بنام خدا

وزارت علوم، تحقیقات و فناوری

مرکز منطقه ای اطلاع رسانی علوم و فناوری

گزارش نهایی طرح تحقیقاتی

استخراج واژه‌های عمومی (غیر موضوعی) برای حوزه‌ی دامپزشکی از یک سو و حوزه‌ی علوم انسانی (ادبیات فارسی) از سوی دیگر و بررسی آن‌ها

مجری :

شاپوررضا برنجیان

مربی گروه پژوهشی زبانشناسی رایانه‌ای

دی ماه ۱۳۹۹

فهرست مندرجات

عنوان **صفحه**

چکیده ۴

فصل اول مقدمه ۴

بیان مسئله ۵

ضرورت تحقیق ۵

هدف تحقیق ۷

سوالات پژوهش ۷

فصل دوم پیشینه‌ی تحقیق ۷

مبانی نظری ۸

فصل سوم روش شناسی پژوهش ۹

انواع واژه‌ها بر اساس غیر موضوعی ۹

انواع فهرست واژه‌های غیرموضوعی ۱۰

کاربرد نتایج ۱۷

تعریف واژه‌های غیرموضوعی ۱۸

روش انجام پژوهش ۱۹

اشکالات موجود در فهرست‌های بدست آمده ۲۰

روش و ابزار گردآوری داده‌ها ۲۲

فصل چهارم یافته ها ۲۲

خروجی‌های طرح ۲۳

موارد استفاده‌ی طرح ۲۳

فصل پنجم بحث و نتیجه‌گیری ۲۳

تعداد واژه‌های غیر موضوعی ۲۶

پیشنهادات ۲۷

منابع و ماخذ ۲۷

سپاسگزاری

تهیه، تنظیم و تدوین این طرح بدون همکاری و یاری افراد گوناگون امکان پذیر نبود، لذا در اینجا بر خود لازم می‌دانم تا از تک تک آن بزرگواران تشکر و سپاسگزاری نمایم.

۱) جناب آقای دکتر محمد جواد دهقانی، ریاست محترم مرکز منطقه‌ای که در تمام مراحل کار با دقت و حوصله نظارت مستمر داشته و همواره راهنما و مشوق بوده‌اند.

۲) جناب آقای دکتر محمدرضا صالحی، معاونت محترم پژوهشی و فناوری مرکز منطقه‌ای که در تمام مراحل کار اعم از تهیهٔ پروپوزال و ... با دقت و حوصله همواره راهنما و مشوق بوده و نظرات و پیشنهادات سازنده‌ای ارائه فرموده و امکان اجرایی شدن این طرح را فراهم نموده‌اند.

۳) اعضای محترم شورای علمی که در هنگام تصویب طرح نظرات سازنده‌ای ارائه نمودند.

شاپوررضا برنجیان

مربی گروه پژوهشی زبانشناسی رایانه‌ای

استخراج واژه‌های عمومی (غیر موضوعی) برای حوزه‌ی دامپزشکی از یک سو و حوزه‌ی علوم انسانی (ادبیات فارسی) از سوی دیگر و بررسی آن‌ها

چکیده

آنچه در این پژوهش مورد بررسی قرار می‌گیرد تعیین واژه‌های غیرموضوعی، در هر یک از دو حوزه‌ی دامپزشکی و علوم‌انسانی می‌باشد.

جامعه‌ی آماری این پژوهش را ۲۳۶ مقالهٔ علمی پژوهشی از مجلات آی-اس-سی (۱۱۸ مقاله در حوزه‌ی دامپزشکی و ۱۱۸ مقاله در حوزه‌ی ادبیات فارسی) تشکیل می‌دهد. ما برای تهیهٔ فهرست واژه های غیرموضوعی از روش نیمه خودکار استفاده نمودیم. روش تحقیق این پژوهش توصیفی- تحلیلی و بررسی پیکره بنیاد است. بر اساس یافته‌های پژوهش، واژه‌های غیرموضوعی علوم‌انسانی توسط متخصص موضوعی ۲۳۸ واژه و واژه‌های غیر موضوعی دامپزشکی ۵۴۰ واژه می‌باشند.

نتایج بدست آمده نشان دادند که همه‌ی مجلات از یک نوع واژه‌های غیرموضوعی بهره نگرفته‌اند. و هر کدام با درجه‌های متفاوتی از شدت و ضعف این واژه‌ها را بکار برده‌اند. از واژه‌های غیرموضوعی مشترک بکاررفته در نشریات، می‌توان به اصوات، قیود، صیغه‌های افعال، حروف بکار رفته به جای اعداد، ضمایر، صفات، اعداد، نشانه‌های اختصاری اشاره کرد. همچنین تفاوت‌های مشهودی در نوع واژه‌های غیر موضوعیِ هر دو حوزه مشاهده می‌شد. از جمله تفاوت‌های موجود در بین واژه‌های غیرموضوعی می‌توان به "مصادر افعال" و "داروها" اشاره نمود که اولی در حوزه دامپزشکی و دومی در حوزه علوم انسانی جزء واژه‌های غیرموضوعی به حساب می‌آیند.

کلیدواژه: واژه‌های غیرموضوعی، بازیابی اطلاعات، واژه‌های‌عمومی، واژه‌های غیر مجاز، بازدارنده.

فصل اول

۱-مقدمه

در سامانه‌های پیشرفته ذخیره‌سازی و بازیابی اطلاعات، شیوه‌های جستجو از توجه ویژه‌ای برخوردار هستند، زیرا چگونگی جستجو و بکار بردن واژه‌ها و شناخت مناسب آن‌ها، تاثیر مستقیمی هم بر سرعت بازیابی و هم بر حجم اِشغال فضای اضافی و هم رضایت کاربران دارد. یک دسته از این واژه‌ها

که امروزه مورد بحث زبانشناسان و اطلاع‌رسان‌ها قرار می‌گیرد واژه‌های غیرموضوعی است، حذف این واژه‌ها می‌توانند در کارائی بهترِ نمایه کردن اسناد و مدارک نتیجه ساز باشند، از این رو تهیه‌ی یک لیست از واژه‌های غیر موضوعی در هر زبان و هر حوزه از علوم، مستلزم دانستن معیارها و روش‌های مختلف تولید این لیست می‌باشد. تا کنون در تعدادی از زبان‌ها از جمله در انگلیسی، لیست‌هایی از واژه‌های تهی انتشار یافته است، با این وجود، هنوز یک لیست واژه‌های تهی استاندارد ازمتون زبان فارسی استخراج نشده است. در این گزارش تلاش بر آن است تا با تکیه بر توسعه‌ی نظام‌های بازیابی اطلاعات، دستور العمل‌ها و معیارهای علمی زبانشناسی، جهت تهیه‌ی سیاهه‌ی واژه‌های غیر موضوعی در زبان فارسی ارائه گردد.

۱-۱- بیان مسئله

با ورود رایانه به عرصه اطلاع رسانی، نظام‌های ذخیره‌سازی و بازیابی‌اطلاعات دستخوش تغییر و دگرگونی اساسی شده‌اند. این تغییرات از آن جهت مورد توجه است که، می‌بایست این نظام‌ها که قبلاً با نظام‌های دستی و سنتی وفق داده شده و پیشرفت‌های قابل توجهی نیز مشاهده گردیده بود، اکنون می‌بایست خود را با قابلیت‌های کم و بیشِ تکنولوژی مدرن هم‌سو نموده و پیشرفت‌های خود را مجدداً از سرگرفته و بکار برد (حری، ۱۳۷۳). بحث اصلی در اینجا، چگونگی بازیابی اطلاعات از رایانه و بکارگیری صحیح و حتی‌الامکان بدون نقص واژگانِ موضوعی و تشخیص آن‌ها از واژگان غیرموضوعی در حوزه‌های علوم‌انسانی و دامپزشکی می‌باشد.

۱-۲- اهمیت تحقیق

امروزه با توجه به حجم انبوه اطلاعات، ابزارها و شیوه‌های سنتی و قدیمی برای کنترل اطلاعات کافی نیست و بی شک استفاده از روش‌های دستی در این زمینه غیراقتصادی بوده و مقرون به صرفه نمی‌باشند (مانند: Google که حدوداً دارای بیش از ۱/۴۰۰/۰۰۰/۰۰۰ صفحه اطلاعات می‌باشد) بنابراین برای انجام صحیح تجزیه و تحلیل متون با قالب زبان طبیعی اطلاعات، به سازماندهی خودکار اطلاعات نیازداریم، به عبارتی دیگر برای سرعت بخشیدن به تجزیه و تحلیل صحیحِ داده‌ها، نیاز به نظام‌های هوشمند داریم. یکی از شاخص‌های مهم در طبقه‌بندی ارزش و اعتبار نظام‌های بازیابی، نحوهٔ پردازش و تجزیه و تحلیل متون می‌باشد، نظام‌های بازیابی اطلاعات به سه گروه هوشمند (خودکار) و غیر هوشمند (غیر خودکار) و نیمه هوشمند (نیمه‌خودکار) تقسیم می‌شوند، در نظام‌های هوشمند، پردازش به صورت اتوماتیک و خودکار صورت می‌گیرد. نظام‌های هوشمند برای اولین بار بین سال‌های ۱۹۶۲ و ۱۹۶۵ در دانشگاهِ هاروارد طراحی و به کار گرفته شدند (سالتون، ۱۹۶۵). یکی از اشکالات

اینگونه نظام‌ها وجود اطلاعات و سندها و دیتاهای کاذب می‌باشند، نظام‌های غیرخودکار و دستی هم همانگونه که قبلاً نیز گفتیم، با توجه به حجم بالای داده‌ها اقتصادی و مقرون به صرفه نیست، اما نظام‌های نیمه خودکار که در حقیقت ادغامی است از هر دو نظام، می‌توانند مزایای هردو را داشته باشند و در عین حال نواقص هردو را رفع نمایند. ما از روش نظام‌های نیمه‌خودکار بهره برده‌ایم.

۱-۳- ضرورت تحقیق

امروزه اکثریت جستجوها با توجه به گستردگی رایانه در کتابخانه‌ها و افزایش کتابخانه‌های آنلاین، و مجازی و دیجیتالی و افزایش استفاده جستجوگران بطور اعم از شبکه جهانی وب، موجب توجه بیش از حد مسئولین و دست اندرکاران اینگونه کتابخانه‌ها و مسئولین سایت‌ها و شبکه‌های اطلاع‌رسانی به تئوری‌های ذخیره‌سازی و بازیابی‌اطلاعات مختلف گردیده است. در این میان نوع و چگونگی جستجو در رایانه‌ها، از توجه ویژه‌ای برخوردار می‌باشند، زیرا چگونگی جستجو می‌تواند عامل موثری باشد درجهت دستیابی یا عدم دستیابی جستجوگران به منابع مورد نیازشان. نا گفته پیداست که این عمل ارتباط مستقیم با نوع نمایه سازی نیز دارد. نمایه‌سازی و بازیابی‌اطلاعات همواره بر پایه کلید واژه‌ها استوارند، البته توجه به چگونگی ذخیره و بازیابی کلید واژه‌ها همواره مورد توجه و بحث و گفتگو بین اطلاع‌رسان ها و متخصصان علم اطلاعات و دانش‌شناسی بوده و هست، زیرا رضایتمندی جستجوگران و کاربران بستگی تام به انتخاب واژه‌های مناسب در هنگام جستجو دارد. این در حالی است که انتخاب واژه‌های مناسب در جهت جستجوی مطلوب برای بسیاری از کاربران دشوار می‌باشد (فتاحی ۱۳۸۵) و بکارگیری واژه‌های نامناسب موجب افزایش بازیافت اسناد و مدارک کاذب می‌گردد. یکی از اِشکالات ذخیره سازی و بازیابی اطلاعات آن است که اکثر مواقع نمایه سازان کلید واژه‌های اصلی که در اسناد و مدارک دارای وزن و معنی می باشند را ثبت کرده و کاربران نیز فقط همانگونه کلید واژه‌ها را در جستجوها بکارمی برند (اینگونه نمایه سازی را نمایه سازی کنترل شده می‌گویند) و این مسئله موجب بازیابی تعداد زیادی از مدارک بی‌ربط می‌شوند که مورد نیاز و هدف کاربران نیست. در حالی که ممکن است جنبه های دیگری ازموضوع مورد نظر کاربران باشد که اینگونه جستجوها قادر به پاسخ گویی به آن موضوعات نمی‌باشند (فتاحی ۱۳۸۵). امّا فنون پردازش زبان طبیعی[2] به طول پرس‌وجوها بستگی دارد، هرچه پرس‌وجوها طولانی‌تر باشند، سودمندی پردازش زبان طبیعی بیشتر خواهد بود (مهراد؛ ۱۳۸۷) جهت دست یابی به این نیازها وافزایش میزان دقت در نتایج بازیابی می‌توان واژه‌هایی را به کلید واژه‌ها افزود که این واژه‌ها "واژه های‌تهی"، "واژه‌های عمومی" یا "واژه‌های غیر موضوعی" نامیده می‌شوند.

[2] – NLP (Natural language processing)

این واژه‌های غیر موضوعی در پردازش بازیابی اطلاعات دو اثر گوناگون و متفاوت دارند:

۱- می‌توانند تأثیر در کارایی بازیابی اطلاعات داشته باشند. زیرا آن‌ها دارای بسامد بالا هستند.

۲- به کاستنِ تأثیر تفاوت‌های بسامدی در میان واژه‌های متداولِ منفی و اثر گذاری بر ارزش پردازش، گرایش دارند.

حال با توجه به موارد فوق‌الذکر و بررسی‌های نگارنده‌ی این سطور، تا کنون پژوهشی در مورد «استخراج واژه‌های عمومی (غیر موضوعی) برای حوزه‌ی دامپزشکی از یک سو و حوزه‌ی علوم انسانی (ادبیات فارسی) از سوی دیگر و بررسی آن‌ها» صورت نگرفته، در ضمن لازم به توضیح است که؛ تفاوت این طرح با سایر تحقیقات انجام گرفته عبارتند از:

۱-استخراج واژه‌های غیر موضوعی بر اساس نظر متخصصان موضوعی دو حوزه‌ی دامپزشکی و علوم انسانی

۲-استخراج واژه‌های غیر موضوعی بر اساس تواترِ بیشتر از ۳۰۰ بارِ هر دو حوزه‌ی دامپزشکی و علوم انسانی

۳-تهیه‌ی فهرست واژه‌های غیر موضوعیِ مطلقِ زبان فارسی

۴-تهیه‌ی فهرست واژه‌های غیر موضوعی بر اساسِ مطلق و متخصصان موضوعی، در هر دو حوزه.

۵-مقایسهٔ واژه‌های غیر موضوعی دو حوزهٔ دامپزشکی و علوم انسانی و بررسی آن‌ها

۶- ذکر تعداد بسامد کلمات.

۱-۴- هدف تحقیق

تحلیل کلماتِ یک متن نشان می‌دهد که، گروهی از کلمات بی اهمیت وجود دارد که به فراوانی در متن ظاهر می‌شود. گروهی نیز وجود دارد که به ندرت در متن ظاهر می شوند و ممکن است نشان دهنده‌ی محتوای اطلاعاتی متن نباشند (ویکری و ویکری، ۱۳۸۰)،

هدف ما در این تحقیق استخراج واژه‌های عمومی‌خاص (غیر موضوعی‌خاص) برای حوزه‌ی دامپزشکی و واژه‌های غیرموضوعی‌خاص برای حوزه‌ی علوم انسانی (ادبیات فارسی) و مقایسه‌ی آن‌ها با یکدیگر می‌باشد.

۱-۵- سئوالات پژوهش:

سوالات این پژوهش عبارتند از:

۱-سیاههٔ واژه‌های غیر موضوعی خاص حوزه‌ی دامپزشکی کدام است؟

۲-سیاههٔ واژه‌های غیر موضوعی خاص حوزه‌ی علوم انسانی (ادب فارسی) کدام است؟

۳-آیا واژه‌های غیرموضوعی در حوزه‌های مختلف متفاوت هستند یا خیر؟

فصل دوم

۲- پیشینه تحقیق

نخستین بار(فوکس)[3] فهرستِ واژه‌های تهی درزبان انگلیسی را تهیه کرد، او یک فهرست شامل۴۲۱ واژه انگلیسی تهیه نمود (ابوالخیر۲۰۰۶). تاکنون در تعدادی از زبان‌ها از جمله (انگلیسی، فرانسه، آلمانی، عربی، روسی) فهرست‌هایی از واژه‌های تهی انتشار یافته است.

۲-۱- واژه‌های تهی درزبان فارسی

برای اولین بار ساووی (۲۰۰۸) نسبت به تهیه فهرست واژه‌های تهی در زبان فارسی اقدام نمود، اما این فهرست چون ترجمه فارسی از فهرست واژه‌های تهی زبان انگلیسی و عربی است لذا اشکالات اساسی دارد که از آن جمله می‌توان به موارد زیر اشاره نمود؛ **اول** آنکه، معیارهایی جهت تهیه این فهرست ارائه نشده، بنا بر این با معیارهای استخراج فهرست واژه‌های تهی زبان فارسی مطابقت ندارد و بعضی از واژه‌های موجود در فهرست اصلاً درزبان فارسی معنی ندارد؛ مانند (خیاه، بعری، تول، وگو، و...). **دوم** اینکه چون بر اساس پیکره زبان استخراج نشده، لذا هزار واژه دارای بسامد بالارا نیز در سایت مورد نظر نشان نداده است و بعضی از کلمات را نیز به شکل‌های مختلف آورده است مانند: (بله، بلی، آره، آری، و...) این فهرست فاقد بسیاری از واژه‌های تهی دائمی (مطلق) است که در زبان فارسی وجود دارد. و چون این فهرست، بر اساس یک حوزه‌ی موضوعی خاص تهیه شده لذا قابل تعمیم به سایر حوزه‌ها نیست و چنانچه این عمل انجام گیرد نتیجه مطلوب حاصل نخواهد شد. **سوم** آنکه بعضی از واژه‌های خارجی دخیل در فارسی مانند (مرسی)، و همچنین ضمایر متصل مانند (-شان، -ایم، -اش، و...) را نیز جزء واژه‌های تهی آورده است؛ همچنین "ه" {های غیر ملفوظ} را نیز که گاهی شناسه صفت فاعلی، و گاهی شناسه صفت مفعولی است جزء واژه‌های تهی آورده است. اگرچه بر اساس شکل نوشتاری خط فارسی این پسوندها گاهی می‌توانند جدا نوشته شوند، اما باید توجه داشت که آن‌ها بخشی از کلمه محسوب می‌شوند و حذف کردن آن‌ها می‌تواند تغییراتی در معنی ایجاد کند و یا آن‌ها را بی‌معنی نماید. بااین حال تولید یک فهرست استاندارد و کامل واژه‌های تهی در زبان فارسی به خاطر فقدان مجموعه‌ها و مطالعات آماری که استفاده از آن را توصیه نموده یا مخالف استفاده از آن باشد، کاری بسیار سترک به

[3] - Fox

حساب می‌آید، بعلاوه تولید این سیاهه‌ها اعم از اینکه در پردازش اطلاعات استفاده شوند یا نشوند در عمل مختلف هستند (مهراد۱۳۸۷).

در ایست واژه‌های **آکادمی داده** که ۸۱۰ واژهٔ غیر موضوعی (ایست واژه) ذکر شده است موارد زیر مشهود می‌باشد:

الف – اسامی مانند: آدم، روزه، دیوانه و ... به عنوان غیر موضوعی ذکر شده است.

ب – مصادر نیز به عنوان واژه‌های غیرموضوعی ذکر شده‌اند. این در حالی است که در حوزه‌ی زبانشناسی و زبان و ادب فارسی مصادر می‌توانند به عنوان واژه‌های موضوعی ذکر شوند. لازم به ذکر است که کلیهٔ افعال صرف شده جزء واژه‌های غیرموضوعی محسوب می‌شوند.

به نظر نگارنده این اشتباه از آنجا ناشی می‌شود که در تهیهٔ واژه‌های تهی می‌بایست ابتدا حوزهٔ مربوطه تعیین گردد زیرا در حوزه‌های مختلف واژه‌های مختلف در فهرست واژه‌های غیر موضوعی قرارمی‌گیرند.

در فهرست واژه‌های غیر موضوعیِ **خانهٔ بیگ دیتا** (Big data) که ۱۲۵۰ واژهٔ غیرموضوعی ذکر شده است. نیز موارد زیر مشهود است:

الف – معلوم نیست به چه دلیل واژهٔ «برای» را صرف کرده و مجدداً آن‌ها را که حدوداً پنج واژه می‌باشند جزء واژه‌های غیر موضوعی ذکر کرده‌اند.

ب – پیشوند «بر» را در اول واژه‌هایی چون «بر روی»، «بر خلاف»، «بر عکس» و ... حذف نکرده و آن‌ها را جزو واژه‌های غیرموضوعی آورده‌اند.

ج – واژهٔ «گفت» را جزو واژه‌های غیرموضوعی آورده‌اند ولی واژه‌هایی مانند: گفتم، گفتی، گو، گفتند و ... را جزو این فهرست نیاورده‌اند.

د – واژهٔ «باید» و همچنین با آوردن «ن» نفی «نباید» را دوباره جزو واژه‌های غیرموضوعی آورده اند.

هاشم‌زاده، نخعی، مرادی مقدم (۱۳۹۲) و بلندیان (۱۳۸۵) بر اساس کاربرد قانون زیف پژوهش خود را انجام داده‌اند، به این صورت که کاربرد قانون زیف را در خصوص تهیهٔ واژه‌های غیرموضوعی بکار برده‌اند. بلندیان (۱۳۸۵) از تحقیق خود چنین نتیجه‌گیری می‌کند که؛ هرگاه متن مقاله بر اساس فراوانیِ تکرار واژه‌ها شمارش شود، سیاهه‌ای به وجود خواهد آمد که در آن واژه‌ها به ترتیب فراوانی، از زیاد به کم منظم می‌شوند. بنابراین مقالات از الگوی پیش‌بینی‌پذیر زیف تبعیت می‌کنند. همچنین در متن همه‌ی مقالات واژه‌های بدون بار معنایی، دارای بالاترین فراوانی هستند. این واژه‌ها را می‌توان در

قالب حروف‌اضافه، ربط، اعداد، افعال، صفت‌ها، قیدها، ضمایر و ...دسته بندی کرد. همچنین، تفکیک به صورت شیوه‌ی "ماشینی با دخالت عامل انسانی" نسبت به شیوه‌ی "صرفاً ماشینی" نتیجه بهتری را به دست می‌دهد. هاشم‌زاده و همکاران (۱۳۹۲) نیز نتایج زیر را بدست آورده‌اند؛ از تعداد کلِ واژه‌های متن که "۷۳۷۵۰" واژه می‌باشند، "۳۵۴۷۷" واژه (۴۸/۱۰ ٪) واژه‌های بازدارنده می‌باشند. که پس از تعدیل (حذف واژه‌های معنادارتوسط متخصص موضوعی) این تعداد به "۳۲۴۲۸" واژه (۴۳/۹۷ ٪) رسیده است. بنابر گفته‌های خود هاشم‌زاده نتایج این پژوهش تا اندازه‌ای با یافته‌های پژوهش‌های سنجی (۱۳۸۷) و فاکس (۱۹۹۰) مطابقت دارد.

سنجی و داورپناه (۱۳۸۸) نیز علاوه براینکه از نظام خودکار استفاده کرده‌اند، نسبت به تهیه و شناسایی واژه‌های غیرمفهومی روی مدارک کتابداری و اطلاع‌رسانی کارکرده و این فهرست در حوزهٔ موضوعی کتابداری و اطلاع‌رسانی می‌باشد. سنجی نیز در تحقیق خود به این نتایج رسیده است که؛ از مجموع "۲۴۸۵۵۲" واژه‌ی بکار رفته در مقاله‌های مورد بررسی در هرسه رشته"۹۷۲۸۰" واژه (۱۲۹۱ واژه بدون احتساب بسامد) به عنوان واژه های غیرموضوعی در سه رشته‌ی مورد مطالعه شناخته شدند. از لحاظ نوع دستوری می‌توان بیان داشت که قیدها و افعال و حروف اضافه ، اعداد، ضمایر، ادات و ... بیشترین حجم از واژه‌های غیر مفهومی را به خود اختصاص داده‌اند. همچنین با احتساب علائم سجاوندی، از "۳۸۰۲۱۷" واژه تعداد واژه‌های بازدارنده"۱۳۰۰۶۱" واژه خواهد بود (۴۶/۴۱ ٪).

در ایست‌واژه‌های **آکادمی داده** ۸۱۰ واژهٔ غیر موضوعی (ایست واژه) ذکر شده است.

در دیتا هارت (۱۳۹۷) آمده است: مراحل پیش پردازش در علوم متن کاوی دارای اهمیت بسیاری هستند، یکی از مراحل متن کاوی، حذف ایست‌واژه‌ها است، ایست‌واژه‌ها کلماتی هستند که بار مفهومی زیادی را حمل نمی‌کنند. بنابراین در مرحلهٔ پیش پردازش حذف می‌شوند. سایت دیتا هارت ۵۴۲ ایست‌واژهٔ فارسی را ارائه داده است. در این سایت آمده است؛ ایست‌واژه‌ها به حروف اضافه‌ای گفته می شوند که مفهومی را منتقل نمی‌کنند. در پردازشِ متون به دنبال کلماتی هستیم که در رسیدن به مدل طبقه بندی دقیق‌تر، ما را یاری کند. در دیتا هارت ذکری از چگونگی تهیهٔ این ایست‌واژه‌ها به میان نیامده و شیوه‌ی تهیه‌ی آن‌ها معلوم نیست.

دو فهرست ایست‌واژه فارسی در دسترس است، یکی فهرست واژه شامل ۸۱۴ ایست واژه که در پروژهٔ دیتاستِ همشهری جمع‌آوری شده است. این فهرست از طریق لینک زیر در دسترس است.

http://dataacademy.ir/upbad/public.

فهرست دیگر شامل ۵۴۲ ایست واژهٔ فارسی می‌باشد که از طریق لینک زیر در دسترس است.

۲-۲- مبانی نظری

مقوله‌های غیر موضوعی، از موضوعات پیچیده‌ای هستند که تلاش و اندیشه زبانشناسان بسیاری را به خود معطوف کرده و تحلیل‌های متعددی در باره‌ی آن‌ها ارائه شده است. در این قسمت ابتدا به تعریف "واژه‌های غیر موضوعی" می‌پردازیم فتاحی می‌نویسد: واژه‌های عمومی (تهی) یا غیرموضوعی، واژه‌هایی هستند که معمولاً به تنهایی مورد جستجو قرار نمی‌گیرند، زیرا به خودی خود معنا و مفهوم خاصی ندارند. واژه‌های عمومی همواره، همراه با واژه‌های موضوعی می‌آیند تا جنبه خاصی از آن موضوع را نشان دهند. مانند: "مقدمه ای بر....." ، "آشنایی با ....." ، "درباره....." (فتاحی؛ ۱۳۸۵). ویژگی همگانی مقوله‌های تهی این است که، هیچکدام حامل "معنی اطلاعاتی" در اسناد نیستند بلکه دلیل کاربرد آن‌ها در جملات و عبارات فقط وجود نقش گرامری (دستوری) آن‌هاست. (زو۲۰۰۶) ، (روگاوان ۱۹۸۶) ، (ریکاردو۱۹۹۹).

بطور کلی می‌توان گفت: واژه‌های تهی کلمات بسیار متداول و معمولی هستند که در اسناد با معنی اندکی ظاهر می‌شوند و تنها حامل یک وظیفه ونقش نحوی هستند ولی فاقد موضوع و محتوا می باشند (ابوالخیر؛ ۲۰۰۶).

فتاحی واژه‌های نیمه‌موضوعی را نیز چنین تعریف میکند: واژه‌های نیمه‌موضوعی واژه‌هایی هستند که به طور معمول به تنهایی مورد جستجو قرار نمی‌گیرند بلکه مانند واژه‌های غیرموضوعی همراه کلید واژه‌های موضوعی می‌آیند. مانند: "ریسک..."، "حادثه..."، "پیشگیری از..."(فتاحی ۱۳۸۵). واژه‌های نیمه‌موضوعی یا نیمه‌تهی، اگر چه در ظاهر می‌توانند بعنوان کلیدواژه بکار روند اما چون دارای عمومیت هستند بنا بر این ترجیحاً با سایر واژه‌های مکمل بکار می‌روند.

فصل سوم

۳- روش شناسی پژوهش

۳-۱- انواع واژه‌ها براساس موضوعی و غیرموضوعی

واژه‌ها براساس موضوعی سه نوع می‌باشند که عبارتند از:

۳-۱-۱- واژه‌های موضوعی: واژه‌هایی هستند که به تنهایی معنا دارند و به صورت کلیدواژه مورد جستجو قرار می‌گیرند.

۳-۱-۲- واژه‌های نیمه موضوعی: واژه‌هایی هستند که اگر چه در ظاهر می‌توانند بعنوان کلیدواژه بکار روند اما چون دارای عمومیت هستند بنابراین ترجیحاً با سایر واژه‌های مکمل بکار می‌روند. مانند: "به عبارت دیگر ..." (قید تفسیر)، "دور از حضور شما" (قید ادب).

۳-۱-۳- واژه‌های غیرموضوعی: همانگونه که قبلاً نیز گفته‌شد واژه‌های عمومی (تهی) یا غیرموضوعی، واژه‌هایی هستند که معمولاً به تنهایی مورد جستجو قرار نمی‌گیرند، زیرا به خودی خود معنا و مفهوم خاصی ندارند. واژه‌های غیرموضوعی را "تهی، غیرمجاز، عمومی، بازدارنده، غیرمفهومی، ایست واژه" نیز نامیده‌اند؛ این عناوین به واژه‌هایی اطلاق می‌شوند که در اسناد و مدارک فاقد ارزش معنایی می‌باشند، ولی از نظر نحوی و دستوری بسیار مهم و تعیین کننده هستند و می‌توان گفت که اسکلت و استخوان‌بندی و داربستِ جملات بوده و به آن‌ها قوام و دوام می‌بخشند، و بدون آن‌ها جملات فاقد معنا و مفهوم می‌باشند. واژه‌های عمومی همواره، همراه با واژه‌های موضوعی می‌آیند تا جنبه خاصی از آن موضوع را نشان دهند.

۳-۲- انواع واژه‌های غیرموضوعی

واژه‌های غیر موضوعی سه نوع می‌باشند که عبارتند از:

واژه‌های غیرموضوعی بسیط: واژه‌هایی هستند که فقط از یک واژه تشکیل شده‌اند. مانند: "برای"، "آن"، "پس" و ... .

واژه‌های غیرموضوعی ترکیبی: واژه‌هایی هستند که از دو واژه یا بیشتر تشکیل شده‌اند مانند: "این که"، آن یکی"، "آن چنان که" و ... .

واژه‌های غیرموضوعی مرکب: عبارت‌های ناقصی هستند که از نظر نحوی برای تکمیل ساختار جملات بکار می‌روند. مانند: "مقدمه‌ای بر"، "به این معنا که" و ... .

۳-۳- انواع لیست واژه‌های غیر موضوعی

دستورالعمل‌ها (یا معیارهای) تهیه‌ی لیست واژه‌های تهی که استاندارد باشند فوق‌العاده مهم است، زیرا امروزه کار برد این لیست‌ها در پردازش زبان، جهت ذخیره‌سازی و بازیابی‌اطلاعات و مدارک الکترونیکی لازم و ضروری است. تا کنون لیست‌های متعددی از واژه‌های تهی که بطور سنتی از تجزیه و تحلیل بسامدی همه‌ی کلمات ازیک پیکره‌ی بزرگ زبانی استخراج شده برای زبان‌های مختلف

(انگلیسی، فرانسه، آلمانی و ...) تهیه شده است. نتایج بدست آمده از پیکره‌های مختلف اغلب کاملاً شبیه همدیگر هستند و عموماً به عنوان استاندارد به کار می‌روند.

اگر مجموعه‌ها مختص به رشته خاصی نباشد، در این صورت فهرست واژه‌های تهی باید شامل موارد زیر باشند.

۱-تمامی صیغه‌های صرف شده‌ی افعال، ۲- کلیه‌ی افعال معین، ۳- تمامی پسوندها و پیشوندها، ۴- واژه‌بست‌ها، ۵- ضمایرمتصل، ۶- حروف، ۷- اصوات، ۸- خطاب واژه‌ها، ۹- اعداد، ۱۰- واژه‌های جمع عربی و جمع با پسوندهای (ها، ات، جات، ان و ...)، ۱۱- ضمایر منفصل، ۱۲- کلیه قیود، ۱۳- صفات، ۱۴- ممیزها (واحدهای شمارش)، ۱۵- شبه جمله‌ها، ۱۶- حروف لاتین، ۱۷- نمادهای خاص، ۱۸- واژه‌های تکراری (منظور از واژه‌های تکراری واژه‌هایی هستند که از تکرار یک واژه ساخته می‌شوند مانند: "دورادور"، "سراسر" و ...).

اما اگر اطلاعاتی که قرار است تجزیه و تحلیل شوند مختص به یک رشته باشند، این فهرست باید بازبینی شده و یا توسط متخصص موضوعیِ رشته‌ی مورد نظر ارزیابی گردد. باید یادآوری کنیم که بسیاری از پژوهشگران امتیاز استفاده از عبارت‌های ثابت مانند عناصر موجود در فهرست واژه های تهی را خاطرنشان ساخته‌اند، بویژه استفاده از فهرست کوتاهی از عبارت‌های تهی[4] را توصیه می‌کنند (مهراد؛ ۱۳۷۸). پربسامدترین واژه‌های مشترک میان حوزه‌های مختلف یا هم پوشانی میان واژه‌های عمومی حوزه‌های مختلف دال بر آن است که این **واژه‌ها کاملا عمومی (عمومی مطلق)** هستند و وابستگی به حوزه موضوعی خاصی ندارند. مانند: (تاریخ...)، (مقدمه‌ای بر...)، (درباره...) در مقابل، واژه‌های عمومی دیگری نیز هستند که در حوزه‌های موضوعی خاصی کار برد دارند. مانند: (اخبار.....)، (اشکال.....) که اینگونه واژه‌ها را **واژه‌های عمومی خاص حوزه‌های تخصصی** می نامیم (فتاحی ۱۳۸۵) .

شیوه‌ی تهیه واژه‌های غیرموضوعی

بطور کلی جهت تهیه‌ی واژه‌های غیر موضوعی ابتدا می‌بایست پیکره‌ی[5] زبانِ مورد نظر آماده شود تا بتوان از آن استفاده نمود. فاکس (۱۹۹۰) برای تهیه‌ی واژه‌های غیرموضوعی برای زبان انگلیسی حدوداً پیکره‌ای متشکل از ۶۶۴۰۶۱۲۱ واژه مورد بررسی قرار داد و تعداد ۴۲۱ واژه استخراج کرد این واژه‌ها از نوع واژه‌های **عمومی مطلق** می‌باشند زیرا پیکره‌های مورد استفاده یک پیکره‌ی عمومی بوده.

---

[4] – Empty phrases

[5] - Corpus

بدیهی است برای تهیه‌ی واژه‌های **عمومی خاصِ** حوزه‌های تخصصی، می‌بایست پیکره آن حوزه مورد بررسی قرار گیرد. اینگونه واژه‌های عمومی مسائل ویژه خود را دارند و گاهی واژه‌هایی که جزو واژه‌های موضوعی می‌باشند، در اینگونه پیکره‌ها به عنوان واژه تهی محسوب می‌شوند، مانند واژه "فیزیک" در اسناد و مدارک و مقالات مربوط به حوزه‌ی "فیزیک".

عموماً از این پیکره‌ها، واژه‌ها را بر اساس تعداد بسامد بالا[6] یا تکرار این واژه‌ها در پیکره بعنوان واژه تهی استخراج می‌کنند زیرا اینگونه واژه‌ها اگر چه گاهی دارای بار موضوعی می‌باشند، اما ثبت آن‌ها موجب اِشغال فضای اضافی خواهد شد.

علاوه بر موارد فوق، بررسی و تصمیم گیری درموارد زیر جهت تهیه‌ی یک لیست واژه‌های تهی استاندارد در زبان فارسی الزامی بنظر می‌رسد. فهرست این واژه‌ها را صورت‌ها و مقولات دستوری زیر تشکیل می‌دهند:

۳-۳-۱ : قیدها.

۳-۳-۱-۱: قید مختص (کلمات تنوین دار.مانند: مندرجاً، مشروحاً، تدریجاً)

۳-۳-۱-۲ : قید مشترک (خوانا، زیبا)

۳-۳-۱-۳: گروه قیدی (در اواسط این دوره)

۳-۳-۱-۴: قید مرکب (هرروز، هرجا)

۳-۳-۱-۵: قید زمان (فروردین، عصر)

۳-۳-۱-۶: قید مکان (بالا، پایین، توی حیاط)

۳-۳-۱-۷: قید مقدار(کم، زیاد)

۳-۳-۱-۸ : قید چگونگی (خوب، زشت، کج، یواش)

۳-۳-۱-۹: قید حالت (خندان، دلیرانه، گریان)

۳-۳-۱-۱۰: قید آرزو یا تمنا (کاشکی، انشاءالله)

۳-۳-۱-۱۱ : قید تعجب (ای عجب، سبحان الله، شگفتا)

۳-۳-۱-۱۲ : قید تکرار (دیگر، دیگربار، بازهم)

۳-۳-۱-۱۳: قید تفسیر (به این معنی که، به عبارت دیگر)

۳-۳-۱-۱۴ : قید ترتیب (پیاپی، پی در پی)

۳-۳-۱-۱۵ : قید پرسش (کجا، آیا، چرا، هیچ)

---

[6] – High Frequency

۳-۳-۱-۱۶ : قید نفی (هرگز، به هیچ وجه، خیر)

۳-۳-۱-۱۷ : قید تصدیق و تاکید (الحق، بلی، براستی، هرآینه، حتما)

۳-۳-۱-۱۸ : قید تردید (شاید، به گمانم، ممکن، احتمالا)

۳-۳-۱-۱۹ : قید تشبیه (گویی، گفتی، پنداری)

۳-۳-۱-۲۰ : قید علت (ازاین رو، به این جهت، لهذا، به این دلیل، زیرا که)

۳-۳-۱-۲۱ : قید تبری و ادب (دور از حضور شما، دور از جناب، خدانکرده)

۳-۳-۱-۲۲ : قید اختصار (باری، فی الجمله، الغرض، خلاصه)

۳-۳-۱-۲۳ : قیدهای دیگری نیز وجود دارند که تعدادی از آن‌ها عبارتند از:

قید شمار (دوبار)، قید ترتیب (اول، دوم)، قید تاکید (بلاشک، بی گمان) قید شک (گویا) قید علت (به چه علت) قید وسیله (دستی، تلفنی)، قید تقریب (درحدود)، قید انحصار (فقط، تنها) قید ارزش (پنج تومان)، البته اقسام متمم‌های قیدی به مراتب بیشتر می باشد (فرشیدور؛ ۱۳۸۲)

۳-۳-۲ : افعال:

۳-۳-۲ -۱ : افعال ردیفی (بگیربشین، بزن بریم)

۳-۳-۲ -۲ : افعال اسنادی "افعالی هستندکه معنی کاملی ندارند" (کشتن، گردیدن)

۳-۳-۲ -۳ : افعال کمکی "معین" (بودن، شدن)

۳-۳-۲ -۴ : افعال بصورت مصدر (گفتن، رفتن)

۳-۳-۲ -۵ : افعال شبه کمکی (بایستن، شایستن، توانستن) (انوری؛ ۱۳۸۵)

۳-۳-۳ : صفات.

۳-۳-۳ -۱ : صفت پرسشی (چکونه، کدام، چند)

۳-۳-۳ -۲ : صفت تعجبی (چه سعادتی)

۳-۳-۳ -۳ : صفت مفعولی (شنیده، گرفته)

۳-۳-۳ -۴ : صفت نسبی (نمکین، فولادین، خودی، خودمانی، دروغین، پیشین)

۳-۳-۳ -۵ : صفت لیاقت (خواندنی، خواستنی)

۳-۳-۳ -۶ : صفت نفی (ناشایست، بی خود، بی دردسر)

۳-۳-۳ -۷ : صفت اشاره (این، آن، همین، همان، چنین، چنان، همچون، این قدر)

۳-۳-۳ -۸ : صفت ساده (خوب، بد، روشن، سرد، سبز، سرخ)

۳-۳-۳ -۹ : صفت مرکب (سبزرنگ، سرخ روی، سفید بخت)

۳-۳-۳ -۱۰ : صفت شمارشی اصلی (یک، دو، سه)

الف: صفت شمارشی ترتیبی (اول، هفتم، دهمین)

ب: صفت شمارشی کسری (سه دهم، چهار هفتم)

۳-۳-۳ -۱۱ : صفت مبهم (فلان، بهمان، برخی، هر)

۳-۳-۴ : واحد شمارش (ممیزها) . (نفر، اصله، باب، راس، قلاده)

۳-۳-۵ : ضمایر

۳-۳-۵ -۱ : ضمایر شخصی منفصل (من، تو، او، وی، مرا)

۳-۳-۵ -۲ : ضمایر مشترک (آن دیگری، این همه، خود، خویش، خودش)

۳-۳-۵ -۳ : ضمایر پرسشی (چندمین، چه سان، چه، کدام)

۳-۳-۵ -۴ : ضمایر مبهم (عده‌ای، گروهی، پاره‌ای)

۳-۳-۵ -۵ : ضمایر تعجبی (چه عالی!)

توجه : لازم بذکر است که صفات وضمایر "اشاره ،پرسشی، تعجبی و مبهم" معمولاً واژه‌های شبیه هم می‌باشند که برای تشخیص آن‌ها می‌توان گفت: اگر همراه اسم بیایند، صفت نامیده می‌شوند، ولی اگر به تنهایی و بدون اسم بیایند، آن‌ها را ضمیر می‌نامند (انوری۱۳۸۵).

۳-۳-۶ : حروف .

۳-۳-۶-۱ : حرف ربط:

الف : حرف ربط ساده (و، اگر، اما)

ب : حرف ربط مرکب (آنجاکه، آنگاه‌که)

۳-۳-۶-۲ : حرف اضافه:

۳-۳-۶-۲-۱ : حرف اضافه ساده (از، با)

۳-۳-۶-۲-۲ : حرف اضافه مرکب (غیراز، بدون)

۳-۳-۶-۳ : حرف ندا (آهای، هی)

۳-۳-۷ : شبه جمله ها (آفرین، آمین، افسوس)

۳-۳-۸ : واژه های تکراری (پاره پاره، تکه تکه، شلپ شلپ)

۳-۳-۹ : نمادهای خاص

۳-۳-۱۰ : حروف لاتین

همانگونه که گفتیم بررسی و تصمیم گیری درموردِ موارد فوق جهت تهیه‌ی یک لیست واژه‌های غیر موضوعیِ استاندارد در زبان فارسی الزامی است. اگرچه به نظر نگارنده‌ی این سطور، بر اساس تعاریف، این موارد می‌بایست جزو واژه‌های غیر موضوعی محسوب گردند.

جهت آگاهی بیشتر، به شیوه‌ی تهیه تعدادی از فهرست‌های تهیه شده اشاره می‌کنیم؛ ساووی (۱۹۹۹) برای زبان فرانسه فهرست واژه‌های غیرموضوع را استخراج کرد و این فهرست ۲۱۵ واژه را شامل می‌شود، وی نیز مانند فوکس (۱۹۹۰) و ابوالخیر (۲۰۰۶) از روش نیمه خودکاراستفاده کرده است واژه‌های موجود در این فهرست نیز واژه‌های «موضوعی مطلق» می‌باشند. ساووی و راسولوفو (۲۰۰۳) همچنین اولین فهرست واژه‌های غیرموضوعی را برای زبان عربی تهیه کردند. همانطور که قبلاً نیز ذکر شد، ساووی (۲۰۰۸) برای تهیه لیست واژه‌های غیرموضوعی در زبان فارسی نیز اقدام نمود به نظر می‌رسد وی برای تهیهٔ این فهرست از روش ویژه‌ای استفاده نکرده بلکه به ترجمهٔ آن‌ها بسنده کرده، بنابر این، چون این لیست ترجمه فارسی از لیست واژه‌های غیرموضوعی مطلق زبان فرانسه، انگلیسی و عربی بودند، لذا اشکالات اساسی داشت، که قبلاً در صفحه ۸ بطور مفصل ذکر گردید.

تقوا، بکلی و سده (۲۰۰۳) واژه‌های غیرموضوعی فارسی را برای استفاده در ریشه‌یاب فارسیِ خود تهیه کردند. این سیاهه نیز واژه‌های غیر موضوعی مطلق را شامل می‌شود.

سنجی (۱۳۸۵) در رسالهٔ کارشناسی ارشد خود، سیاهه‌ای از واژه‌های غیر موضوعی در زبان فارسی برای نمایه سازی خود کارِ متن‌های فارسی در رشتهٔ کتابداری و اطلاع رسانی انجام داده است، وی نیز جامعه‌ی آماری خود را از ۶۳ مقاله از مقاله‌های مندرج در آخرین شمارهٔ منتشر شده در مجله‌های علمی و پژوهشیِ کتابداری و اطلاع رسانی در سال ۱۳۸۵ تهیه کرده است. وی علائم سجاوندی را نیز جزء واژه‌های غیر موضوعی محسوب کرده. وی برای تهیهٔ این سیاهه، ابتدا مقاله‌ها را در محیط word تایپ نموده تا امکان تفکیک واژگان متن را داشته باشد، سپس به صورت ماشینی و با استفاده از فرامینِ موجود در نرم افزار word به تفکیک واژگان اقدام نموده است. سپس واژگان تفکیک شدهٔ هر متن را براساس معیارهای زبانشناسی، قواعد دستوری و آیین نگارش فارسی از لحاظ نوع و بارمعنایی به صورت دستی بررسی و ویرایش نموده است. برای شمارش واژگان مرتب شدهٔ هر متن از دستور word count در نرم افزار word استفاده شده است.

چون حروف‌اضافه در متون، در پیدا کردن الگوی پنهان در متن نقش ایفا نمی‌کند، بنابر این بهتر است در مرحلهٔ پیش پردازش حذف شوند. در حذف ایست واژه‌ها به دنبال حذف کلمات زائد برای رسیدن به پردازش متن بهینه هستیم.

امروزه در اکثر روش‌های پردازش متون، مرحلهٔ حذف ایست‌واژه‌ها به عنوان کلمات زائد انجام می پذیرد. فهرست کلمات ایست‌واژه‌ها یکی از معضلات برای زبان فارسی است، زیرا برای این زبان همچنان یک فهرست واژهٔ کامل ارائه نشده است و از فهرست واژه‌های جمع‌آوری شده به صورت دستی استفاده می‌شود.

در تمامی روش‌های متن کاوی و به طبع آن در روش‌های طبقه بندی متون انجام مراحل آماده سازی متن اجتناب ناپذیر است. و برای انجام مراحل آماده سازی متن، حذفِ ایست واژه‌ها الزامی است. با انجام این کار از بار اجرای الگوریتم به مقدار زیادی کاسته می‌شود.

۳-۴- روش انجام پژوهش

چون پیکرهٔ ما از دو حوزه‌ی خاص تشکیل شده و حداکثر مقالات در دسترسِ پژوهشگر بصورت فونت word ، ۲۴۰ مقاله در حوزهٔ دامپزشکی بوده‌اند، لذا، جهت نمونه‌گیری و محاسبهٔ حجم نمونه از جدول کرجسی و مورگان به آدرس: https://www.spss-iran.com/morgan-table) استفاده نمودیم. نمونه‌گیری تصادفی ما در این تحقیق، نمونه‌گیری خوشه‌ای است، به این صورت که ابتدا حوزه ی موضوعی را مشخص نموده (در این پژوهش حوزه‌ی مورد نظر، دامپزشکی و ادبیات فارسی است) و سپس ۲۳۶ مقالهٔ word (از هریک از حوزه‌ها ۱۱۸ مقاله) از مقالات موجود را انتخاب کرده و با استفاده از نرم افزار text Analyzer و در سایت https://www.online-utility.org/text/analyzer.jsp مورد بررسی قرار دادیم و بطور متوسط هر مقاله‌ی علوم انسانی تقریباً به ۲۲۲۰ واژه و هر مقاله‌ی دامپزشکی تقریباً به ۸۴۲ واژه تجزیه شد (علت این اختلاف حجم بالای مقالات علوم انسانی نسبت به مقالات حوزه دامپزشکی بودند). سپس تمامی نتایج بدست آمده در هر یک از حوزه‌ها را با هم ادغام کردیم و در حوزه‌ی علوم انسانی ۸۱۸۳۵ واژه و در حوزه‌ی دامپزشکی ۲۶۷۵۶ واژه بدست آمد. سپس به بررسی و تجزیه تحلیل هر یک از حوزه‌ها پرداختیم.

در ابتدای این بررسی‌ها به اشکالاتی بر خوردیم که اکثراً ناشی از تایپ مقالات در فضای نرم افزار word بوده‌اند که در زیر به تعدادی از آن‌ها با ذکر مثال‌های واقعی می‌پردازیم.

۳-۵- اشکالات موجود در فهرست‌های بدست آمده

همانگونه که قبلاً نیز اشاره کردیم در هنگام بررسی فهرست واژه‌های این دو حوزه به اشکالاتی بر خوردیم که خودِ این اشکالات باعث می‌شد تا صحتِ تعداد واژه‌های موجود در هر حوزه مورد تردید قرار

گیرد و این تردید موجب اختلال در صحت نتایج به دست آمده می‌گردد.

۱-گاهی واژه‌ها همراه با علائم سجاوندی بدون فاصله ظاهر می‌شدند (مانند: از، یا از").

۲-گاهی بعضی از واژه‌ها که دارای حرف «ی» بودند بصورت «ی عربی» نوشته شده‌اند (مانند: بازی)، که این امر باعث شده بود واژه‌هایی که دارای حرف «ی» بودند، به همان شکل تکرار شوند.

۳-در بعضی از واژه‌ها برای ثبت نیم فاصله بجای (ctrl+shif+2) از (ctrl+-) استفاده شده که این عمل در اکسل به صورت خط ظاهر می‌شود و در شکلِ واژه‌ها تغییر ایجاد می‌کند. (مانند: واژه-ها).

۴-پسوندها و پیشوندها به صورت چسبان و نیم فاصله و فاصله نیز دچار اشکال در جمع بندی تعداد یک واژه‌ی واحد می‌گردد. (مانند: واژه-ها، واژهها، واژه ها، و ...)، (احترامامیز، احترام-آمیز، احترامامیز، احترام‌آمیز) (احمدبن‌عبدالله، احمدبنعبدالله) (ادبیاتغربنویسندگانیچون) و ...

۵-نوشته شدن بعضی واژه‌ها با حمزه «أ» بجای «ا» (مانند: أحلام، احلام – أحمد، احمد – اختصاصاً، اختصاصا).

۶-افزودن یا کاستنِ حمزه (مانند: أحیاء، احیا / احیاگران، أحیاءگران)

۷-با افزودن یا کاستنِ واکه‌های کوتاه (مانند: اخصِّ، اخص)

۸-نوشتن «ۀ» بجای «ه» (مانند: ادامۀ بجای ادامه)

۹- نوشتن بعضی از حروف بجای حروف دیگر به دلیل نزدیکی حروف در صفحه کلید (مانند: «ز» بجای «ر» در استاز بجای استار).

با توجه به مشکلات فوق‌الذکر نسبت به یکدست‌سازی و تهیه‌ی فراوانی واژه‌ها اقدام نمودیم و پس از تصحیحات متوالی و بررسی در چندین مرحله، نتایج نهایی به‌شرح زیر بدست آمد.

واژه‌های تهی استخراج شده از پیکره‌ی مورد نظر توسط متخصصان موضوعی در حوزه‌ی علوم انسانی، **۲۳۸** واژه می‌باشند و واژه‌های تهی استخراج شده از پیکره‌ی مورد نظر در حوزه‌ی دامپزشکی توسط متخصص موضوعی **۵۴۰** واژه می‌باشند که از این تعداد، **۹۹** واژه به صورت مشترک (هم در حوزه‌ی دامپزشکی و هم در حوزه‌ی علوم انسانی) می‌باشند. این سیاهه‌ها توسط متخصصان موضوعی بررسی گردیده و مورد تایید قرارگرفتند.

۳-۶- روش و ابزار گردآوری داده‌ها

همانگونه که قبلاً نیز گفته شد، اگر مجموعه‌ها و اسناد، مختص به رشته خاصی نباشند، می‌بایست، فهرست واژه‌های غیرموضوعی شامل *قیود*، *حرف اضافه*، *حرف ربط*، *صفات*، *واحدهای شمارش*، *ضمایر*، *علائم سجاوندی*، *شبه جمله‌ها*، *اصوات و ...* باشند. اما اگر اطلاعاتی که قرار است تجزیه و تحلیل شوند،

مختص به یک رشته یا حوزه‌ی خاص باشند، این فهرست باید بازبینی شده و یا توسط متخصص حوزه‌ی موضوعیِ مورد نظر (متخصص موضوعی) ارزیابی شود.

بطور کلی جهت تهیه‌ی واژه‌های غیرموضوعی ابتدا می‌بایست پیکره‌ی[5] رشته‌ی مورد نظر مشخص شود تا بتوان از آن استفاده نمود.

عموماً برای استخراج واژه‌های غیرموضوعی از پیکره‌ها استفاده می‌کنند،

ما ابتدا انواع کلماتی که در اسناد و مدارک جزو واژه‌های «غیرموضوعی» محسوب می‌شوند را استخراج نمودیم و چون پیکره‌ای که فقط اسناد رشتهٔ خاصی را داشته باشد وجود ندارد، لذا برای نمونه‌گیری تصادفی در این تحقیق از نمونه‌گیری خوشه‌ای استفاده کردیم.

در ادامه ابتدا فهرستی از عناصر واژگانی مذکور در حوزه‌ی دامپزشکی را تهیه کرده و سپس، واژه هایی را که از تجزیه تحلیل بسامدی واژه‌ها بدست آمده‌اند را به آن‌ها افزودیم، این فهرست یک سیاههٔ کامل از واژه‌های غیرموضوعی رشتهٔ دامپزشکی را تشکیل داده و بسیار به حوزه‌ی موضوعی وابسته است بنابراین، برای حوزه‌های موضوعیِ دیگر چندان قابل استفاده نمی‌باشد، این حرف به این معناست که؛ واژه‌های غیرموضوعی مستخرج با این روش را نمی‌توان در تمامی حوزه‌های علمی زبان فارسی بکاربرد. (به عنوان مثال؛ اگر در حوزه‌ی علوم مهندسی، واژه‌ی "الگوریتم" جزو واژه‌های موضوعی تلقی شوند، این واژه را نمی‌توان در حوزه‌ی علوم انسانی (ادب فارسی) جزو واژه‌های موضوعی محسوب کرد).

ما برای تهیه‌ی فهرست واژه‌های غیرموضوعی از روش **نیمه خودکار** استفاده نمودیم، زیرا در روش خودکار، عملی بودن این روش بیشتر از آنکه به یک شاخه‌ی موضوعی خاص وابسته باشد، به مجموعه‌ی مدارک مورد پردازش بستگی دارد، بنابراین، باید در روش‌های پردازش تغییراتی اعمال شود و از قضاوت انسانی نیز استفاده شود. روش‌های استفاده می‌توانند بر اساس اهداف نظام‌ها با هم متفاوت باشند.

در روش نیمه خودکار که در این پژوهش بکار بردیم ابتدا فهرستی از واژگان و تعداد بسامد آن‌ها تهیه کردیم. تهیه‌ی این فهرست همانند روش خودکار بود، سپس مراحل زیر برروی آن‌ها انجام گرفت.

۱-واژه‌های با بسامد بالاتر از ۳۰۰ بار تکرار، به عنوان فهرست واژه‌های **غیرموضوعیِ بسامد بالا** انتخاب شدند.

۲-این واژه‌ها نیز از نظر موضوعی و معناداری مورد پالایش قرار گرفتند.

---

[5] - Corpus

توجه: قبلاً در پروپوزال برای انجام روش پژوهش در نظر داشتیم واژه‌های با بسامدِ زیرِ ۱۰۰ بار تکرار ، به عنوان فهرست واژه‌های **غیرموضوعی بسامد پایین** انتخاب شوند. اما با توجه به حجم بالای داده‌ها، واژه‌های زیر ۱۰۰ بار تکرار در علوم انسانی به بیش از ۴۳۰۰۰ واژه و در دامپزشکی به بیش از ۱۱۸۵۰ واژه می‌رسید؛ که اکثر آن‌ها واژه‌های موضوعی بودند، بر این اساس از ذکر واژه‌های غیرموضوعی **بسامد پایین** صرف‌نظر گردید.

فهرست واژه‌های بالای ۳۰۰ تواتر، بیشتر شاملِ مجموعه‌ای از علائم، کلمات، عبارات، پیشوندها، پسوندها، قیود، ضمایر، حروف اضافه و ... بودند.

۳- در این فهرست تعداد بسامد کلمات نیز ذکر شدند.

۴- سیاهه‌ی مربوطه ویرایش شده و واژه‌های بی معنا حذف شدند.

۵- واژه‌های تکراری حذف شدند.

۶- واژه‌هایی که به زبان و خط لاتین می‌باشند حذف شدند.

۷- پسوندهای جمع حذف شدند.

۳-۷- فهرست مجلات

فهرست و اطلاعات کتابشناختی مجلاتی که داده‌ها از مقالات آن‌ها استخراج گردیدند به قرار زیر می‌باشند.

مجله دامپزشکی ایران، دانشگاه شهید چمران اهواز دوره پانزدهم، شماره ۲، تابستان ۱۳۹۸
مجله دامپزشکی ایران، دانشگاه شهید چمران اهواز دوره چهاردهم، شماره ۱، بهار ۱۳۹۷
مجله دامپزشکی ایران، دانشگاه شهید چمران اهواز دوره دهم، شماره ۳، پاییز ۱۳۹۳
مجله دامپزشکی ایران، دانشگاه شهید چمران اهواز دوره یازدهم، شماره ۴، زمستان ۱۳۹۴
مجله دامپزشکی ایران، دانشگاه شهید چمران اهواز دوره دوازدهم، شماره ۱، بهار ۱۳۹۵
مجله دامپزشکی ایران، دانشگاه شهید چمران اهواز دوره دوازدهم، شماره ۲، تابستان ۱۳۹۵
مجله دامپزشکی ایران، دانشگاه شهید چمران اهواز دوره دوازدهم، شماره ۳، پاییز ۱۳۹۵
مجله دامپزشکی ایران، دانشگاه شهید چمران اهواز دوره دوازدهم، شماره ۴، زمستان ۱۳۹۵
دو فصلنامه علوم درمانگاهی دامپزشکی ایران دانشگاه شهر کرد دوره ۱۱ شماره ۱ بهار و تابستان ۱۳۹۶
دو فصلنامه علوم درمانگاهی دامپزشکی ایران دانشگاه شهرکرد دوره ۱۱ شماره ۲ پاییز و زمستان ۱۳۹۶
دو فصلنامه علوم درمانگاهی دامپزشکی ایران دانشگاه شهر کرد دوره ۱۲ شماره ۱ بهار و تابستان ۱۳۹۷
دو فصلنامه علوم درمانگاهی دامپزشکی ایران دانشگاه شهرکرد دوره ۱۲ شماره ۲ پاییز و زمستان ۱۳۹۷

نشریه ادب و زبان، دانشکدۀ ادبیّات و علوم انسانی، دانشگـــاه شهید باهنـــر کرمـــــان، سال ۱۷، شماره ۳۵، بهار و تابستان ۱۳۹۳
نشریه ادب و زبان، دانشکدۀ ادبیّات و علوم انسانی، دانشگـــاه شهید باهنـــر کرمـــــان، سال ۱۵، شماره ۳۲، پاییز و زمستان ۱۳۹۱
نشریه ادب و زبان، دانشکدۀ ادبیّات و علوم انسانی، دانشگـــاه شهید باهنـــر کرمـــــان، سال ۱۶، شماره ۳۳، بهار و تابستان ۱۳۹۲
مجله مطالعات ایرانی، دانشکدۀ ادبیّات و علوم انسانی، دانشگـــاه شهید باهنـــر کرمـــــان، سال ۱۳، شماره ۲۶، پاییز و زمستان ۱۳۹۳
نشریه ادبیات تطبیقی، دانشکدۀ ادبیّات و علوم انسانی، دانشگـــاه شهید باهنـــر کرمـــــان، سال ۷، شمارۀ ۱۳، پاییز و زمستان ۱۳۹۴
نشریه ادبیات تطبیقی، دانشکدۀ ادبیّات و علوم انسانی، دانشگـــاه شهید باهنـــر کرمـــــان، سال چهارم، شماره ۷، پاییز و زمستان ۱۳۹۱
نشریه ادبیات تطبیقی، دانشکدۀ ادبیّات و علوم انسانی، دانشگـــاه شهید باهنـــر کرمـــــان، سال ششم، شماره ۱۰، بهار و تابستان ۱۳۹۳
نشریه ادبیات پایداری، دانشکدۀ ادبیّات و علوم انسانی، دانشگـــاه شهید باهنـــر کرمـــــان، سال پنجم، شماره نهم، پاییز و زمستان ۱۳۹۲
نشریه ادبیات پایداری، دانشکدۀ ادبیّات و علوم انسانی، دانشگـــاه شهید باهنـــر کرمـــــان، سال ششم، شماره دهم، بهار و تابستان ۱۳۹۳

فصل چهارم

۴-یافته‌ها

یافته‌های این پژوهش به شرح زیر می‌باشند
در حوزه‌ی علوم انسانی تعداد واژه‌هایی که فقط یکبار بکار برده شده‌اند، ۲۳۸۰۰ واژه و آن‌ها که
دو بار بکار برده شده‌اند، ۶۲۶۵ واژه،
سه بار بکار برده شده‌اند، ۲۹۶۹ واژه
۴ بار بکار رفته‌اند ۱۹۳۱ واژه
۵ بار بکار رفته‌اند ۱۳۱۹ واژه
۶ بار بکار رفته‌اند ۱۰۰۳ واژه
۷ بار بکار رفته‌اند ۷۷۸ واژه
۸ بار بکار رفته‌اند ۶۱۳ واژه
۹ بار بکار رفته‌اند ۵۲۲ واژه می باشند

و چون آستانه‌ی مورد نظر ما ۳۰۰ در نظر گرفته شده بود، کلِ واژه هایی که از ۳۰۰ به بالا تواتر داشتند ۲۴۱ (دویست و چهل و یک) واژه بوده‌اند.

در حوزه‌ی دامپزشکی نیز تعداد واژه‌هایی که فقط
یکبار بکار برده شده‌اند، ۵۷۷۲ (پنج هزار و هفتصد و هفتاد و دو) واژه و آن‌ها که
۲ بار بکار رفته‌اند ۲۰۰۰ واژه
۳ بار بکار رفته‌اند ۱۰۱۲ واژه
۴ بار بکار رفته‌اند ۷۰۰ واژه
۵ بار بکار رفته‌اند ۴۹۰ واژه
۶ بار بکار رفته‌اند ۳۵۰ واژه
۷ بار بکار رفته‌اند ۳۰۰ واژه می‌باشند
و از آنجا که آستانه‌ی مورد نظر ما کلِ واژه‌های بالای ۳۰۰ تواتر بوده‌اند لذا تعداد کل اینگونه واژه‌ها در حوزه‌ی دامپزشکی ۱۵۹ (یکصد و پنجاه و نه) واژه بوده‌اند.

۴-۱- خروجی‌های طرح

ما **بر اساس نظر متخصصان** موضوعی توانستیم از کل واژه‌ها دو فهرست‌واژه (یکی برای حوزه‌ی دامپزشکی «۵۴۰ واژه» و دیگری جهت حوزه‌ی علوم انسانی «۲۳۸ واژه») تهیه نماییم.

سپس این فهرست‌ها را **همراه با فهرست واژه‌های تهی مطلق** که قبلاً تهیه کرده بودیم، ادغام نموده و دو فهرست کامل تهیه نمودیم، برای حوزه‌ی علوم انسانی «۴۱۵ واژه» و برای حوزه‌ی دامپزشکی «۶۹۶ واژه» تهیه شد.

بر اساس آستانه‌ی ۳۰۰ تواتر، ما دو فهرست واژه (یکی برای حوزه‌ی دامپزشکی «۱۵۹ واژه» و دیگری برای حوزه‌ی علوم انسانی «۲۴۱ واژه») تهیه نمودیم.

سپس این فهرست‌ها را همراه با فهرست واژه‌های تهی مطلق (جمعاً در ۷ فهرست واژه) رایت شده در یک CD به دفتر معاونت محترم پژوهشی مرکز منطقه‌ای اطلاع رسانی علوم و فناوری تحویل گردید.

۴-۲- موارد استفاده‌ی طرح

این طرح پس از اجرا می‌تواند مورد استفادهٔ تمامی مراکز اطلاع رسانی قرارگیرد.

استفاده از نتایج این پژوهش می‌تواند برای تهیه ریشه‌یاب‌های فارسی، نمایه سازی دستی و نمایه سازی خودکار و همچنین در تحلیل خودکار کلمات یک متن، مورد استفاده قرارگیرد.

بکارگیری واژه‌های غیرموضوعی، باعث صرفه‌جویی در زمان و همچنین باعث کاهش حجم فایل‌ها و بهبود دقت در بازیابی اطلاعات میگردد. و همچنین می‌توان از این طریق به تشخیص صحیح کلید واژه‌ها رسید و در استخراج خودکارِ کلید واژه‌های اسناد از آن استفاده نمود.

فصل پنجم

۵-بحث و نتیجه گیری

پس از استخراج داده‌ها، ابتدا واژه‌های بالاتر از آستانه را به عنوان واژه‌های غیرموضوعی در هر دو حوزه استخراج نمودیم. در حوزه‌ی علوم انسانی ۲۴۱ واژه و در حوزه‌ی دامپزشکی ۱۵۹ واژه می‌باشند، این سیاهه‌ها دارای اشکالاتی می باشند که از آن جمله می توان به موارد زیر اشاره نمود. **یکی** اینکه در این فهرست تعدادی واژه‌های موضوعی "از جمله، بیماری، ماهی و ..." مشاهده می‌شود، **دوم** اینکه کلمات به صورت جمع "مانند: مطالعات و اثرات" مشاهده می‌شوند. **سوم** اینکه پسوندها و پیشوندها "مانند: می، ها و...." مشاهده می‌شوند.

در ادامه فهرستی را که از تجزیه تحلیل بسامدی واژه‌های مقالات حوزۀ دامپزشکی بدست آمده‌اند را با حذف تکراری‌ها و یک بازبینی کلی، تهیه نمودیم، این فهرست یک سیاهۀ کامل از واژه‌های غیر موضوعی رشتۀ دامپزشکی را تشکیل می‌دهد. سپس فهرستی را که از تجزیه تحلیل بسامدی واژه‌های مقالات ادب فارسی بدست آمده‌اند را با حذف تکراری‌ها و یک بازبینی کلی، تهیه نمودیم، این فهرست نیز یک سیاهۀ کامل از واژه‌های غیرموضوعی رشتۀ ادبیات فارسی را به دست داده است. این فهرست‌ها در سه مرحله مورد بررسی تخصصی قرار گرفت و در نهایت از ۱۲۰۳۱ واژه‌ی دامپزشکی، ۵۴۰ واژه‌ی غیر موضوعی و از ۴۳۸۸۲ واژه‌ی علوم انسانی ۲۳۸ واژه‌ی غیرموضوعی بدون احتساب کلیه‌ی صیغه‌های صرف شده‌ی افعال، افعال معین، پسوندها و پیشوندهای منفصل و ... (که در صفحات قبلی ذکر شده‌اند) استخراج گردید.

نتایج بدست آمده از این بررسی نشان می‌دهد که، قیدها (قید مختص، قید مشترک، قید مرکب، قید زمان، قید مکان، قید مقدار، قید چگونگی، قید حالت، قید تمنا، قید تعجب، قید تکرار، قید ترتیب، قید تصدیق و ...)، افعال (افعال ردیفی، افعال اسنادی، افعال کمکی، افعال شبه کمکی و ...)، صفات (صفت پرسشی، صفت تعجبی، صفت مفعولی، صفت مفعولی، صفت لیاقت، صفت اشاره، صفت ساده، صفت شمارشی و ...)، ضمایر (ضمایر شخصی، ضمایر مشترک، ضمایر پرسشی و ...)، حروف (حرف ربط،

حرف اضافه و ...) در حوزۀ دامپزشکی جزء واژه‌های غیر موضوعی محسوب می‌شوند اما در حوزۀ ادب فارسی گاهی جزء واژه‌های موضوعی محسوب می‌گردند. بر این اساس تعداد واژه‌های غیر موضوعی در حوزۀ دامپزشکی به مراتب بیشتر از علوم انسانی (ادبیات فارسی) می‌باشد و علت این تفاوت را در ساختار تحقیق وپژوهش‌های ادبی می‌توان جستجو کرد، زیرا واژه‌هایی از جنس حروف اضافه و نقش‌نما، مانند"را" در ادبیات وجود دارند که گاهی جزء واژه‌های موضوعی محسوب می‌شوند. مانند **مقالۀ**: "بابا سالار، اصغر. (۱۳۹۲). کاربردهای خاصّ «را» در برخی متون فارسی. مجلۀ ادب فارسی، دوره ۳، شماره ۱، بهار و تابستان ۱۳۹۲، دانشگاه تهران."

بنابراین نتایج حاصله وابستگی تامِ فهرست‌ها به حوزه‌ی موضوعی را نشان می‌دهند. زیرا در بعضی موارد بعضی از کلمات که به عنوانِ واژه‌های غیرموضوعی ذکر می‌شوند، ممکن است ابتدا جزو واژه‌های موضوعی به‌نظر برسند، اما چون این فهرست وابستگیِ تام به حوزه‌ی موضوعی دارد، لذا در یک حوزه‌ی خاص جزو واژه‌های غیرموضوعی تلقی می‌گردد. و برعکس تعدادی از واژه‌ها که به عنوانِ واژه‌های غیرموضوعی در بعضی از فهرست‌ها مشاهده می‌شوند، ممکن است در فهرست حوزه‌ی دیگر، بدلیل وابستگی به حوزه‌ی موضوعی خاص، به عنوانِ واژه‌ی‌موضوعی تلقی گردند.

لازم به توضیح است که در زبان فارسی واژه‌های غیرموضوعیِ استانداردی وجود ندارد.

نتیجه‌ی این پژوهش نشان داد که واژه‌های غیرموضوعی حوزه‌های مختلف متفاوت هستند.

۵-۱- خروجی‌ها و فهرست‌های تهیه شده

بطور کلی در این پژوهش چهار نوع فهرست واژه‌های غیر موضوعی تهیه گردید،

- یکی واژه‌های غیر موضوعی مطلق که از حروف اضافه، حروف ربط و قیدها و تعدادی از صفات تشکیل شده است، اینگونه واژه‌ها همیشه بعنوان غیر موضوعی بکار می‌روند.

- دیگری واژه‌هایی هستند که در مجموعه‌های اسناد و مدارک و بر اساس پیکره‌ی زبانی توسط متخصصان موضوعی در حوزه‌های خاصی تهیه می‌شوند

- سومی فهرستی است که با استفاده از آستانه‌ی تواتر واژه‌ها تهیه گردیده است

- چهارمی، فهرستی است که با استفاده از هر دو فهرست قبلی (سیاهه‌ی تهیه شده بر اساس نظر متخصصان موضوعی و سیاهه‌ی واژه‌های تهیِ مطلق) تهیه گردید که در اینجا ما آن را **فهرست مشترک** می‌نامیم.

بنابراین بنظر می‌رسد تهیه یک فهرست ثابت از آن‌ها جهت هر زبانی بویژه زبان فارسی لازم است. این **فهرست مشترک** می‌تواند همواره در اسناد و پیکره‌های مختلف هر زبان متفاوت باشد.

این واژه‌ها در پردازش بازیابی اطلاعات می‌توانند دو اثر متفاوت داشته باشند. یکی اینکه بدلیل دارا بودن بسامد، حذف آن‌ها کارایی بازیابی و نمایه سازی را افزایش می‌دهد زیرا باعث کاسته شدن فضای اشغال در حافظه رایانه می‌شوند، از سوی دیگر، با افزودن واژه‌ها و عبارت‌های تهی و بسط جستجو، کاربران می‌توانند به نتایج کمتر اما دقیقتر و مطلوبتری برسند (فتاحی ۱۳۸۵). این تناقض نشانگر آن است که واژه‌های غیر موضوعی بر دو نوع هستند یکی واژه‌های **مستقل** دیگری واژه‌های **وابسته** ،بدین معنی که تعدادی از این واژه‌ها بطور مستقل غیر موضوعی هستند و در هر حالت بعنوان واژه‌های غیر موضوعی محسوب می‌شوند، مانند: (در، از، و ...). اما واژه‌های وابسته‌ی غیر موضوعی، واژه هایی هستند که فی نفسه غیر موضوعی هستند ولی چنانچه با یک واژه موضوعی مناسب ظاهر شوند، بعنوان یک واژه کمکی در عبارت موضوعی بکار می‌روند، مانند (مقدمه ای بر ...) که اینگونه واژه‌ها را دکتر فتاحی «نیمه موضوعی» می‌نامد. جهت کار برد مناسب واژه‌های غیر موضوعی بهتر است که، اولاً؛ هر گروه از واژه‌ها در جای مناسب بکار روند، درثانی؛ ترکیبی مشترک از هر دو لیست تهیه و بعنوان یک لیست استاندارد (برای یک پیکره مشخص در یک حوزه‌ی معُین) بکار برده شود.

همانگونه که قبلاً نیز ذکر شد باید توجه داشته باشیم که تهیه‌ی چنین فهرستی اولاً مختص یک حوزه و بسیار محدود بوده و حتی قابل تسری به سایر پیکره‌های همان حوزه نیز نخواهد بود و کاربردش فقط می‌تواند در همان پیکره توجیه پذیر باشد، ثانیاً برای تهیه‌ی آن می‌بایست حتماً دو مرحله‌ی فوق انجام پذیرد (تهیه‌ی سیاهه‌ی مطلق و تهیه‌ی سیاهه توسط متخصص موضوعی).

با توجه به بررسی‌های انجام گرفته، در نهایت می‌توان چنین نتیجه گرفت که:

۱-تعداد واژه‌های تهی، در هر زبان به مراتب بیشتر از آن تعدادی است که تاکنون در سایت‌های Big data ، اکادمی داده و توسط فاکس و سنجی و داورپناه تهیه و ارائه شده‌اند.

۲- تعداد اینگونه واژه‌ها با توجه به پویایی زبان، همواره در حالِ تغییر و افزایش می‌باشند.

۳- تعدادِ اینگونه واژه‌ها به تعدادِ واژه‌هایی بستگی دارد که در فرهنگها (دیکشنری‌ها) فهرست نمی‌شوند. به عبارت دیگر، در هر زبان تعداد واژه‌های تهی به اندازه‌ی تعداد واژه‌هایی هستند که بصورت مدخل در واژه نامه‌ها قرار نمی‌گیرند.

۴- نمی‌توان در هیچ زبانی، سیاهه‌ی واژه‌های تهی را تهیه نمود.

۵-۲-تعداد واژه‌های غیر موضوعی

به نظر می‌رسد تعداد این واژه‌ها براساس تعاریف ارائه شده، همواره بیشتر از آن تعدادی هستند که بتوان در یک سیاهه جمع‌آوری کرد، به‌عنوان مثال اگر تعداد افعال بسیط را ۳۰۰ فعل در نظر بگیریم و در هشت زمان و شش صیغه صرف شوند، حدوداً ۱۴۴۰۰ واژه غیر موضوعی خواهیم داشت، حال اگر این تعداد را با افعال مرکب در نظر بگیریم بی‌شک این تعداد فقط در مورد افعال، عدد قابل توجهی خواهد بود. و اگر به این تعداد قیود، صفت‌ها، افعال معین، ضمایر، اصوات، و ... را نیز بیافزاییم، خود سیاهه‌ای به اندازه‌ی یک فرهنگ را شامل خواهد شد.

حال، با احتساب موارد مذکور، تعداد واژه‌های غیر موضوعی بسیار بیشتر از مواردی است که تا به امروز توسط متخصصان اطلاع رسانی ذکر شده است.

۵-۳- کاربرد نتایج

واژه‌های غیرموضوعی زمینه‌ساز تسهیل در بازیابی صحیح اطلاعات می‌باشد، بازیابی صحیح و بی نقص اطلاعات نیازمند شناخت واژه‌های موضوعی است و حلقه‌ی مفقود در توسعه و گسترش ذخیره سازی و بازیابی اطلاعاتِ رایانه‌ای همانا مستلزم شناختی درست و صحیح از واژه‌های غیرموضوعی است.

گاهی مقولهٔ واژه‌های غیر موضوعی به عنوان یک روش منفی در مقالات طبیعی گفتار، تلقی می شود. واژه‌های غیرموضوعی سیاهه‌ای از واژه‌ها (حروف اضافه، ضمایر و غیره) در فهرستی از اصطلاحات هستند که ارزش معنایی اندکی دارند و وقتی که در مدرکی یافت شوند، دور از واژه‌های موضوعی نگهداشته شده و نادیده گرفته می‌شوند همچنین می‌توان گفت، تقریباً تمام کاربردهای بازیابی اطلاعات پیش از پردازش اسناد و پرس‌وجوها، واژه‌های غیرموضوعی را نادیده می‌گیرند. این عمل معمولاً کارکرد نظام را افزایش می‌دهد (مهراد و ناصری۱۳۸۷). بطور کلی حذف واژه‌های غیرموضوعی اندازه‌ی اسناد را عوض کرده و بر وزن و ارزش پردازش آن‌ها تأثیر می‌گذارد، همچنین این عمل می‌تواند در کارایی شایسته واژه‌ها و ماهیت آن‌ها و این واقعیت که این واژه‌ها حامل معنی نیستند و تنها حامل نقش نحوی و صورت‌ها و مقولات دستوری هستند تأثیر گذار باشد. همچنین، حذف واژه‌های غیرموضوعی می‌تواند موجب افزایش کارایی پردازش از حدوداً ۳۰ ٪ به ۵۰ ٪ گردد. نمونهٔ آن در مجموعه‌ی یک متن طولانی بهتر نمایش داده می‌شود. در ضمن، در تمامی روش‌های متن کاوی و به طبع آن در روش‌های طبقه بندی متون انجام مراحل آماده سازی متن اجتناب ناپذیر است. و برای انجام مراحل آماده سازی متن،

حذفِ ایست واژه‌ها الزامی است. با انجام این کار از بار اجرای الگوریتم به مقدار زیادی کاسته می‌شود.

مهراد و فلاحتی (۱۳۸۴: ۱۱) معتقدند بیشتر سامانه‌های بازیابی اطلاعات از فایل معکوس یا فهرستی از واژگان که به صورت الفبایی مرتب شده‌اند استفاده می‌نمایند و واژه‌های غیرموضوعی از این فهرست کنار گذاشته می‌شوند.

در نهایت شناخت یک فهرست واژه‌های غیرموضوعی جهت حذف آن‌ها و پردازش متن در یک سامانه بازیابی اطلاعات ضروری است. علاوه بر این، می‌توان از نتایج این طرح در تهیه‌ی فهرست مارکوفِ معنا دار برای زبان، نمایه سازی ماشینی، خلاصه سازی خودکار و ماشین ترجمه استفاده نمود.

۵-۴-پیشنهادات

در انتها پیشنهاد می‌شود که این تحقیق در مورد سایر حوزه‌ها از جمله، مهندسی، علوم و ... نیز انجام پذیرد و در نهایت با یکدیگر مقایسه گردند.

منابع :

۱- انوری، حسن؛ احمدی گیوی، حسن، (۱۳۸۵) "دستور زبان فارسی ۲"، انتشارات موسسه فرهنگی فاطمی، تهران، ۱۳۸۵

۲- باطنی، محمدرضا، (۱۳۵۶) "توصیف ساختمان دستوری زبان فارسی" موسسه انتشارات امیر کبیر، تهران، ۱۳۵۶

۳- بلندیان، صدیقه "۱۳۸۵" تحلیل متن مقالات فارسی کتابداری و اطلاع رسانی و امکان نمایه سازی ماسینی آنها بر اساس قانون زیف، پایان نامه کارشناسی ارشد، دانشگاه فردوسی مشهد.

۴-حری، عباس (۱۳۷۳) "مسائل و مشکلات ذخیره پیش همارا و بازیابی پس همارا در نظام کامپیوتری" فصلنامه کتاب، پائیز و زمستان ۱۳۷۳

۵-داور پناه، محمد رضا؛ بلندیان، صدیقه، (۱۳۸۶) تحلیل متن مقالات فارسی و امکان نمایه سازی ماشینی آن‌ها براساس قانون زیف، فصلنامه پژوهش در مسائل تعلیم و تربیت ویژه نامه کتابداری و اطلاع رسانی دور دوم.

۶- سنجی، مجیده؛ داور پناه، محمدرضا (۱۳۸۸)، شناسایی واژه‌های غیر مفهومی در نمایه سازی خودکار مدارک فارسی، فصلنامه کتابداری و اطلاع رسانی، جلد ۱۲، شماره ۴،

۷-صفوی، کورش، (۱۳۸۷) "درآمدی بر معنی شناسی" پژوهشگاه فرهنگ و هنر اسلامی، انتشارات سوره مهر ، تهران.

۸- فتاحی، رحمت الله، (۱۳۸۵) "شناسایی و تحلیل واژگان عمومی در منابع وب: رویکردی نو به بسط عبارت جستجو با استفاده از زبان طبیعی در موتورهای کاوش "مطالعات تربیت و روانشناسی دانشگاه فردوسی مشهد، دوره ۷ شماره ۱ سال ۱۳۸۵.

۹-فرشیدور، خسرو، (۱۳۸۲) "دستورمفصل امروز" انتشارات سخن ، تهران،

۱۰- مشکوه الدینی، مهدی، "دستورزبان فارسی برپایه ی نظریه ی گشتاری" انتشارات دانشگاه فردوسی مشهد، ۱۳۷۹

۱۱- مهراد، جعفر؛ مریم، ناصری، (۱۳۸۷)"پردازش زبان طبیعی وبازیابی اطلاعات" مرکز منطقه‌ای اطلاع رسانی علوم وفناوری، نشر چاپار، شیراز، ۱۳۸۷

۱۲- مهراد، جعفر؛ فلاحتی قدیمی فومنی، محمد رضا " ۱۳۸۴ " معنا شناسی و بازیابی اطلاعات، مشهد، انتشارات کتابخانه رایانه ای ، شیراز، کتابخانه منطقه ای علوم و تکنولوژی .

۱۳- ویگری، برایان؛ ویکری، الینا (۱۳۸۰) "علم اطلاع رسانی در نظر و عمل،" ترجمه: عبدالحسین فرج پهلو، انتشارات دانشگاه فردوسی مشهد.

۱۴- هاشم زاده، محمد جواد؛ نخعی، زینب؛ مرادی مقدم، حسین "۱۳۹۲" کاربرد و تعدیل قانون زیف و الگوی آماری زو در بازشناسی واژه‌های باز دارنده زبان فارسی با استفاده از خوشه زبانی مقالات علمی پژوهشی رشته کتابداری و اطلاع رسانی، پژوهشنامه کتابداری و اطلاع رسانی دانشکده علوم تربیتی و روانشناسی دانشگاه فردوسی مشهد.

14-Abu El-Khair,Ibrahim “Effects of stop words Elimination for Arabic Information Retrival : A Comparative Study “ International Journal Of computing and Information Sciences ,vol 4 ,no3, December 2006

15-Fox, C. (1990). *A stop list for general text*. Retrieved November 20, 2010, from http://www.informatik.uni-trier.de/ley/indice/a-tree.pdf

16- http : // www. Open source . org / Licenceses / bsd License /.

17-Lazar , Gilbert. “Grammaire Du Persan Contemporan” Paris Librairie C.Klincksieck , 1957 .

18- Ricardo , B.Y , Berthier R.N , (1999) “ Moderne Information Retrival “ Addison Wesley LongmanPublishing Boston .

19-Reghavan. V . V ,Wong S.K.M(1986) “A Critical Analysis aof Vector Space Model For Information Retrival “Journal Of The American Society For Information Science .

20- Savoy , Jacques (1999) “http : // members . unine . ch / Jacques , Savoy / Clef Persian ST .txt / .

21- Savoy, J .Rassolofo F (2003). A stemming procedure and stopword list for general French corpora. Journal of the American Society for Information Science, 50, , 944-952.

22--Taghva, K., Beckley, R.,&Sadeh, M. (2003a). A list of Farsi stopwords.Retrieved January 15, 2009, from http://www.isri.unlv.edu/publications/isripub/taghva2003-01.ps

23- Zou Feng , Wang , F .Deng , x (2005) “Evaluation Of Stop Word List In Chinese Language”

24- Zou Feng , Wang , F .Deng , x . (2006) “ Automatic Indentification Of Chinese Stop Word “A Special Issue On Advaces In N.L.P of the Journal Research On Computing Science .