که به دلایل تدابیر امنیتی سایت خبری چیده بود باعث گردید که در روند پیمایش این ۲۵۰ صفحه تغییراتی ایجاد کنیم. این تغییرات به این صورت بود که به جای اینکه در ابتدا تمامی ۲۵۰ صفحه یک زیر دسته چک شود، در هر مرحله ۵ صفحه از هر زیر دسته خبری پیمایش می‌شود که به طور کلی طی ۵۰ مرحله، ۲۵۰ صفحه از هر زیر دسته خبری را پیمایش می‌کند. البته باید متذکر شد که تعداد صفحات بعضی از این زیر دسته‌های خبری، ممکن است از ۲۵۰ صفحه کمتر باشد که به ناچار این صفحات پیمایش می‌شود، اما آدرس جدیدی برای ما پیدا نخواهد شد. الگوریتم استفاده شده برای این بخش در زبان برنامه‌نویسی پایتون در ادامه بیان شده است.

```
range_list = []
```

```
for j in range(50):
    range_list.append(range (5*j , 5*j+5))
[range_list: [ range(5) , range(5,10) , … , range(245,250) ]
```

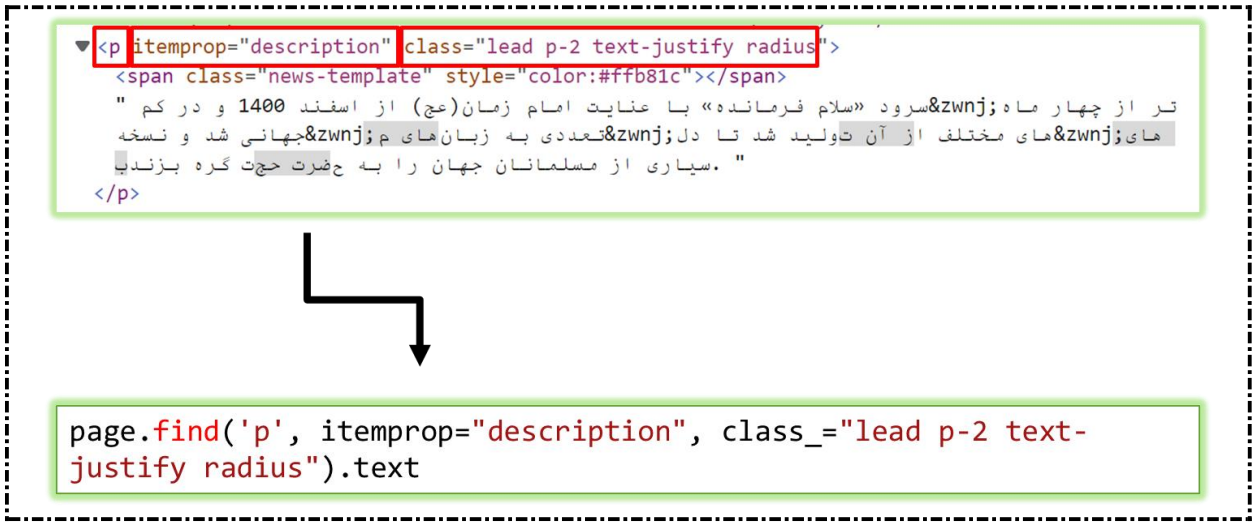
پس از جمع‌آوری تمامی آدرس‌های خبر متنی، باید عنوان، خلاصه، متن اصلی، برچسب و تاریخ خبر مربوط به هر آدرس جمع‌آوری شده را استخراج کنیم. برای این کار ابتدا لازم است قالب صفحه خبری متنی تحلیل (parse) شود که برای این کار از کتابخانه بیوتیفول‌سوپ (beautifulsoup) در پایتون استفاده کردیم تا قالب سایت خبری، برای خزنده ما قابل تحلیل باشد اما قبل از انجام این کار، با در نظر داشتن این موضوع که سایت خبری معتبر بوده و برای رعایت نکات لازم برای سئوی (seo) سایت خود، لینک‌های ذخیره‌شده‌ی ما که طی ۲ هفته، جمع‌آوری شد، نباید از دسترس خارج شده باشند، اما جهت اطمینان ما هرکدام از آدرس‌ها را قبل از تحلیل، بررسی می‌کردیم که اگر کد وضعیت آدرس برابر ۲۰۰ بود یعنی اینکه صفحه در دسترس بود، ادامه روند کارمان طی بشود. در ادامه به صورت تفکیک شده به ساختار تک‌تک بخش‌های مورد استفاده صفحه و تکه کدهایی که برای جمع‌آوری هر یک از ویژگی‌های عنوان، خلاصه، متن اصلی، برچسب و تاریخ خبر مربوط به هر آدرس مورد استفاده قرار می‌گیرد، اشاره خواهیم کرد.

در شکل ۱۴ ساختار بخش عنوان خبر صفحه و تکه کد مربوطه جهت جمع‌آوری این بخش را مشاهده می‌کنید؛



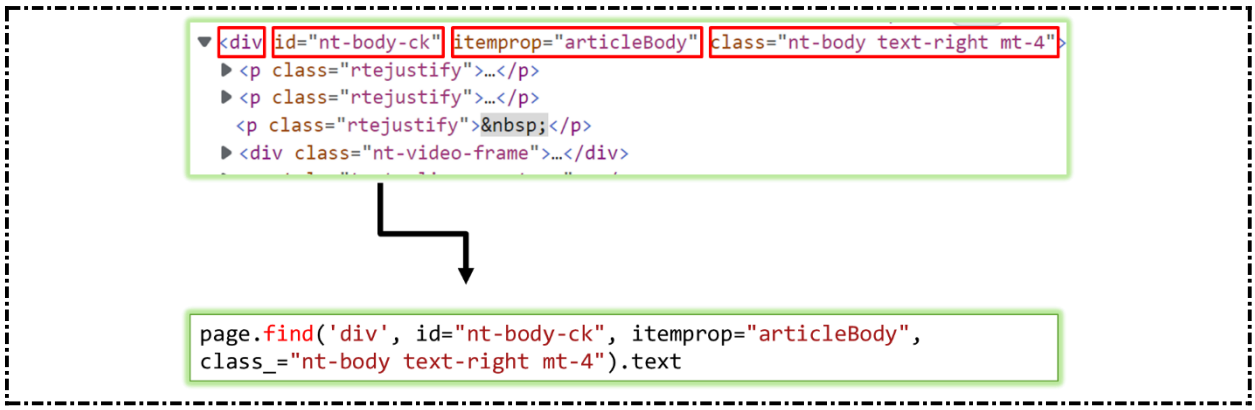
شکل ۱۴. عنوان خبر

در شکل ۱۵ ساختار بخش خلاصه خبر صفحه و تکه کد مربوطه جهت جمع‌آوری این بخش را مشاهده می‌کنید؛



شکل ۱۵. خلاصه خبر

در شکل ۱۶ ساختار بخش متن اصلی خبر صفحه و تکه کد مربوطه جهت جمع‌آوری این بخش را مشاهده می‌کنید؛

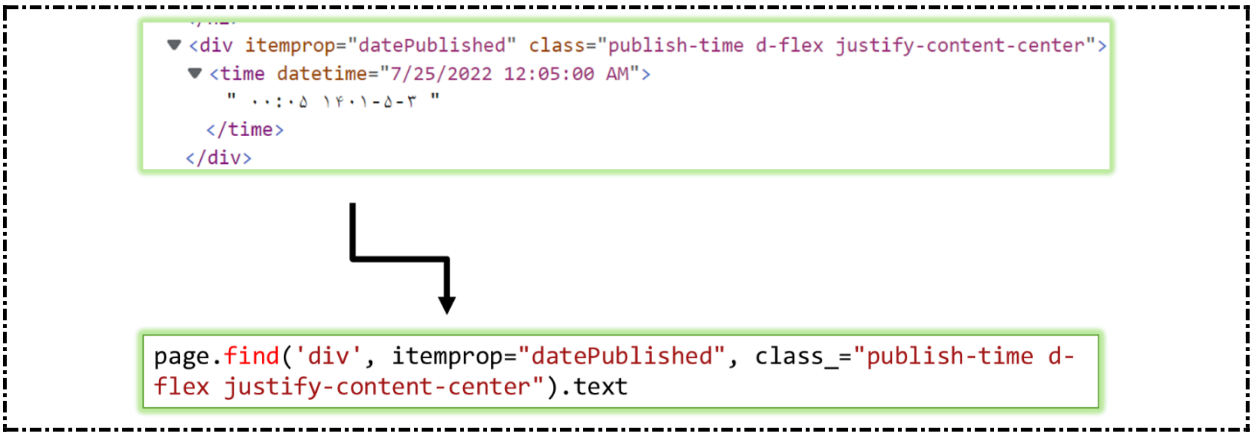


شکل ۱۶. متن اصلی خبر

در شکل ۱۷ ساختار بخش برچسب خبر صفحه و تکه کد مربوطه جهت جمع‌آوری این بخش را مشاهده می‌کنید؛

شکل ۱۷. برچسب خبر

در شکل ۱۸ ساختار بخش تاریخ خبر صفحه و تکه کد مربوطه جهت جمع‌آوری این بخش را مشاهده می‌کنید؛



شکل ۱۸. تاریخ خبر

۵. ارزیابی

به منظور ارزیابی مجموعه داده مورد نظر با سایر مجموعه داده‌ها به تفکیک و در دو جدول مجزا و با در نظر گرفتن معیارهای جامعیت، تعداد داده‌های مناسب، وجود ویژگی‌های مفید، دارا بودن ویژگی‌های منحصربه‌فرد و همچنین ذخیره‌سازی در قالب استاندارد برای جدول شماره ۱ و با عنوان ویژگی‌ها و همچنین با در نظر گرفتن معیارهای دسته‌بندی متن، استخراج متن، خلاصه‌سازی متن و تشخیص عنوان برای جدول شماره ۲ و با عنوان کاربردها سعی کردیم که مقایسه‌ای نسبت به مجموعه داده IHU-PersianNewsDataSet-Javadzade-et-al با سایر مجموعه داده‌های موجود انجام دهیم.

در جدول ۱ به مقایسه ویژگی‌های مجموعه داده‌های معرفی شده می‌پردازیم. با توجه به موارد بیان شده درمی‌یابیم که ایرادات هرکدام از مجموعه دادگان‌ها به چه صورت است. همچنین به این نکات دست می‌یابیم که مجموعه دادگان IHU-PersianNewsDataSet-Javadzade-et-al ایرادات وارده بر سایر موارد را پوشش داده است.

جدول ۱. مقایسه ویژگی‌های مجموعه داده‌های دیگر با مجموعه داده IHU-PersianNewsDataSet-Javadzade-et-al

	جامعیت	تعداد داده‌های مناسب	وجود ویژگی‌های مفید	ویژگی منحصربه‌فرد	ذخیره استاندارد
دادگان ۱۷۵ هزارتایی	✓	✓	✗	✓	✗
دادگان هزارتایی	✗	✗	✗	✗	✗
توسعه اندیشه نوین	✗	✓	✓	✗	✓

دادگان همشهری	✗	✓	✓	✓	✓
IHU-PersianNewsDataSet-Javadzade-et-al	✓	✓	✓	✓	✓

در رابطه با کاربرد مجموعه دادگان، قضیه به همین منوال است. منظور از جامعیت همان امکان به‌کارگیری یک مجموعه در یکی از موارد پژوهشی در حوزه پردازش زبان طبیعی است. از نمونه این کاربردها می‌توان به دسته‌بندی (موضوعی) متن، استخراج متن، خلاصه‌سازی متن و تشخیص عنوان اشاره نمود. این کاربردها در طراحی مجموعه دادگان معرفی شده مورد توجه قرار گرفته است و به‌گونه‌ای طراحی شده است که در تمامی موارد کاربرد دارد. در جدول ۲ به بررسی کاربرد هرکدام از مجموعه دادگان می‌پردازیم.

جدول ۲. مقایسه کاربردهای مجموعه داده‌های دیگر با مجموعه داده IHU-PersianNewsDataSet-Javadzade-et-al

	دسته‌بندی متن	استخراج متن	خلاصه‌سازی متن	تشخیص عنوان
دادگان ۱۷۵ هزارتایی	✓	✓	✗	✓
دادگان هزارتایی	✗	✓	✓	✓
توسعه اندیشه نوین	✗	✓	✗	✓
دادگان همشهری	✓	✓	✗	✗
IHU-PersianNewsDataSet-Javadzade-et-al	✓	✓	✓	✓

۶. نتیجه‌گیری

یکی از الزامات توسعه سیستم‌های مبتنی بر پردازش زبان طبیعی دارا بودن یک مجموعه داده‌ای است که دارای طیف وسیعی از ویژگی‌ها و کاربردها باشد که این امر امروزه به یکی از چالش‌های بسیار مهم در این حوزه تبدیل شده است. در این کار ما با استفاده از طراحی یک خزشگر سعی کرده‌ایم که مجموعه داده‌ای را طراحی کنیم که شامل پنج ویژگی عنوان، خلاصه، متن

اصلی، برچسب و تاریخ انتشار خبر می‌باشد. گفتنی است که برچسب‌های این مجموعه داده همان‌طور که گفته شده طیف وسیعی از موضوعات در دنیای کنونی را در بر گرفته و شامل ۱۳ برچسب اصلی ورزشی، هنر و رسانه، فرهنگ، علم و پیشرفت، سیاسی، سیاست خارجی، زندگی، خانواده، جامعه، تعلیم و تربیت، بین‌الملل، اقتصادی و استان‌ها می‌باشد. از جمله کارهایی که بر روی این مجموعه داده قابل انجام است می‌توان به دسته‌بندی متن، استخراج متن، خلاصه‌سازی متن و تشخیص عنوان اشاره کرد و همچنین از ویژگی‌های بارز این مجموعه داده می‌توان به جامعیت، تعداد داده‌های مناسب، وجود ویژگی‌های مفید، دارا بودن ویژگی‌های منحصربه‌فرد و همچنین ذخیره‌سازی در قالب استاندارد اشاره کرد. خزشگر مذکور با استفاده از زبان برنامه‌نویسی پایتون طراحی گردیده و داده‌ها را جمع‌آوری کرده است و سپس پس از نرمال‌سازی داده‌ها در قالب دو فایل csv و xml ذخیره‌سازی گردیده است. این مجموعه داده محصول گروه پردازش زبان دانشگاه جامع امام حسین (ع) می‌باشد که از طریق لینک مذکور در بخش مقدمه و با رعایت حق کپی‌رایت جهت استفاده پژوهشگران محترم قابل دریافت می‌باشد.

منابع و مراجع

۱. میلاد یوسفی،۲۰۲۰، مجموعه داده اخبار فارسی برگرفته از

https://github.com/milad-4274/persian_news

رؤیت شده در تاریخ ۱۴ اردیبهشت ۱۴۰۱

۲. مجموعه داده هزار خبر فارسی با مشخصات هر خبر برگرفته از

http://dataheart.ir/article/3493/%D8%AF%DB%8C%D8%AA%D8%A7%D8%B3%D8%AA-%D9%87%D8%B2%D8%A7%D8%B1-%D8%AE%D8%A8%D8%B1-%D9%81%D8%A7%D8%B1%D8%B3%DB%8C-%D8%A8%D8%A7-%D9%85%D8%B4%D8%AE%D8%B5%D8%A7%D8%AA-%D9%87%D8%B1-%D8%AE%D8%A8%D8%B1

رؤیت شده در تاریخ ۱۴ اردیبهشت ۱۴۰۱

۳. گروه توسعه اندیشه نوین،۲۰۱۶، مجموعه داده کاملی از اخبار فارسی برگرفته از

https://tanoco.ir/datamining/%D8%AF%DB%8C%D8%AA%D8%A7%D8%B3%D8%AA-%DA%A9%D8%A7%D9%85%D9%84%DB%8C-%D8%A7%D8%B2-%D8%A7%D8%AE%D8%A8%D8%A7%D8%B1-/%D9%81%D8%A7%D8%B1%D8%B3%DB%8C

رؤیت شده در تاریخ ۱۴ اردیبهشت ۱۴۰۱

۴. مجموعه داده کامل همشهری نسخه ۱ شامل ۱۶۶ هزار سند برگرفته از

http://dataheart.ir/article/3487/%D9%85%D8%AC%D9%85%D9%88%D8%B9%D9%87-%D8%AF%D8%A7%D8%AF%D9%87--%DA%A9%D8%A7%D9%85%D9%84-%D9%87%D9%85%D8%B4%D9%87%D8%B1%DB%8C-%D9%86%D8%B3%D8%AE%D9%87-1-%D8%B4%D8%A7%D9%85%D9%84-166-%D9%87%D8%B2%D8%A7%D8%B1-%D8%B3%D9%86%D8%AF-%D8%AF%D8%B1-%D9%81%D8%B1%D9%85%D8%AA-%D8%A7%DA%A9%D8%B3%D9%84-%D9%88-csv

رؤیت شده در تاریخ ۱۴ اردیبهشت ۱۴۰۱

۵. رستگار، امیر (۲۰۱۹). کراولر چیست و چگونه یک نمونه از آن را بسازیم.
https://virgool.io/@rastegar.amir3/%DA%A9%D8%B1%D8%A7%D9%88%D9%84%D8%B1-%DA%86%DB%8C%D8%B3%D8%AA-%D9%88-%DA%86%DA%AF%D9%88%D9%86%D9%87-%DB%8C%DA%A9-%D9%86%D9%85%D9%88%D9%86%D9%87-%D8%A7%D8%B2-%D8%A2%D9%86-%D8%A8%D8%B3%D8%A7%D8%B2%DB%8C%D9%85-dyjxsq3kdoco
۶. پایگاه اطلاع رسانی پلیس فتا (بی تا). URL چیست؟.
https://cyberpolice.ir/node/140810
۷. آی پاور میزبان مطمئن خدمات مبتنی بر وب شما (بی تا). ریدایرکت چیست؟ نحوه ریدایرکت کردن یک آدرس به آدرس دیگر در cPanel.
https://www.irpower.com/kb/redirect-in-cpanel
۸. خانه بیگ دیتا (بی تا). پارس کردن صفحات وب با کتابخانه beautifulsoup پایتون.
https://bigdata-ir.com/%D9%BE%D8%A7%D8%B1%D8%B3-%D8%B5%D9%81%D8%AD%D8%A7%D8%AA-%D9%88%D8%A8-%D8%A8%D8%A7-%DA%A9%D8%AA%D8%A7%D8%A8%D8%AE%D8%A7%D9%86%D9%87-beautifulsoup-%D9%BE%D8%A7%DB%8C%D8%AA%D9%88/#

Design and production of Persian news data set IHU-PersianNewsDataSet-Javadzade-et-al Imam Hossein Comprehensive University

Hossein Hosseini[1], Mohammad Ghalenoei[2], Mohammad Mahdi Mokhtari[3], Mohammad Ali Javadzade[4]

[1]Master student of Imam Hossein Comprehensive University hosseinhosseini@ihu.ac.ir

[2]Master student of Imam Hossein Comprehensive University mghalenoei@ihu.ac.ir

[3]Master student of Imam Hossein Comprehensive University m_m_mokhtari@ihu.ac.ir

[4]Assistant Professor of Imam Hossein Comprehensive University javadzade@ihu.ac.ir

Abstract

Although the lack of data is one of the important challenges for research in the field of natural language processing, but this challenge is more acute in the Persian language, so finding a high-quality and comprehensive dataset in the Persian language is a difficult task. In addition to that, having some problems such as the ability to categorize and not complying with the storage standard are among the problems of the existing datasets, each of which can affect the learning rate of the model, the results, and the error rate in the experiments. For this reason, all these reasons made us seek to collect and prepare a dataset that covers all such problems and reduces the amount of error when using data in different models. In this research, we have designed and used a crawler to collect textual data. By crawling on one of the news bases, it has been able to collect data sets in five columns: title, summary, text, tag, and publication date. The textual data has been normalized with the help of one of the Persian language libraries in the Python programming language and stored in csv and xml formats and made available to fellow researchers. The tags in this dataset include 13 main tags of sports, art and media, culture, science and progress, political, foreign policy, life, family, society, education and training, international, economic and provinces. Among the tasks that can be done on this data set are text classification, text extraction, text summarization and title recognition. Also, one of the prominent features of this data set is its comprehensiveness, the amount of suitable data, the existence of useful features, having unique features, as well as storage in a standard format. This dataset is a product of the Language Processing Department of Imam Hossein Comprehensive University and can be downloaded and used through the link mentioned in the footnote of the next page and with respect to copyright.

Keywords: dataset, Persian news, natural language processing, Persian news dataset, machine learning, text classification, text extraction, text summarization, title recognition