
بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده مهندسی کامپیوتر

پایان‌نامه کارشناسی

گرایش مهندسی نرم‌افزار

عنوان

پیشنهاد و ارزیابی یک معیار برای سنجش نزدیکی وبلاگ‌های فارسی براساس محتوا

نگارش

آیدین نصیری‌شرق

استاد راهنما

دکتر حسن ابوالحسنی

دی‌ماه ۱۳۸۸

دانشگاه صنعتی شریف

دانشکده مهندسی کامپیوتر

قبول شده به عنوان قسمتی از ملزومات درجه کارشناسی

مهندسی کامپیوتر نرم افزار

دانشگاه صنعتی شریف

عنوان پایان نامه: پیشنهاد و ارزیابی یک معیار برای سنجش نزدیکی وبلاگ‌های فارسی بر اساس محتوا

تهیه کننده: آیدین نصیری‌شرق

استاد راهنما: دکتر حسن ابوالحسنی

سال اجرا: ۱۳۸۸

نام استاد پروژه: **محل امضاء:** **مورخ:**

نمره پروژه: **به عدد:** **به حروف:**

تقدیم به:

پدر بزرگوارم، که با بردباری و حمایت های بی دریغش، همواره رهگشای من بود

مادر فداکارم، که با دانش و درایت دلسوزانه اش، همواره رهنمای من بود

و **خواهر عزیزم آیدا**، به امید آن که انگیزه ای برای ادامه ی راهش باشد.

تقدیر و تشکر

بر خود لازم می‌دانم از زحمات استاد عزیزم جناب آقای دکتر حسن ابوالحسنی نهایت تشکر و قدردانی را به‌عمل آورم. حسن نظر و درایت ایشان از همان ابتدا در انتخاب عنوان این پایان‌نامه، راهنمایی‌های ارزنده شان در اجرای مراحل مختلف، و نهایتاً رهنمودهای مدبرانه‌ی ایشان در جمع‌بندی نتایج همچون چراغی روشن‌گر در تمام مسیر بر سر راه من بوده است.

نیز لازم است از جناب آقای دکتر صفری که در تمام مراحل اجرای این پایان‌نامه دلسوزانه از من حمایت کردند و در دشواری‌های راه، انگیزه‌ی من برای ادامه‌ی کار را تقویت کردند، صمیمانه تشکر کنم.

از جناب آقای مهندس سجاد شیرعلی شهرضا نیز که با ارائه‌ی ایده‌های ظریف در راه گسترش نتایج و هم‌چنین معرفی منابع، حامی من در اجرای این پایان‌نامه بودند، سپاسگذاری می‌کنم. نکات ارزنده‌ای که ایشان در بازبینی از این پایان‌نامه و پیش از نهایی‌شدن آن ارائه داشتند، تأثیر به‌سزایی در کیفیت نهایی این پایان‌نامه داشته است.

آقایان روزبه پورنادر ، بهنام پورنادر و حامد ملک از شرکت فارسی‌وب شریف نیز شایسته‌ی قدردانی هستند. داده‌های علمی مربوط به «واژگان پرکاربرد زبان فارسی» و تجربیات کار با زبان فارسی از دست‌آوردهای دو دوره‌ی کارآموزی این‌جانب در آن شرکت و تحت حمایت‌های صمیمانه‌ی این عزیزان به‌دست آمده است.

در پایان از ۳ موجود غیرانسانی به‌نام‌های «بستار اینترنت»، «وب‌سایت گوگل» و «وب‌سایت ویکی‌پدیا» نیز تشکّر می‌کنم. وقتی پروسه‌ی تهیه‌ی این پایان‌نامه را با پروسه‌ی پایان‌نامه‌ی دکترای مادرم (قریب ۱۰ سال پیش) مقایسه می‌کنم، می‌بینم به‌حق این ۳ عزیز کمک غیر قابل وصف‌ای در بالابردن کارایی، سرعت و دقّت کارم داشته‌اند.

چکیده

وبلاگ‌ها از جدیدترین پدیده‌های افزایش ارتباطات و تولید محتوا در دنیای اینترنت می‌باشند. تاحدی که برخی از صاحب نظران عمده محتوای تولیدشده در وب ۲٫۰ را توسط کاربران عادی و از طریق وبلاگ‌ها می‌دانند ([صابری، ۸۶] و [Oreilly 05]).

غالب تحلیل‌های انجام شده بر روی وبلاگ‌ها، در شاخه‌های مهندسی، براساس پیوندها و ساختار آن‌ها است؛ چرا که وبلاگ‌ها یک بستر تحلیلی و مدل آماری مناسب برای تحلیل به‌صورت یک گراف یا شبکه‌ی اجتماعی می‌باشند ([جمالی، ۸۶]). بررسی‌های محتوایی وبلاگ‌ها نیز عمدتاً در شاخه‌های علوم انسانی و برپایه‌ی نمونه‌برداری تک‌به‌تک و تحلیل انسانی شکل گرفته‌است (نظیر [ضیایی‌پرور، ۸۶] در وبلاگ‌های فارسی و [Hurring 05] در وبلاگ‌های انگلیسی).

عمده مشکل بررسی محتوایی وبلاگ‌های فارسی به ساختار زبان فارسی و تفاوت نگارش‌ها باز می‌گردد. مشکلات ناشی از تفاوت گفتار عامیانه با نوشتار و سلیقه‌ای بودن رسم‌الخط از یک سو و مشکلات ساختاری زبان فارسی (نظیر املای یک‌سان شُکر و شِکَر) از سوی دیگر باعث شده تا تحلیل محتوایی خودکار وبلاگ‌های فارسی زبان، موضوعی باشد که کمتر به آن پرداخته شده است.

در این پروژه با بررسی خودکار محتوایی وبلاگ‌ها، صرفاً به‌عنوان گنجینه‌ی واژگان مستقل، معیاری برای بیان شباهت و نزدیکی دو وبلاگ ارائه خواهد شد. سپس به‌عنوان یک ارزیابی از موفقیت معیار، با استفاده از یک محک شامل قریب ۹۰۰۰ وبلاگ که در همین پروژه تهیه شده است، عمل‌کرد این معیار روی دسته‌بندی وبلاگ‌ها بر حسب یک فهرست موضوعی، سنجیده خواهد شد. در نهایت با ذکر چالش‌های در پیش رو، پیشنهاداتی برای کارهای آتی ارائه خواهیم داد.

واژگان کلیدی

وبلاگ، وبلاگ‌های فارسی، محتوای وبلاگ، بررسی محتوا، نزدیکی وبلاگ‌ها، دسته‌بندی وبلاگ‌ها

فهرست مطالب

فهرست جداول

فهرست اشکال

۱ فصل یکم)

آشنایی با وبلاگ‌های فارسی

بخش ۱ این فصل ابتدا به معرفی وبلاگ و ترجمه‌ی لغوی آن اختصاص دارد. در ادامه با ذکر تاریخچه‌ی مختصری از وبلاگ و وبلاگ‌نویسی، دامنه‌ی نویسندگان وبلاگ‌ها و اهداف آن‌ها پویش می‌شود. در بخش ۲ با گذر به دنیای وبلاگ‌های فارسی و معرفی آن‌ها، ویژگی‌های خاص وبلاگ‌های فارسی مطرح خواهد شد. در بخش ۳ با معرفی ویژگی‌های خط و زبان فارسی و اشاره‌ای به دشواری‌های پردازش متون فارسی، چشم‌اندازی از مراحل آتی ترسیم خواهیم شد. نهایتاً در بخش ۴ یک جمع‌بندی از این فصل ارائه خواهد شد.

۱.۱ معرفی وبلاگ

با پیدایش وب ۲٫۰، عملاً مشارکت مخاطبان در دنیای اینترنت سهم عمده‌ی تولید محتوا را به خود اختصاص داده است. یکی از ساده‌ترین و رایج‌ترین ابزار مشارکتی کاربران در اینترنت وبلاگ‌ها هستند. در تعریف وبلاگ (که معادل فارسی «وب‌نوشت» برای آن پیش‌نهاد شده است) در ویکی‌پدیای فارسی[1] می‌خوانیم:

> *وب‌نوشت، وبلاگ یا بلاگ یک وب‌گاه یا صفحاتی از یک وب‌گاه اینترنتی متشکل از مطالب کوتاه به نام پست (Post) با محتوای اندیشه، اطلاعات و خاطرات روزانه شخصی و پیوندهایی است که به ترتیب زمانی از جدید به قدیم قرار گرفته‌اند. نویسنده وبلاگ ممکن است یک یا چند نفر باشد.*
>
> *واحد مطالب در وبلاگ، «پست» است، در حالی که واحد مطالب در وب‌گاه صفحه (Page) است. معمولاً در انتهای هر مطلب، برچسب تاریخ و زمان، نام نویسنده و پیوند ثابت به آن یادداشت ثبت می‌شود. فاصلهٔ زمانی بین مطالب وب‌نوشت لزوماً یکسان نیست و زمان نوشته‌شدن هر مطلب به خواست نویسندهٔ وب‌نوشت بستگی دارد. مطالب نوشته شده در یک وب‌نوشت همانند محتویات یک وب‌گاه معمولی در دسترس کاربران قرار می‌گیرد. در بیشتر موارد وب‌نوشت‌ها دارای روشی برای دسترسی به بایگانی یادداشت‌ها هستند (مثلاً دسترسی*

[1] http://fa.wikipedia.org/wiki/وب‌نوشت

به بایگانی بر حسب تاریخ یا موضوع). بعضی از وب‌نوشت‌ها امکان جستجو برای یک واژه یا عبارت خاص را در میان مطالب به کاربر می‌دهند.

با این تعریف، می‌توان به‌صورت خلاصه وبلاگ را چنین معرفی کرد:

تعریف ساده‌ی وبلاگ:

فضایی روی وب (WWW) که نویسنده یا نویسندگانی با ثبت یادداشت‌هایی، آن را مرتباً به‌روز می‌کنند و غالباً در صفحه‌ی نخست مطالب با نظم تاریخی معکوس قرار دارند. وبلاگ عموماً دارای بایگانی و بخش نظرخواهی از کاربران (مجزا برای هر یادداشت) است.

واژه‌ی «Weblog» (وبلاگ) اولین‌بار در ماه دسامبر سال ۱۹۹۷ توسط Jorn Barger استفاده شده است. این واژه از ترکیب دو لفظ Web (صورت مختصر World Wide Web) و Log، از ریشه‌ی Logos یونانی، به‌معنی ثبت‌کردن پدیدآمده‌است. از این رو به‌صورت لغوی می‌توان وبلاگ را «واقعه‌نگاری روی وب» ترجمه کرد. صورت ساده‌ی این واژه که blog (بلاگ) است نیز در سال ۱۹۹۹ توسط Peter Merholz مصطلح گردید.

۱.۱.۱ تاریخچه‌ی وبلاگ‌ها

با توجه به جامعیت و فراگیر بودن دنیای وب، نمی‌توان بنیان‌گذار اوّلیه‌ی مشخصی را برای وبلاگ‌ها درنظر گرفت. آن‌چه مسلم است، ساختار کنونی وبلاگ نشأت گرفته از usenetهای قدیمی یا فروم‌ها و BBSهای اواخر دهه‌ی ۸۰ میادی‌اند. با این حال، برخی گزارش‌ها، Justin Hall (از دانشجویان کالج Swarthmore)، برخی Dave Winer (با گزارشات Scripting News) و برخی حتی Tim Berners-Lee (مخترع وب) را اوّلین نویسندگان وب در دنیا می‌دانند. وجه اشتراک این وبلاگ‌های اولیه در ساختار آن‌ها بوده‌است. این ساختار مجموعه‌ی منظمی از لینک‌ها و توضیحات متناسب با سلیقه‌ی نویسنده بوده که در بازه‌های زمانی به‌روز می‌شده‌است.

تا سال ۱۹۹۹ نرخ رشد وبلاگ‌ها بسیار کم بود با راه اندازی چندین سرویس رایگان و یا ارزان قیمت دررابطه با وبلاگ و وبلاگ نویسی ، تعداد آنان رو به افزایش نهاد . سرویس هائی نظیر Pitas, Livejournal، Blogger و EditThisPage.com ، نمونه هائی در این زمینه می باشند . استفاده از سرویس های

فوق،مستلزم دانش فنی بالائی نبود و علاقه مندان به ایجاد وبلاگ می توانستند به سرعت اقدام به ایجاد وبلاگ مورد نظر خود نمایند. تا اواسط سال ۲۰۰۰ ، بیش از یکهزار وبلاگ ایجاد و این رقم تا اواسط سال ۲۰۰۲ به بیش از نیم میلیون رسید [بهبهانی، ۸۳] [کرمی، ۸۴]. در سال ۲۰۰۵، واژه‌ی blog و مشتقات آن به واژه‌نامه‌ی آکسفورد راه یافتند[۲]. اکنون پس از گذشت قریب ۱۰ سال از آغاز شیوع وبلاگ‌نویسی عمومی، به گزارش سایت thefuturebuzz.com[۳] تا سال ۲۰۰۹ بیش از ۱۳۳ میلیون وبلاگ توسط سایت Technorati.com اندیس‌گزاری شده‌اند و در ماه مارچ سال ۲۰۰۸ حدود ۳۴۶ میلیون نفر وبلاگ به ۸۱ زبان مختلف می‌خوانده‌اند.

۱.۱.۲ محتوای وبلاگ‌ها و هدف از ایجاد وبلاگ

آن‌چه مسلّم است وبلاگ به‌خودی‌خود یک قالب محتوای دانش است که هر کاربر اینترنت می‌تواند به‌سادگی و بدون پرداخت هیچ هزینه‌ای اقدام به ساختن یک وبلاگ برای خود، و اشاعه‌ی دانش و سلیقه‌ی خود بکند. این سادگی و آزادی عمل باعث شده تا محتوای وبلاگ‌ها طیف وسیعی را در بر گیرد.

[بهبهانی، ۸۳] مزایا و دستاوردهای ایجاد وبلاگ را در ۵ دسته به‌صورت زیر طبقه‌بندی می‌کند[۴]:

۱. ***انتخاب مطلب (داده):*** *میزان تولید و نشر اطلاعات در سطح جهان به سرعت در حال پیشرفت بوده و روندی کاملاً تصاعدی را طی می‌نماید. بدیهی است در چنین وضعیتی، حتی‌امکان مطالعه بخش اندکی از آنان نیز وجود نداشته و ما مستلزم استفاده از روش ها و مکانیزم‌هایی به منظور فیلترینگ اطلاعات و یافتن اطلاعات مورد نظر در یک رابطه خاص بدون از دست دادن منبع ارزشمند و محدود زمان می باشیم. وبلاگ ها با تمرکز بر روی یک موضوع خاص می توانند بستری مناسب برای ارائه اطلاعات را فراهم نمایند. با مطالعه و خواندن مطالب منتشر شده بر روی یک وبلاگ که توسط فردی با علایق مشترک با شما تهیه و منشتر شده است، احتمال یافتن مطالب مورد نظر در زمانی معقول فراهم می گردد. با ترکیب و جمع بندی مطالب منتشر شده در*

---

[2] http://www.askoxford.com/worldofwords/bubblingunder/archive/bubbling_03/

[3] http://thefuturebuzz.com/2009/01/12/social-media-web-20-internet-numbers-stats/

[۴] علی‌رغم این‌که ممکن‌ست تصور شود، این دسته‌بندی بیش‌تر جنبه‌ی اجتماعی داشته و خارج از مباحث فنّی این پایان‌نامه است، اما تصور نگارنده این‌ست که دقّت در این اهداف می‌تواند نویسندگان وبلاگ‌ها - خصوصاً وبلاگ‌های فارسی - را بهتر شناسایی کرده و در مراحل فنی - خصوصاً مربوط به زبان و نگارش فارسی - با شناخت مخاطب به درک بهتر دامنه‌ی مسئله کمک شایانی بکند.

ارتباط با یک موضوع خاص از چندین وبلاگ انتخابی، می‌توان به مجموعه ای از اطلاعات مورد علاقه، دست یافت.

۲. ***مدیریت دانش و تجارب شخصی:*** *محتویات وبلاگ به منزله یک بایگانی از افکار و اندیشه های وبلاگ نویسان آن بوده که در مقاطع زمانی متفاوتی نوشته شده و در صورت نیاز به اطلاعاتی خاص می‌توان با استفاده از مراکز جستجو و بر اساس یک کلید واژه خاص به آنان مراجعه نمود. وجود لینک‌های متعدد مرتبط با یک موضوع خاص که توسط مؤلف یک وبلاگ مشخص می‌گردد، امکان دنبال نمودن وضعیت موجود در رابطه با یک موضوع خاص را در اختیآر علاقه‌مندان قرار می‌دهد .*

۳. ***ارتباط دو سویه:*** *همانگونه که قابل پیش‌بینی بود، وبلاگ‌ها به محیط و یا رسانه محاوره ای برای مباحث عمومی و تخصصی تبدیل و امکان تعامل اطلاعاتی بین وبلاگ نویسان و خوانندگان از یکطرف و خوانندگان با خوانندگان از طرف دیگر فراهم می گردد. ویژگی فوق از ماهیت دوطرفه بودن وب به نحو احسن استفاده و آن را در جهت اهداف خود بکار می گیرد.*

۴. ***جامعه شبکه‌ای:*** *پدیده وبلاگ نویسی فرصت‌ها و پتانسیل‌های مناسبی را در جامعه شبکه‌ای، ایجاد می‌نماید. نویسندگان وبلاگ‌ها به مرور زمان توسط خوانندگان خود شناخته خواهند شد. بدین ترتیب آنان در معرض فرصت‌هائی قرار خواهند گرفت که شاید هرگز تصور آن را نمی کردند. در جامعه شبکه‌ای هر شخص می تواند دارای سهمی در تولید و ارائه اطلاعات داشته باشد و خود نیز می تواند از دستاوردهای اطلاعاتی دیگران استفاده نماید. جامعه شبکه‌ای دارای ایستگاه‌هایی (انسان) است که در آن هر یک سهمی در تولید و ارائه اطلاعات و استفاده از اطلاعات دیگران را بر عهده خواهند داشت.*

۵. ***روتینگ اطلاعات:*** *وبلاگ‌ها دارای تاثیری مثبت در خصوص چرخش آزادنه اطلاعات در یک جامعه اطلاعاتی می باشند. خواننده و نویسنده یک وبلاگ اغلب به یک جامعه و یا سازمان یکسان تعلق نداشته و برای ارتباط بین آنان مرز خاصی وجود نخواهد داشت. بدین ترتیب ما شاهد تقابل افکار، اندیشه ها در یک مقیاس گسترده و جهانی بوده که زمنیه یک جامعه اطلاعاتی را ایجاد می نماید. ایجاد چنین روابطی در دنیای خارج از وبلاگ امری مشکل و گاهاً غیر ممکن است.*

وی هم‌چنین ۳ دلیل عمده برای فراگیر شدن وبلاگ‌ها ارائه می‌دهد:

۱. ***عدم وجود انحصار:*** *عملیات یک وبلاگ به یک نرم افزار خاص و انحصاری و یا امکاناتی که صرفا" در اختیار یک سازمان خاص است، وابسته نمی‌باشد. وبلاگ‌نویسان در حوزه عملکرد وبلاگ خود دارای آزادی عمل مناسبی می باشند و در هر زمان می توانند شکل ظاهری ، لی اوت و یا محتویات آن را تغییر داده و یا ویژگی های جدیدی را بدون کسب اجازه یک مقام خاص به آن اضافه نمایند. وبلاگ ها بستر لازم برای خلاقیت انسان را در تمامی زمنیه ها ارائه نموده و هر یک از ساکنین این کره خاکی قادر به شکوفائی خلاقیت خود و آفرینش محصولات متفاوت اطلاعاتی، خواهند بود.*

۲. ***تعداد و تنوع گسترده کاربران:*** *تمامی افرادی که در یک جامعه مدرن اطلاعاتی زندگی می کنند ، تمایل به انتشار تجارب خود ، استفاده از تجارب دیگران ، دریافت بازخورد سریع نسبت به موارد منتشر شده ، ارزیابی نتایج و بهبود دانش و تجارب خود را دارند(فردایمان بهتر از امروز و امروزمان بهتر از دیروز) . وبلاگ ها یک شبکه ارتباطی قدرتمند را ایجاد نموده و پس از طرح یک ایده و یا موضوع جدید در وبلاگ، شاهد حرکت سریع آن در شبکه ارتباطی خواهیم بود. بدین ترتیب درصورت ارائه یک مطلب ارزشمند و صحیح ، امکان استفاده از آن در سریعترین زمان ممکن برای دیگران فراهم شده و در صورتی که مطلب منتشر شده نادرست باشد ، نویسنده آن با دریافت سریع بازخوردهای مورد نظر و بررسی آنان ، می تواند اشتباه خود را در اسرع وقت تصحیح نماید.*

۳. ***ارتباط شغلی وبلاگ نویسان:*** *تعداد زیادی از وبلاگ نویسان خود پیاده کنندگان نرم افزار می باشند. در چینین مواردی آنان به عنوان وبلاگ نویس قادر به تشریح ویژگی های جدید یک محصول با کیفیتی مطلوبتر و با استفاده از فن آورهای متعدد خواهند بود .*

به‌عنوان تکمیل‌کننده‌ای بر این دسته‌بندی‌ها، ۱۰ روش وبلاگ‌نوشتن را که [رسولی، ۸۵] گردآوردی کرده، به‌صورت خلاصه بیان می‌کنیم. این دسته‌بندی، در فصل ارزیابی معیار این پروژه به‌عنوان ملاکی برای طبقه‌بندی و تفکیک وبلاگ‌ها براساس سبک محتوا مورد استفاده قرار می‌گیرد.

۱. ***بلاگ کردن برای ابراز و تصحیح عقاید و خصوصیات شخصی:*** *این شاید یکی از رایج‌ترین شکل‌های استفاده از وبلاگ است. این وبلاگها، شبیه به دفتر خاطرات روزانه یا گزارش روزانه طبیعی است.*

۲. ***بلاگ کردن برای ایجاد ارتباط با دیگر بلاگرها:*** *در اصطلاح، بلاگرها می‌گویند یکی از دلایلی که ما بلاگ کردن را آغاز کردیم این بود که خواستار ارتباط با دوستان و فامیل بودیم و به جای e-mail فرستادن (یا دایره پیچ در پیچ mail) فقط بلاگ می‌کردیم.*

۳. ***بلاگ کردن برای تعامل اجتماعی:*** *صرف نظر از دوستان و آشنایان، مردم نیز می‌توانند بوسیله وبلاگها صاحب یك شبکه ارتباطی- اجتماعی خوب شوند حتی اگر بلاگرها در نقاط مختلف کره زمین و کشورهای گوناگون و با فرهنگ‌های متفاوت باشند، می‌توانند مواردی را که به آنها علاقمند هستند، طراحی و در وبلاگ بگذارند و به طور حتم، افراد بسیاری وجود دارند که به موضوع مطرح شده علاقه داشته باشند.*

۴. ***بلاگ کردن به منظور تحصیل:*** *در داخل این چتر حجیم تحصیلی، مردم می‌توانند از وبلاگها برای روشهای آموزشی نیز استفاده کنند همچنین مربی و معلم می‌توانند در پروژه‌های مشترك با دانشجویان و دانش‌آموزان همکاری کنند و یا برای بر قراری ارتباط با انجمن‌ها و... از آن استفاده کنند.*

۵. ***بلاگ کردن برای جستجوی کار:*** *این کارکرد هنوز در حد رایج قرار نگرفته ولی بلاگرها می‌توانند تجربیات خود را به معرض نمایش همگان بگذارند.*

۶. ***بلاگ برای ثبت وقایع یك مسافرت:*** *گزارش مسافرتها همیشه مورد توجه بسیاری از مردم بوده است. این به خاطر آن است که مسافرین از بلاگها برای ثبت جزئیات و موارد مثبت سفر خود استفاده می‌کنند. از وقتی که کافی نت‌ها در همه جای جهان راه‌اندازی شده‌اند، بلاگرها می‌توانند در طول سفر هم، وبلاگ خود را به‌روز کنند.*

۷. ***بلاگ برای تجارت:*** *امروزه بسیاری از داد و ستدهای تجاری از طریق وب انجام شود و بسیاری از بلاگرها می‌توانند از این طریق درخواست خرید یا فروش محصولات خود را با کمترین هزینه (تقریبا بدون هزینه) در معرض تماشای دیگران قرار دهند.*

۸. ***بلاگ برای امور سیاسی:*** *بلاگهای سیاسی تبدیل به یك اجبار در رسانه‌ها و خبرگزاریهای سیاسی دنیا تبدیل شده اند و حتی سیاستمداران راههایی را برای استفاده از قدرت بلاگ‌ها برای نبردهای سیاسی و یا در ارتباط ماندن با موکلان خود استفاده می‌کنند و حالا مردم عادی هم می‌توانند حرف خود را در مورد مطالب سیاسی به راحتی بگویند.*

۹. ***بلاگ کردن برای اطلاع و آگاهی و اعلامیه‌های دولتی:*** *سازمانها و شرکتهایی که می‌خواهند سطح آگاهی مردم را بالا ببرند و خبرهای دولتی را گسترش بدهند با استفاده از وبلاگ‌ها منفعت زیادی می‌برند. هیأتها، کلوپها، انجمنها، کمیته‌ها و دیگر گروهها مثل موارد ذکر شده می‌توانند از بلاگ برای جذب مردم و بر قراری ارتباط استفاده کنند.*

۱۰. ***بلاگ کردن برای سود و منفعت:*** *بسیاری از بلاگرها از وبلاگ برای سود و منفعت خود استفاده می‌کنند و با راه‌اندازی شبکه‌های وبلاگی، از این روش کسب درآمد می‌کنند.*

۱.۲ وبلاگ‌های فارسی

۱.۲.۱ تاریخچه‌ی وبلاگ‌های فارسی

اکثر منابع متفق‌القول‌اند که اولین وبلاگ در زبان فارسی در شهریورماه سال ۱۳۸۰ هجری شمسی (۲۰۰۱ میلادی) توسط «سلمان جریری» (از دانشجویان رشته مهندسی کامپیوتر دانشگاه صنعتی‌شریف) آغاز شده است. نرم‌افزار ساخت و مدیریت این وبلاگ توسط خود جریری نوشته شده بود.

۱۸ روز پس از جریری، «حسین درخشان» (روزنامه‌نگار) با انتشار وبلاگش، که هم‌چنان به‌صورت دستی مدیریت می‌شد، وبلاگ‌نویسی‌اش را آغاز کرد و در ۱۴ آبان‌ماه همان‌سال، درخشان با انتشار مقاله‌ای آموزش وبلاگ‌نویسی به‌کمک سایت Blogger.com را آموزش داد[۵].

۱.۲.۲ آمار مربوط به وبلاگ‌های فارسی در اینترنت

به گفته‌ی [ضیایی‌پرور، ۸۶] در سال ۱۳۸۴، وبلاگ‌های فارسی که حدود ۲۵۰هزار وبلاگ بوده‌اند رتبه‌ی چهارم وبلاگ‌های فعال را داشته‌اند. در حالی‌که در سال ۱۳۸۶ و به‌دلیل رشد وبلاگ‌ها به سایر زبان‌ها (نظیر فرانسوی، اسپانیولی و ...) رتبه‌ی ایران به دهم تنزّل پیدا کرده است. به‌نقل از ویکی‌پدیا[۶]، در سال ۲۰۰۴، زبان فارسی نوزدهمین زبان رایج اینترنت بوده است و اکنون با استناد به آمار سایت NITLE Blog Census[۷] اکنون نیز زبان فارسی رتبه‌ی دهم زبان‌های دنیا را در وبلاگ‌ها داراست.

متأسفانه به‌دلیل تحریم کشور ایران، آمارهای به‌روز مربوط به زبان فارسی و درصد کاربران اینترنت در ایران، در وب‌سایت مرجع InternetWorldStats[۸] وجود ندارد و در این سایت حتی اسمی از ایران به‌عنوان یک کشور آسیایی هم برده نشده است. هم‌چنین شباهت نویسه‌های زبان فارسی با زبان عربی و نفوذ مؤثرتر کشورهای عرب‌زبان در مراجع آمارگیری (خصوصاً وابسته به ایالات متحده آمریکا) باعث شده تا بسیاری از اسناد

---

[۵] منبع: ویکی‌پدیا http://en.wikipedia.org/wiki/Blogging_in_Iran

[۶] http://en.wikipedia.org/wiki/Global_Internet_usage

[۷] قابل مشاهده در آدرس http://www.knowledgesearch.org/census/lang.html . باید در نظر داشت این سایت آمار وبلاگ‌های فارسی اندیس‌گذاری شده را ۱۹۰۰۰ وبلاگ نشان می‌دهد که قابل قیاس با ادعای چند صد هزار وبلاگ نیست. منتهی نرخ یک‌درصدی وبلاگ‌های فارسی نسبت به انگلیسی، به احتمال قوی می‌تواند معیار مناسب‌تری باشد.

[۸] به آدرس http://www.internetworldstats.com/stats3.htm

فارسی‌زبان در زمره اسناد زبان عربی (که اکنون رتبه‌ی هشتم[۹] در میان زبان‌های رایج در اینترنت را به‌خود اختصاص داده) قرار بگیرند.

از سوی دیگر، نبودن یک نهاد یا سازمان متولّی داخل کشور می‌تواند دلیل دیگری برای عدم دسترسی به داده‌های دقیق باشد. و به‌قول برخی وبلاگ‌نویسان، تنها در مسائل مربوط به فیلترینگ، نهادهای دولتی یادی از دنیای وبلاگستان می‌کنند؛ طوری‌که هنوز در فرهنگ بسیاری از شهرها و خانواده‌های ایران، «وبلاگ‌نویس» فردی‌است که تنها به‌صرف آزاد بودن تریبون اینترنت به این رسانه روی آورده و «وبلاگ‌نویسی» تلویحاً استفاده از آزادی افراطی، به‌ویژه در مورد انتقادات سیاسی است.

علی‌ای‌حال، با گزارش‌هایی که از وب‌سایت سرویس‌دهنده‌ی وبلاگ‌نویسی نظیر Persianblog.ir، blogfa.com و پس از آن‌ها mihanblog.com، blogsky.com، Parsiblog.com و ... می‌توان تعداد وبلاگ‌های فارسی را حتی قریب یک الی دو میلیون وبلاگ تخمین زد. منتهی این‌که چه‌تعداد از این وبلاگ‌ها هنوز نوزاد هستند (کم‌تر از ۱۰ پست دارند) و چه‌تعداد از این وبلاگ‌ها مرده‌اند (بیش از یک‌سال از آخرین به‌روزرسانی‌شان می‌گذرد) به‌سادگی قابل ارزیابی نیست. علاوه بر این، هیچ تخمینی از تعداد وبلاگ‌های فارسی که در شبکه‌های اجتماعی (نظیر Yahoo! 360 یا Facebook) به‌صورت خصوصی یا عمومی دنبال می‌شوند وجود ندارد.

به‌عنوان یک پیش‌بینی درباره‌ی نرخ رشد وبلاگ‌های فارسی در آینده، می‌توان چنین گفت که موج ابتدایی (آغازگر) وبلاگ‌های فارسی اکنون سپری شده است و از نظر اجتماعی، دیگر «داشتن وبلاگ» یک معیار فردی برای به‌روز بودن به‌شمار نمی‌رود. دلیل این امر از یک سو فراگیر شدن وبلاگ‌ها و از سوی دیگر ورود تکنولوژی‌های جدید است. برای مثال در سال‌های ۸۱ تا ۸۳، تنها واسط تعاملی برای فارسی‌زبانان در دنیای اینترنت، پس از Chat، داشتن وبلاگ بود؛ اما اکنون وفور شبکه‌های اجتماعی با امکاناتی به‌مراتب بیش‌تر از وبلاگ از یک‌سو و وجود تالارهای گفتمان متعدد در مسائل گوناگون و حتی بازی‌های آنلاین از سوی دیگر، باعث شده تا انگیزه‌ی کاربران تازه‌کار اینترنت برای ایجاد وبلاگ بسیار کم‌تر از قبل شود. متأسفانه نبودن یک

[۹] منبع: http://www.internetworldstats.com/stats7.htm

متولّی رسمی برای ارزیابی این تغییرات، ما را در دسترسی به اطلاعات دقیق و بررسی دلایل این تغییرات ناکام گذاشته است.

۱.۲.۳ تفاوت‌های وبلاگ‌های فارسی و غیر فارسی

دکتر حمید ضیایی‌پرور، که از اولین محققان دولتی (با حمایت وزارت فرهنگ و ارشاد اسلامی) در بررسی وبلاگ‌های فارسی می‌باشد، مهم‌ترین تفاوت وبلاگ‌های فارسی و غیرفارسی را چنین برمی‌شمارد[۱۰]:

> *مهمترین تفاوتی که میان این وبلاگ‌های فارسی و غیرفارسی از نظر محتوایی دیده می‌شود ناشی از محدودیت‌هایی است که در فضای سیاسی و اجتماعی کشور وجود دارد. این مسأله سبب می‌شود جوان‌های ما عمدتا دل‌نوشته‌های خود را منعکس کنند و مسائلی را بیان می‌کنند که در کشورهای دیگر جوان‌ها منعی ندارند آن‌ها را بگویند، بنابراین ما با حجم انبوهی حرف‌های خودمانی و داستان‌های روزمره، شعر و ادبیات مواجه‌ایم*
>
> *[...] از بعد ظاهری نیز وبلاگ‌های خارجی از سادگی برخوردارند، اما وبلاگ‌های ایرانی از لحاظ گرافیک و فرمت‌ها تنوع بسیاری دارند. همچنین در زمینه‌ی کامنت‌گذاری در وبلاگ‌های ایرانی این امر بیشتر به عنوان سرگرمی نگریسته می‌شود و کسانی‌که دارای وبلاگ‌های علمی و تخصصی هستند از این گذاشتن کامنت صرف‌نظر می‌کنند در حالی که در وبلاگ‌های غربی عمدتا بخش کامنت‌ها باز است و در این قسمت پیام‌های علمی و کارشناسانه رد و بدل می‌شود.*

به این تفاوت‌ها، می‌توان تفاوت در فرهنگ و انگیزه را نیز به‌طور دقیق اضافه کرد. به‌طور مثال تعداد زیادی از وبلاگ‌ها که از نظر محتوایی، در دسته «علمی» یا «آموزشی» قرار می‌گیرد، برخی مناسبت‌های عام نظیر نوروز یا یلدا را در یک پست جداگانه تبریک می‌گویند[۱۱]. از این رو، شاید بتوان ادعا کرد که عدم وجود عاملی برای محدود کردن و قاعده‌مند کردن پست‌ها در راستای هدف خاص، باعث می‌شود که «آزادی» در نگارش محتوا و سبک پست‌ها، آن‌ها را از «خاص‌منظوره» بودن و پرداختن به یک هدف واحد دور نگه‌دارد.

---

[۱۰] مصاحبه با ایسنا، فروردین ۱۳۸۴. قابل مشاهده در آدرس: http://isna.ir/ISNA/NewsView.aspx?ID=News-510169&Lang=P

[۱۱] مثال: وبلاگ گروه نرم‌افزار دانشگاه صنعتی شریف: http://software.ce.sharif.edu/2006/12/blog-post_20.html

پرواضح است که این موارد، تحلیل‌های محتوایی را نیز تحت‌الشعاع قرار می‌دهد. برای مثال شاهد هستیم که وبلاگ دانشجویان مهندسی نرم‌افزار یک دانشگاه، به‌صرف یک پست یک در آن جزوه‌ای از یک استاد را منتشر می‌کنند، وبلاگ خود را در شاخه‌ی «آموزشی» ثبت‌نام می‌کنند ولی در سه پست آتی به مسائل و دغدغه‌های روزمره (نظیر مشکلات سلف غذاخوری یا جوّ درس‌خواندن در خوابگاه‌ها) می‌پردازند.

شاید بتوان این محورگریزی را به‌نوعی به فرهنگ و ساختار اجتماعی جامعه‌ی ایران نسبت داد. به‌هر حال، بحث ریشه‌یابی این موضوع از حیطه‌ی این پایان‌نامه خارج است.

۱.۲.۴ بررسی‌های انجام شده روی وبلاگ‌های فارسی

چنان‌چه اشاره شد، هیچ نهاد یا وزارتخانه‌ی رسمی‌ای در ایران متولّی فعّالی در زمینه بررسی وبلاگ‌ها و سایر بانک‌های محتوایی زبان فارسی در وب نیست. تنها موارد انجام و منتشر شده نیز گزارشات دکتر ضیایی‌پور است که به بررسی موردی و غیرخودکار ۱۰۰ وبلاگ پرطرفدار، خصوصاً از نگاه سیاسی، پرداخته است. در سایر موارد، از وبلاگ‌ها به‌عنوان بستری برای استخراج یک گراف یا شبکه لینک‌ها استفاده شده است (نظیر [جمالی، ۸۶]) و یا استناد بر دسته‌بندی انجام شده توسط خود کاربران بوده است.

برخی از بررسی‌های انجام شده نیز بدون ذکر منبع و فاقد توضیح درباره‌ی روش تحقیق هستند. برای مثال [شوقی، ۸۵] با رویکرد «فردمحوری» وبلاگ‌نویس، اکثر وبلاگ‌نویسان را جوان و نوجوان دانسته و کمیّت موضوعات مختلف وبلاگ‌ها را ناشی از سلیقه‌های جوانان خوانده است. وی هم‌چنین هیچ توضیحی درباره‌ی نحوه‌ی جمع‌آوری اطلاعات ارائه نکرده است. در طبقه‌بندی او که بر روی ۱۶۴٬۸۴۵ وبلاگ فارسی در سال ۱۳۸۵ انجام شده، حدود ۵۳ هزار وبلاگ به مسائل شخصی و عمومی می‌پرداخته‌اند، در حالی‌که تعداد وبلاگ‌های زنان، کتاب‌داری و انقلاب اسلامی به‌ترتیب تنها ۶۸، ۴۷ و ۱۷ عدد است. این آمار نشان می‌دهد که بررسی مذکور چندان هم دقیق نبوده است.

۱.۳ زبان فارسی و وبلاگ‌های فارسی

دکتر حمید ضیایی‌پرور در مصاحبه‌ای با خبرگزاری ایسنا[۱۲] درباره‌ی تأثیر وبلاگ‌های فارسی بر زبان فارسی می‌گوید:

> *وبلاگ‌ها مرکز ثقلی برای احیای زبان فارسی در سطح بین‌الملل به شمار می‌آیند [...]. وبلاگ‌ها برخواسته از نیازها و خواسته‌های عموم جامعه هستند. بنابراین متناسب با فرهنگ بومی هر جامعه‌ای می‌تواند در تولید محتوا، گسترش فرهنگ و زبان آن جامعه نقش موثری داشته باشد [...].*
>
> *عمده‌ترین متون فارسی موجود در اینترنت متعلق به وبلاگ‌های فارسی زبان هستند [...]. اگر فرهنگستان زبان و ادبیان فارسی در ایران بخواهد از فرهنگ و زبان بومی حمایت کند یکی از بهترین روش‌های آن، حمایت از تولید کنندگان محتوای فارسی در اینترنت به ویژه وبلاگ‌ها، محسوب می‌شوند[...]. رشد نرم‌افزارها و موتورهای جستجو فارسی مانند گوگل فارسی، از جمله ابزاری است که اهمیت زبان فارسی را در دنیای اینترنت نشان می‌دهد.*

با توجه به فرمالیته بودن وب‌سایت‌های دولتی، اقتصادی/تبلیغاتی بودن وب‌سایت‌های سازمان‌های خصوصی و دولتی، محدود بودن تعداد کاربران اینترنتی تولیدکننده‌ی محتوی به‌زبان فارسی و هم‌چنین عدم حمایت عمومی در تدوین بانک‌های دانش (نظیر ویکی‌پدیای فارسی که فارسی زبان سی و دوم آن است[۱۳])، می‌توان مشاهده کرد که در اکثر جستجوهای عمومی برای عبارات فارسی در موتورهای رایج جست‌جو (نظیر یاهو و گوگل)، حداقل نیمی از نتایج مربوط به وبلاگ‌های فارسی یا خبرگزاری‌ها می‌باشند. در چنین شرایطی، نقش وبلاگ‌های فارسی در تولید و اشاعه‌ی محتوای فارسی در اینترنت به روشنی دیده می‌شود.

پرسشی که در این‌جا مطرح می‌شود این‌ست که محتوای فارسی گسترش یافته شده در وبلاگ‌های فارسی تا چه‌اندازه برطبق «اصول و قواعد خط و زبان فارسی» است؟

برای پاسخ به این پرسش، ابتدا باید «اصول و قواعد خط و زبان فارسی» را بشناسیم.

---

[۱۲] مصاحبه با ایسنا، مرداد ۱۳۸۴، قابل مشاهده در آدرس: http://isna.ir/ISNA/NewsView.aspx?ID=News-572508&Lang=P

[۱۳] منبع: http://meta.wikimedia.org/wiki/List_of_Wikipedias

۱.۳.۱ اصول و قواعد خطّ فارسی

زبان فارسی از جمله زبان‌های بسیار پویا است که در سال‌های اخیر حتی رسم‌الخط (دستور خط) آن نیز به‌شدت دست‌خوش تغییرات متعدد و بعضاً نامناسب شده است. متأسفانه عدم تأثیرگذاری کافی بنیاد دولتی حفظ خطّ فارسی یعنی «فرهنگستان زبان و ادب فارسی[۱۴]» باعث شده تا دستورخط‌های متنوعی برای تدوین کتب و مقالات ارائه شود. نمونه‌ای از این دستورخط‌ها [قدسی، ۸۳] است که تنها به نکاتی در باب خط (و نه زبان) اشاره کرده‌است که در تضاد با دستور خط مصوّب فرهنگستان زبان و ادب فارسی ([فرهنگستان ۸۴]) است. برخی، فراتر از اشاعه‌ی جدانویسی رفته و حتی یک زبان جدید ([زارعیان، ۸۳]) و رسم‌الخط نوین و متفوات وبلاگ alefbe.persianblog.ir را ابداع کرده‌اند.

ابوالحسن نجفی (مترجم، محقق، زبان‌شناس) در مصاحبه‌ای[۱۵] می‌گوید:

> *+ نجفی: **همه کسانی که بر جدانویسی اصرار می‌ورزند، دو استدلال دارند: یکی آنکه خواندن (تلفظ) با نوشتن مطابقت پیدا کند و دیگر آنکه شاگردان مدرسه آسانتر یاد بگیرند.** اما من تردید دارم که اگر جدا بنویسم بهتر بفهمند. ممکن است در آغاز برای دانش‌آموز اول ابتدایی کمی چسباندن «ها» مشکل باشد، اما وقتی یاد گرفت و چشمش به این صورت دید، دیگر مشکلی ندارد. شاید ابتدا کمی زحمت و زمان ببرد. آیا برای اینکه لقمه را بجویم و در دهان شاگرد بگذاریم، باید رسم‌الخط خود را عوض کنیم و سنت خود را زیر پا بگذاریم؟ [...]*
>
> *رسولی (مصاحبه‌کننده): برخی اعتقاد دارند که ما اصلا" سنت رسم‌الخط نداشته‌ایم. آیا واقعاً همین‌طور است؟*
>
> *+ نجفی: خیر. ما سنت رسم‌الخط داشتیم. در متون قدیمی هست و می‌توانیم ببینیم. اختلافاتی هست اما اختلاف در همان حد چسباندن یا جدا کردن «تر» صفت تفضیلی است، یا استفاده از شکل نوشتاری «ی» به عنوان صامت میانجی به جای همزه، مثلا در نامه‌ی پدرم. اما این اختلافات به اندازه اختلاف امروزی نیست. رسم‌الخط امروز یك بلبشوی واقعی است. تفاوت میان اختلاف جزئی و کلی است. اگر ما یك اختلاف جزئی میان دو رسم‌الخط داشته باشیم تا اینکه به‌کلی دو رسم‌الخط متفاوت باشد، خیلی فرق می‌کند.*

[۱۴] وب‌سایت: http://www.persianacademy.ir

[۱۵] مصاحبه از آرزو رسولی. موجود در آدرس: http://www.persian-language.org/Group/Talk.asp?ID=914

رسولی (مصاحبه‌کننده): این اختلافات رسم‌الخطی چه پیامدی دارد؟

+ نجفی: اصلاً مانع از مطالعه می‌شود. خود من با همه علاقه‌ای که به مطالعه رمان و داستان دارم، نمی‌توانم نوشته‌های امروز را بخوانم و مدتهاست هیچ رمان و داستانی نخوانده‌ام، مگر آنکه تکلیفی بر عهده‌ام باشد و اظهارنظری از من خواسته باشند. واقعاً خواندن این رسم‌الخط برایم شکنجه است. یادم رفت نمونه‌هایی برایتان بیاورم که ببینید با این رسم‌الخط چه می‌کنند. ما با اجازه‌ای که دادیم که در رسم‌الخط دخالت شود، هر کس حق خود می‌داند که این دخالت را بکند و اجتهاد کند که چطور باید نوشت. نتیجه آنکه دو نفر مثل هم نمی‌نویسند و رسم‌الخط واحدی نداریم. کسانی که قصدشان خدمت بود، چه خدمتی کردند؟ کسانی که می‌خواستند خط فارسی را اصلاح کنند، آن را خرابتر از آنچه بود، کردند. ***اصلاً یك نكته كلی‌تر بگویم: این خط فارسی اصلاح‌پذیر نیست.***

برخی از تغییرات جدید رسم‌الخط استدلال‌هایی واضح و شفاف دارند. برای مثال تغییر نگارش «بنام خدا» به «به‌نام خدا» به این دلیل است که «بنام» به‌معنی «مشهور» یک واژه‌ی متفاوت با ترکیب «به‌نام» است.

اما برخی از این تغییرات دلیل خاص و مستدلی (مانند استفاده از واژه‌ی نادرست در مثال فوق) ندارند. برای مثال برخی کتاب‌های چاپ شده (شاید در راستای گریز از شباهت‌ها با زبان عربی) تنوین انتهایی لغات را به‌صورت «ن» نگارش می‌کنند؛ نظیر «مثلن» یا «حتمن» و یا برخی کلمات عربی مصطلح نظیر «حتی» و «مستثنی» را به‌صورت «حتا» و «مستثنا» نگارش می‌کنند.

از تغییرات دیگر در این راستا می‌توان به «یای واصل در نقش کسره‌ی اضافه» اشاره کرد. برخی معتقدند نگارش «خانه‌ی ما» صحیح است (در راستای حذف همزه) و برخی هم‌چنان «خانهٔ ما» را صحیح می‌دانند. نکته‌ی مهم این‌جاست که فرهنگستان زبان فارسی در [فرهنگستان، ۸۴]، املای «خانهٔ ما» را توصیه می‌کند و مراجعی نظیر ویکی‌پدیای فارسی نیز از این دستور طبعیّت می‌کنند، حال آن‌که در کتاب‌های درسی دوره‌ی ابتدایی هم‌اکنون املای «خانه‌ی ما» تدریس می‌شود.

دسته‌ی دیگری از این تغییرات مربوط به جدانویسی و چسبیده‌نویسی است. برای مثال املای لغاتی نظیر «کتاب‌خانه» و «چه‌گونه» صورت جداشده‌ی «کتابخانه» و «چگونه» اند.

مثال‌هایی از این دست تنها به سلیقه‌ای شدن خط فارسی برمی‌گردند که متأسفانه با ترویج دستورهای مختلف خط، روزبه‌روز شاهد دگرگونی بیش‌تر متون فارسی هستیم. این دگرگونی نه‌تنها ثبات زبان فارسی را زیرسؤال می‌برد، بلکه به حفظ یکپارچگی اسناد و متون نیز لطمه می‌زند. باید توجه داشت که همان‌طور که استاد نجفی در مصاحبه خود گفتند، از آن‌جا که ساختار واژه در ذهن انسان کاملاً بصری و یکپارچه است (و نه حرف به حرف)، تفاوت در نگارش‌های مختلف خط، فرآیند مطالعه و یادگیری را دشوار، زمان‌گیر و کسل‌کننده می‌کند.

متأسفانه اکنون برای یک‌سان‌سازی رسم‌الخط‌ها دیر شده است و حتی برخی زبانشناسان معتقدند که خط فارسی هرگز به یک صورت واحد باز نخواهد گشت. از این رو، حداقل برای تسهیل دسترسی به اسناد و داده‌های رایانه‌ای، نیاز به کسب تجربیاتی در زمینه‌ی رفع چالش‌های پردازش متون زبان فارسی حس می‌شود که در این پایان‌نامه اقداماتی در همین راستا را پیش‌نهاد خواهیم کرد.

۱.۳.۲ خط فارسی در دنیای رایانه و اینترنت

مطالب گفته‌شده در فصل قبل تنها به دشواری‌ها و تفاوت‌های نگارش خط فارسی (مثلاً در نامه‌های کاغذی با خودکار) اشاره می‌کرد و متأسفانه در دنیای رایانه، این دشواری‌ها دوچندان می‌شود. برای مثال، عدم رعایت فاصله‌گذاری‌های صحیح (مثلاً عدم قرار دادن فاصله قبل از نقطه یا ویرگول، یا گذاشتن نیم‌فاصله بین «می» و «شود» در «می‌شود» و ...)، استفاده از نویسه‌های نادرست (مثلاً استفاده از ویرگول انگلیسی «,» به‌جای فارسی یا گیومه‌ی انگلیسی ”واژه“ به‌جای «واژه» یا «ي» عربی به‌جای «ی» فارسی)، تشخیص نادرست نویسه (نظیر «خانة ما» به‌جای «خانهٔ ما» به‌دلیل ریز بودن فونت) و ... از مشکلات مضاعف زبان فارسی در رایانه می‌باشند. با مقایسه‌ی این دشواری‌ها با سهولت‌های زبان انگلیسی، در می‌یابیم چرا پیش‌رفت‌های آنان در پردازش زبان بسیار سریع‌تر و ساده‌تر از اقدامات مشابه در زبان فارسی است.

اما از سوی دیگر، متأسفانه در این حوزه نیز عدم حمایت، پیگیری و تأثیرگذاری نهادهای دولتی به‌چشم می‌خورد. برای مثال در سال ۱۳۸۵، شورای عالی انقلاب فرهنگی در [شورا، ۸۵] موضوع فراگیر کردن دستور خط فارسی مصوب فرهنگستان زبان و ادب فارسی با بهره‌گیری از رایانه را به شرح ذیل تصویب نموده‌است.

به منظور سامان‌دهی نگارش زبان فارسی در محیط‌های رایانه و صیانت از دستورخط صحیح زبان فارسی، وزارت فرهنگ و ارشاد اسلامی موظف است ظرف مهلت شش ماه نسبت به طراحی نرم افزار نوشتار رایانه‌ای مجهز به سیستم غلط‌یاب خودکار و براساس دستورخط زبان فارسی مصوب فرهنگستان زبان و ادب فارسی اقدام نماید.

و طبق مصاحبه‌ی دکتر حمید ضیایی پرور با روزنامه‌ی همشهری[۱۶]، پس از گذشت دو سال و اندی، هیچ خبری از این بسته‌ی نرم‌افزاری نشده است. در اقدامی مشابه دبیرخانه‌ی شورای عالی اطلاع رسانی، طرح «تسما» (تولید و ساماندهی محتوای الکترونیکی ایران) را در سال ۱۳۸۴ با هدف ساماندهی به محتوای اینترنت آغاز کرد که این پرونده نیز ناتمام مانده است[۱۷].

برای ساماندهی و پردازش محتوای وبلاگ‌ها، نخست لازم‌ست «زبان وبلاگ‌ها» به‌درستی تبیین شود. مهندس بهرام اشرف‌زاده که از اولین بررسی‌کنندگان وضعیت زبان فارسی در وبلاگ‌ها بوده‌اند، در مقاله‌ی [اشرف‌زاده ۸۲] هشت ویژگی مهم و کلیدی در «زبان وبلاگ‌ها» را چنین معرفی می‌کنند:

توصیف کلی زبان وبلاگ ها کار چندان آسانی نیست، زیرا آزاد بودن نویسندگان در شیوه نوشتن و تفاوت های بسیار زیاد در تخصص ها و سلیقه های آنها، باعث شده است که هر وبلاگی زبان خاص خودش را داشته باشد. بااین حال، به نظر می رسد که بتوان ویژگی‌هایی را ذکر کرد که در غالب وبلاگ ها مشترک باشد. این ویژگی ها به شرح زیر است:

۱) از آنجا که وبلاگ‌ها عمدتاً از متن تشکیل شده‌اند، زبان آنها طبیعتاً گونه نوشتاری زبان یا زبان نوشتاری است. اما زبان نوشتاری وبلاگ‌ها به زبان گفتاری بسیار نزدیک است

۲) زبان وبلاگ‌ها عموماً زبان غیررسمی یا محاوره‌ای است. به علاوه، مطالب وبلاگ‌ها عموماً از قول خود نویسنده بیان می شود.

۳) زبان وبلاگ ها، عمدتاً همان زبان فارسی معیار است که در تهران به آن تکلم می شود.

۴) زبان وبلاگ‌ها حاوی واژه های متعددی از حوزه فنی رایانه یا کامپیوتر است.

[۱۶] رسانه‌های جهان و زبان فارسی. مصاحبه با دکتر حمید ضیایی‌پرور. همشهری آنلاین: http://www.hamshahrionline.ir/News/?id=61892

[۱۷] آخرین به‌روزرسانی وبسایت این طرح مربوط به سال ۱۳۸۵ است: http://www.scict.ir/Portal/Home/Default.aspx?CategoryID=865faefc-1cd1-4540-b470-b6c919bba080

۵) در زبان وبلاگ‌ها، واژه های بیگانه متعددی استفاده می شود که بعضاً با خط بیگانه نیز نوشته می شوند.

۶) رسم الخط وبلاگ‌ها عمدتاً غیراستاندارد و متغیر است، و مثل خود زبان محاوره، در معرض نوآوری‌های آنی است.

۷) نوشته‌های وبلاگ‌ها حاوی غلط‌های تایپی و نگارشی نسبتاً زیادی است، هرچند که اغلب وبلاگ های مهم و پرخواننده، نگارش قابل قبولی دارند.

۸) رسم الخط وبلاگ‌ها، تابع محدودیت های محیط الکترونیکی و عدم تطبیق آن با الزامات خط فارسی است.

پرواضح است که وجود این ۸ ویژگی در زبان وبلاگ‌ها، موجب دشواری پردازش متون و تبدیل آن‌ها به یک فرم استاندارد می‌شود. اما درسوی دیگر همین غیراستاندارد بودن می‌تواند در دسته‌بندی وبلاگ‌ها و تفکیک آن‌ها مفید واقع شود. برای مثال نوشتار سیستم‌عامل شرکت مایکروسافت (MS Windows) در وبلاگ‌های مرتبط با رایانه عملاً به‌صورت «ویندوز» است و این نوشتار فینگلیش، به‌خودی‌خود تمایز بین «پنجره» (که در وبلاگ‌های شخصی/ادبی می‌تواند رایج باشد) و «سیستم‌عامل پنجره (Windows)» را مشخص می‌کند. در حالی‌که در زبان انگلیسی همین کلمه‌ی Windows به خودی خود نمی‌تواند معیار باشد.

با توجه به این نکته، احتمال می‌رود که تشخیص دسته‌ی محتوایی وبلاگ از روی واژگان آن نتایج چندان بدی هم ندهد. پاسخ به میزان کارایی این نوع تشخیص، پرسش اصلی این پروژه است.

۱.۳.۳ جمع‌بندی دشواری‌های پردازش متون به زبان فارسی

در این فصل با مفهوم وبلاگ و وبلاگ‌های فارسی آشنا شدیم. با توضیحات ارائه شده، اکنون برای ما خیلی دور از انتظار نیست که به‌هنگام بررسی وبلاگ‌ها، با وبلاگی مواجه شویم که در دسته‌ی ورزش قرار گرفته (با طبقه‌بندی توسط خود کاربران) ولی به بیان خاطرات شخصی و با دستور خط پر از اشتباهی نوشته شده است.

در فصل بعد، قصد داریم نکات ذکر شده پیرامون دشواری زبان فارسی را حتی‌المقدور کلاسه‌بندی کنیم تا بتوانیم یک معیار مناسب برای استخراج گنجینه‌ی واژگانی یک وبلاگ داشته باشیم. هم‌چنین با بررسی تعداد

مناسبی از وبلاگ‌ها در یک دامنه‌ی محدود، قصد داریم شباهت بین گنجینه‌ی واژگان اصلاح شده و دسته‌ی وبلاگ‌ها (شخصی، ورزشی، آموزشی، ...) را با استفاده از معیار ارائه شده ارزیابی کنیم.

۲ فصل دوم)

تعریف پروژه و نحوه جمع‌آوری داده‌ها

در این فصل ابتدا معرفی صورت پروژه، موانع و ساده‌سازی‌های در نظر گرفته در انجام پروژه ذکر می‌شود. سپس با معرفی بستر اجرایی پروژه، منبع و نحوه‌ی جمع‌آوری داده‌های خام پروژه شرح داده می‌شود.

۲.۱ صورت مسئله

۲.۱.۱ تعریف پروژه

در این پروژه قصد داریم شباهت بین وبلاگ‌های فارسی بر اساس محتوایشان را بسنجیم و از آن به‌عنوان ابزاری برای دسته‌بندی وبلاگ‌های فارسی استفاده کنیم.

۲.۱.۲ دستاوردهای پروژه

از آن‌جا که وبلاگ‌های فارسی شاخص‌ترین تولیدکننده‌ی محتوای فارسی در وب هستند، نتایج این پروژه به‌صورت عملی در سایر حیطه‌ها نیز قابل گسترش خواهد بود. برای مثال می‌توان از همین روش برای تشخیص خودکار «دسته»ی یک خبر یا پیش‌نهاد خبرهای مشابه در خبرگزاری‌ها استفاده کرد و به‌نوعی عمل تگ‌گذاری خودکار را اجرا کرد.

۲.۲ موانع اصلی تحقیق و راهکارهای رفع آن‌ها

مشکلات عمده‌ی انجام این پروژه، همراه با راهکار پیشنهادی ما برای حل‌شان در زیر ذکر شده است. این راهکارها صرفاً در راستای رسیدن به هدف و عملی‌شدن پروژه پیشنهاد و اِعمال شده‌اند و غالباً با ساده‌ترکردن صورت مسئله، تنها یک راه‌حل عملی برای رفع مشکل ارائه می‌دهند. واضح‌ست که نزدیک‌ترین هر کدام از این

ساده‌سازی‌ها به صورت اصلی مسئله می‌تواند گام مفیدی در راستای ارتقای کیفیت نهایی بردارد. از همین رو، این موضوع را می‌توان اصلی‌ترین پیشنهاد راهبردی برای کارهای آتی دانست.

۲.۲.۱ تنوع در رسم‌الخط فارسی

در فصل قبل دیدیم که یکی از چالش‌های اساسی پیش روی محتواکاوی وبلاگ‌های فارسی، تفاوت‌های اساسی در رسم‌الخط‌های فارسی است. [مرتضایی، ۸۰] تعدادی از این مشکلات را در قالب ۵ دسته‌ی زیر مطرح می‌کند:

۱. **گوناگونی معادل‌های علمی:** به‌نقل از واژگان کتابداری و اطلاع‌رسانی[۱۸]، متخصصان این رشته ۶ معادل برای Manual، ۹ معادل برای Online، ۱۲ معادل برای Layout و ۱۳ معادل برای Cross Reference بکار برده‌اند. برخی مصوّبات فرهنگستان زبان فارسی (نظیر پرونَجا به‌جای فایل یا پروندان به‌جای زونکن) در این بازار آشفته سهم مهمی ایفا می‌کنند.

۲. **ضبط اسامی:** در برگردان اسامی خاص، عناصر و ترکیب‌های شیمیایی، ابزارها و تجهیزات، محل‌های جغرافیایی و ... از زبان‌های دیگر هیچ قاعده خاصی در فارسی وجود ندارد و در این میان تنها سلیقه‌ی مترجم دخیل است. برای مثال نامی نظیر «Robinson» به‌صورت‌های «رابینسن»، «روبینسون»، «ربینسون» و حتی «روبنسن» نوشته می‌شود. اصطلاحات تخصصی و جدیدالورود به‌دنیای کامپیوتر و رایانه نیز در این بین سهم مناسبی دارند. برای مثل «ماکروسافت»، «مایکروسافت» و «میکروسافت» هر سه نوشتاری رایج (ولو نادرست) برای Microsoft هستند.

۳. **سرهم نویسی، جدانویسی و بی‌فاصله‌نویسی:** این موضوع در فصل قبلی همین پایان‌نامه بررسی شد. بااین‌حال مشکلاتی نظیر «برنامه درسی» یا «برنامه‌ی درسی» یا «برنامه‌درسی» یا «برنامهٔ درسی» هم‌چنان پابرجاست. مثال دیگری از این سبک «آبگرمکن» یا «آب‌گرمکن» یا «آبگرم‌کن» یا «آب‌گرم‌کن» است. بعضی اسامی خاص نظیر «علی‌رضا» یا «علیرضا» نیز در این

[۱۸] هاشمی، ابوالفضل. «واژگان کتابداری و اطلاع‌رسانی». تهران، دبیرخانه هیأت امنای کتابخانه‌های کشور. ۱۳۷۶.

قاعده دخیل هستند و چون در این موارد دارنده‌ی نام مختار است، هردوی این نگارش‌ها را باید صحیح دانست.

۴. **انواع جمع:** تعدّد علامات جمع در فارسی (ها، ان، ات، ین، ون، مکسّر) در زبان فارسی موجب گردیده است در ثبت برخی واژگان، نظیر کلیدواژه‌ها، مشکلی به مشکلات بالا افزوده شود. برای مثال «استادها»، «اساتید» و «استادان»، هر سه در زبان فارسی رایج هستند. مقایسه‌ی این تعدّد با قواعد ساده و مشخص جمع در زبان‌های بیگانه (نظیر انگلیسی یا فرانسوی) می‌توانند دشواری این مشکل را به‌وضوح نشان بدهند.

۵. **صورت‌های مختلف نوشتاری:** همزه، الف مقصوره، تشدید و دوگانگی شکل نوشتاری، از مشکلات رایج در زبان فارسی است. برای مثال «هیئت» و «هیأت»، «آیت‌اله داود عطائی» و «آیت‌الله داوود عطایی»، «پاییز در آیینه» و «پائیز در آینه»، مثال‌هایی از این دوگانگی هستند.

در این پروژه با اِعمال فیلترهایی بر روی واژگان استخراج شده، سعی داریم تا حد امکان مشکلاتی از این قبیل را رفع کنیم. شرح کامل این فیلترها و اثرات آن‌ها در فصل‌های آتی ارائه می‌شود.

۲.۲.۲ پیچیدگی زبان فارسی و عبارات فارسی

پرواضح است که محل قرارگیری کلمات در یک جمله یا عبارت، می‌تواند در نقش و کاربرد آن واژه دخیل باشد. به‌عنوان یک مثال ساده، در زبان فارسی معمولاً کلمه‌ای نظیر «می‌شود» با فاصله از هم نوشته می‌شود (به‌صورت «می شود») و از همین رو پیش‌وند «می» را یک ایست‌واژه در زبان فارسی می‌شمارند. در سوی دیگر همین لفظ «می» در بیت اوّل ساقی‌نامه‌ی حافظ که می‌گوید «*بیا ساقی آن می که حال آورد / کرامت فزاید کمال آورد*»، یک کلمه‌ی مستقل و با معنی کاملاً متفاوت است. چنین مثال‌هایی در زبان فارسی که اِعراب‌گذاری در آن تنها به‌هنگام ضرورت است، به‌وفور یافت می‌شود. نظیر ترکیب سه‌حرفی «مهر» که هم می‌تواند «مُهر» (نشانه) باشد و هم «مِهر» (محبّت)؛ یا «شکر» که هم می‌تواند «شُکر» (سپاس‌گذاری) باشد و هم «شِکَر» (شیرین مانند قند).

به‌طور کلی می‌توان تأثیرات واژه در جمله یا عبارت که ما آن را نادیده می‌گیریم را چنین دسته‌بندی کرد:

۱. **واژه‌هایی که به دو گونه یک‌سان خوانده و نوشته می‌شوند.** نظیر مثال «می» در بالا یا «باز» در دو بیت زیر از غزلیات حافظ که به اصطلاح ادبا، از آرایه‌ی «جناس» استفاده کرده است. چنین واژه‌هایی درصورتی که مستقلاً بررسی شوند، برخی از معانی خود را از دست می‌دهند در حالی‌که در جمله یا عبارت اگر به آن‌ها نگاه شود، معنی دقیق خود را خواهند داشت.

نگشاید دلم چو غنچه اگر ساغری از لبش نبوید باز

بس که در پرده چنگ گفت سخن بهرش موی تا نمویید باز

۲. **واژه‌هایی که به‌دلیل عدم گذاشتن اِعراب مشابه نوشته می‌شوند.** نظیر شُکر و شِکَر در مثال بالا یا بیت زیر از قصاید حافظ. علی‌رغم این‌که گذاشتن اِعراب باعث می‌شود که این تفکیک صورت بپذیرد، اما از آن‌جا که این اعراب‌گذاری همواره انجام نمی‌شود، بود و نبود آن‌ها برای تفکیک صورت بدون اعراب، تأثیری نخواهد داشت.

مذاق جانش ز تلخی غم شود ایمن کسی که شکر شکر تو در دهان گیرد

۳. **واژه‌هایی که به‌واسطه‌ی یک فاصله‌ی خالی در میان‌شان به‌صورت دو واژه‌ی نامرتبط در می‌آید.** ما در این پروژه واژه‌ها را با استفاده از فاصله‌ی خالی از هم جدا می‌کنیم در حالی‌که همین فواصل خالی و ترکیب واژگانی می‌تواند در معنای واژه تأثیرگذار باشد. برای مثال واژه‌ی «کف‌بینی» (رمّالی) در صورتی که با فاصله نوشته شود معنای آن به‌کل عوض می‌شود؛ و یا ترکیب «چرخ‌زنان» در شعر زیر از حافظ که با فاصله به‌صورت «چرخ زنان» نوشته شده‌است. بدین‌سان استفاده از فاصله به‌جای نیم‌فاصله، باعث می‌شود تا در پروژه‌ی ما یک کلمه‌ی مرکّب به‌صورت دو کلمه‌ی مستقل به شمار رود که در معنا و کاربرد می‌تواند تأثیر نامطلوبی ایجاد کند.

کمتر از ذره نه‌ای پست مشو مهر بورز تا به خلوتکده خورشید رسی چرخ زنان

۴. **واژه‌هایی که با توجه به واژه‌های قبل یا بعد از خود معنی می‌گیرند.** برای مثال «جزیره جاوا[۱۹]» نام جزیره‌ایست که پایتخت اندونزی یعنی جاکارتا در آن قرار دارد. از سوی دیگر «زبان برنامه‌نویسی جاوا» یکی از پرکاربردترین زبان‌های برنامه‌نویسی، خصوصاً در ساخت نرم‌افزارهای گوشی‌های تلفن همراه است. با این وصف، صرف مشاهده‌ی «جاوا» نمی‌تواند مفهوم دقیقی برای ما به‌ارمغان بیاورد، در حالی‌که مشاهده‌ی «جزیره جاوا» یا «جاوای اندونزی» می‌تواند به مفهوم اوّل و عبارات «کد جاوا»، «برنامه جاوا» یا «دانلود جاوا» می‌تواند به مفهوم دوم اشاره کند.

با این توصیف‌ها، فرض استقلال واژگانی عملاً فرض بزرگی است که ما در این پروژه بر آن تکیه می‌کنیم؛ چراکه تعیین معنا و اثر واقعی یک واژه در عبارت نیاز به عملیات پردازش زبان طبیعی دارد که هم بسیار دشوار و هم در یک محیط غیراستاندارد (نظیر زبان وبلاگ) محتمل به داشتن درصد خطای بالا است.

جدای از این فرض، ما فرض یکپارچگی واژگانی یک وبلاگ را داریم. بدین‌صورت که تمامی واژه‌های ۱۰ پست اخیر یک وبلاگ را یک مجموعه درنظر می‌گیریم و به نرخ ظهور یک واژه در پست‌های مختلف کاری نداریم. پرواضح‌ست که این فرض می‌تواند بخش مهمی از اطلاعات بالقوه را دور بریزد. برای مثال وبلاگ الف را در نظر بگیرید که در ۵ پست از ده پست آخرش، در هر کدام دو بار از ترکیب «امام حسین» (ع) استفاده کرده‌است. در سوی دیگر وبلاگ ب را در نظر بگیرید که در یکی از ده پست آخرش، ده بار ترکیب «امام حسین» (ع) مشاهده می‌شود. به‌صورت ضمنی می‌توان دریافت که وبلاگ الف احتمالاً یک وبلاگ مذهبی است که اما سوم شیعیان محوریت نیمی از پست‌های آن را دارد، در حالی‌که وبلاگ ب، می‌تواند یک وبلاگ از هر موضوعی باشد که در یک پست، مناسبت‌های مربوط به ایّام محرّم را تسلیت گفته است و یا حتی به مضامین دیگری از آن ترکیب (نظیر راه‌اندازی ایستگاه مترو میدان امام حسین(ع)) اشاره کرده است.

۲.۲.۳ عدم وجود محوریت موضوع

همان‌گونه که در [شوقی، ۸۵] نیز اشاره شده‌است وبلاگ‌ها پایگاه‌های دانش آزاد و بی‌قاعده‌ای هستند که فقدان معیارهای سود و زیان در آن‌ها اغلب باعث می‌شود تا نویسنده (یا نویسندگان) وبلاگ حال و هوای

[۱۹] به‌نقل از http://en.wikipedia.org/wiki/Java

شخصی خود را به موضوع مشخص شده ترجیح بدهند. از این رو پراکندگی مطالب در وبلاگ امری اجتناب ناپذیر است و باید همواره این موضوع را مدنظر قرار داد.

از همین رو، انتظار موفقیّت بالای یک دسته‌بندی موضوعی وبلاگ‌ها انتظاری نادرست است؛ چرا که موضوع وبلاگ‌ها، به تقلید از محتوای وبلاگ‌ها، ممکن‌ست پویایی مهارنشدنی‌ای داشته باشد.

۲.۲.۴ گستردگی دنیای وبلاگ‌ها و نبودن سیاهه‌ی موضوعی برای تمام آن‌ها

در وهله‌ی نخست باید در نظر داشت که تعداد وبلاگ‌های فارسی بسیار زیاد (بیش از یک میلیون) است که متأسفانه لیست کاملی از آن‌ها در هیچ کجای وب موجود نیست. هم‌چنین باید در نظر داشت که اخیراً برخی از سرویس‌دهندگان وبلاگ‌های فارسی (نظیر PersianBlog) سیاهه‌ی موضوعی[۲۰] را از امکانات سایت خود حذف کرده و از این رو دسترس ما به وبلاگ‌ها بر اساس موضوع محدودتر شده است.

برای مقابله با این قبیل مشکلات، در این پروژه تنها وبلاگ‌های سایت بلاگفا را در نظر گرفته‌ایم. خوش‌بختانه فهرست موضوع وبلاگ‌های این سرویس دهنده در آدرس http://www.blogfa.com/Members قرار دارد. علی‌رغم این‌که این فهرست موضوعی توسط خود کاربران تهیه می‌شود، اما توصیه‌های بلاگفا به‌هنگام انتخاب موضوع و ثبت وبلاگ در فهرست موضوعی بسیار مناسب است. برخی از این توصیه‌ها بدین شرح هستند:

- *تنها در صورتی که اکثر مطالب شما (حدود دو سوم مطالب) مرتبط با یک موضوع خاص می‌باشد آن موضوع را برای بلاگ خود انتخاب کنید.*
- *درصورتی که موضوع انتخاب شده با مطالب وبلاگ شما همخوانی نداشته باشد، وبلاگ شما از فهرست موضوعی حذف شده و دیگر امکان انتخاب موضوع را نخواهید داشت .*
- *برخی موضوعات دارای سطوح جزئی تر هم میباشد اگر مطالب بلاگ شما با چند زیر گروه همخوانی دارد و یا اصولا متعلق به زیر گروه خاصی نیست سطح بالایی یا عمومی را انتخاب کنید. مثلا اگر در مورد شاخه های مختلف هنر مانند فیلم و ادبیات و ... می‌نویسید حتما موضوع هنر را انتخاب کنید اما اگر صرفا در مورد سینما می نویسید زیرگروه سینما را انتخاب کنید*

[۲۰] که غالباً آن را directory می‌نامند

از این رو انتظار می‌رود دسته‌بندی انجام شده در این فهرست موضوعی، برای پروژه‌ی ما مناسب باشد.

لازم به یادآوری‌است که محدود کردن دنیای مورد ارزیابی به وبلاگ‌های سایت بلاگفا، به هیچ‌وجه به کلیّت پروژه‌ی ما لطمه‌ای نمی‌زند، چرا که این سایت به‌عنوان یکی از ۳ پایگاه اصلی وبلاگنویسان فارسی‌زبان شناخته شده و به‌دلیل امکانات متنوع و کارایی آسان، انتخاب اوّل تعداد کثیری از کاربران می‌باشد. از این رو به‌احتمال قریب به یقین نمونه‌های آماری این سایت، نتایج بسیار مناسبی به ما خواهد داد.

۲.۲.۵ استخراج متن خام از میان تگ‌های HTML

از آن‌جا که پست‌های وبلاگ‌ها در قالب وبلاگ تلفیق می‌شوند تا HTML نهایی وبلاگ را بسازند، استخراج داده‌های خام پست‌ها کار ساده‌ای نیست چرا که تحت قالب‌های متنوع، داده‌ها فرمت‌های مختلفی می‌گیرند.

از همین رو، برای دسترسی ساده‌تر به داده‌های خام و هم‌چنین تصمیم‌گیری بر اساس حجم متناسبی از داده‌های فارسی، تنها ۱۰ پست آخر هر وبلاگ با استفاده از RSS آن استخراج می‌شود.

۲.۳ بستر اجرای پروژه

۲.۳.۱ معرفی وب‌سایت بلاگفا به‌عنوان بستر اجرایی

با توجه به استاندارد بودن فرمت‌های وبلاگ‌های میزبانی‌شده توسّط بلاگفا و هم‌چنین وجود یک دسته‌بندی موضوع غنی (با حدود ۲۰۰هزار وبلاگ)، وبلاگ‌های سایت بلاگفا به‌عنوان بستر تحقیقاتی این پروژه انتخاب شدند. فراگیر بودن این سرویس‌دهنده‌ی وبلاگ‌های فارسی زبان در طی ۵ سال گذشته[۲۱] و به ادعای خود سایت پربیننده‌ترین سایت وبلاگ‌های فارسی زبان باشد.

۲.۳.۲ دسته‌بندی موجود برای وبلاگ‌های سایت بلاگفا

سامانه‌ی فهرست موضوعی سایت بلاگفا به کاربران این وب‌سایت اجازه می‌دهد تا وبلاگ‌های خودشان را در یکی از ۱۶ دسته‌ی موضوعی از پیش تعیین شده ثبت کنند. برخی از این ۱۶ دسته (نظیر علم و فن‌آوری) خود شامل چندین زیردسته (نظیر پزشکی، محیط‌زیست، کشاورزی، نجوم و ...) می‌شوند که جمعاً ۵۹ دسته و زیردسته را تشکیل می‌دهند. در نمودار ۱: درصد تعداد وبلاگ‌های سایت بلاگفا در دسته‌های مختلف تعداد وبلاگ‌های موجود در هر یک از دسته‌های اصلی موضوعات (تعداد مجموع، شامل خود دسته و زیردسته‌ها) نمایش داده شده‌است.

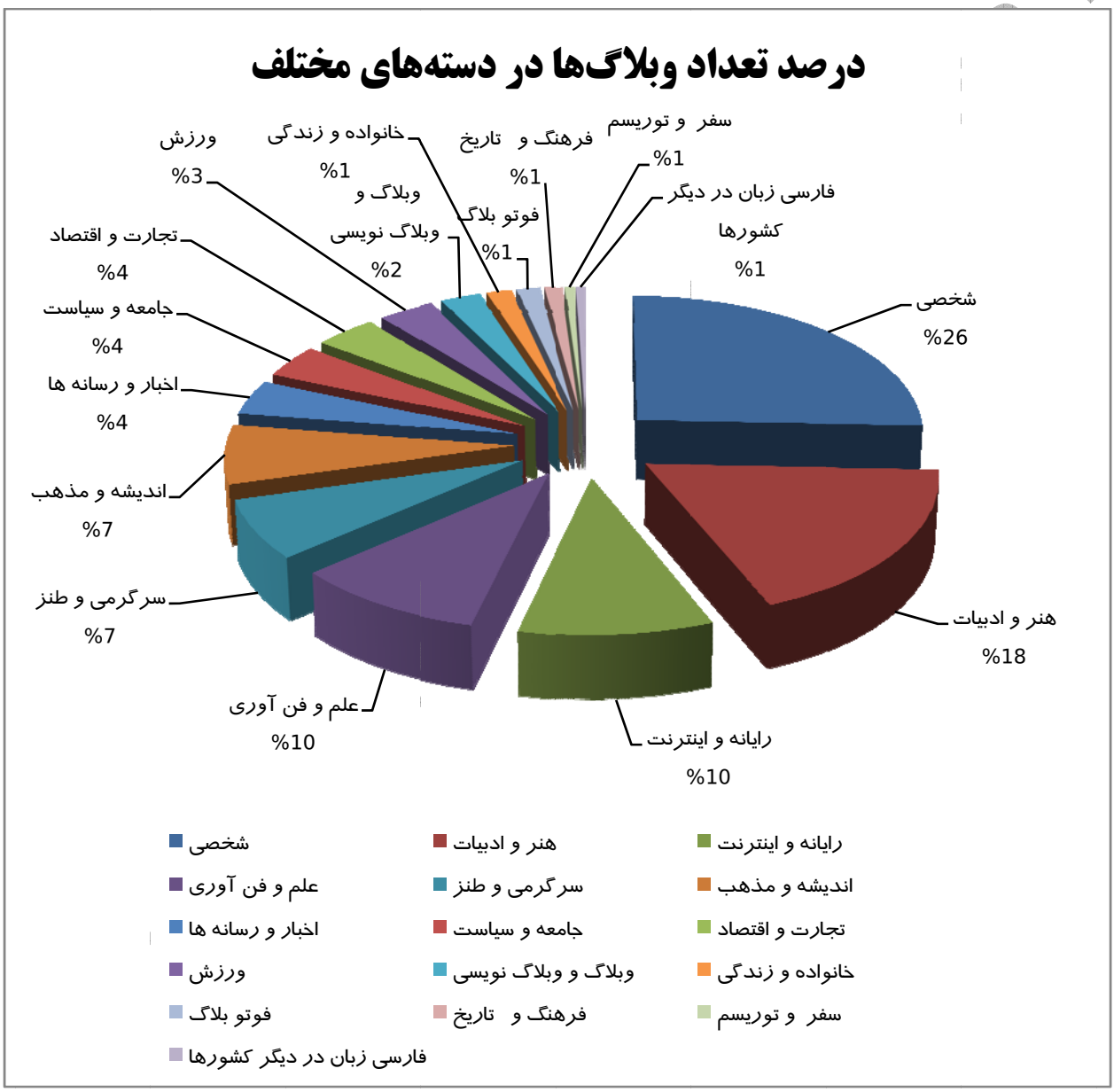
روش ثبت موضوع در بلاگفا به‌این صورت است که کاربر می‌تواند یکی از ۱۶ دسته اصلی یا ۴۳ دسته فرعی (که هر کدام از این دسته‌های فرعی، زیردسته دقیقاً یکی از دسته‌های اصلی هستند) را انتخاب کنند. در دسته‌بندی، یک وبلاگ تنها در یک دسته یا زیردسته ثبت می‌شود (توسط کاربر)، منتهی در لیست وبلاگ‌های اصلی یک گروه مثل علمی، هم وبلاگ‌های دسته‌ی اصلی این گروه، یعنی «علمی: عمومی» و هم وبلاگ‌های تمام زیردسته، مثل «علمی: کشاورزی»، «علمی: نجوم» و ... قرار می‌گیرند.

با این توضیح می‌توان ادعا کرد که دسته‌بندی عمومی برخی از دسته‌ها دچار پراکندگی موضوعی است. مثلاً وبلاگی که در دسته‌ی «هنر و ادبیات: عمومی» قرار گرفته باشد و در هیچ از زیردسته‌های مشخص شده‌ی هنر و

[۲۱] http://news.blogfa.com/Post-149.aspx

ادبیات نباشد، می‌تواند مربوط به «نقاشی»، «مجسمه‌سازی» یا «سفره‌آرایی» باشد که با هم اشتراک واژگانی خاصی ندارد.

در سوی دیگر، تعدادی از دسته‌های مشخص شده، بیان‌گر ویژگی نویسنده و نه محتوای وبلاگ هستند. برای مثال وبلاگ‌های «فارسی‌زبان در دیگر کشورها» عموماً از نظر محتوایی تفاوت خاصی با وبلاگ‌های شخصی و یا وبلاگ‌های مخصوص همان موضوع ندارند و غیرایرانی بودن نویسنده این دسته‌بندی را ایجاد می‌کند.



نمودار ۱: درصد تعداد وبلاگ‌های سایت بلاگفا در دسته‌های مختلف

مشکل سوم، کم بودن تعداد داده‌ها در یک زیردسته است. برای مثال تنها ۶۴ وبلاگ در زیردسته‌ی «یهود» قرار می‌گیرند که برای شناسایی یک زیردسته جدید در بررسی ما کافی نیستند.

با عنایت به نکات گفته شده، تعدادی از آن ۵۹ دسته/زیردسته‌های در بررسی ما حذف می‌شوند. جدول جدول ۱ : لیست کلاس‌های مورد بررسی لیست زیردسته‌های مورد بررسی ما (که از این به بعد به تمام آن‌ها «کلاس» می‌گوییم) را نشان می‌دهد. دسته‌های محذوف با دلیل در قسمت «در نظر گفته شده» پس‌زمینه تیره دارند.

کد دسته	دسته	زیردسته	تعداد	در نظر گرفته شده
01	اخبار و رسانه ها	عمومی	7994	عمومی
0101	اخبار و رسانه ها	اخبار و رویدادها	2122	1
0102	اخبار و رسانه ها	مطبوعات و رسانه ها	803	1
0103	اخبار و رسانه ها	خبرنگاران و روزنامه نگاران	814	1
02	هنر و ادبیات	عمومی	36158	عمومی
0201	هنر و ادبیات	ادبیات و شعر	16467	1
0202	هنر و ادبیات	کتاب	850	1
0203	هنر و ادبیات	سینما و تئاتر	2530	1
0204	هنر و ادبیات	موسیقی	7986	1
0205	هنر و ادبیات	تصویر برداری	684	1
0206	هنر و ادبیات	هنرمندان	2508	1
03	رایانه و اینترنت	عمومی	19649	عمومی
0301	رایانه و اینترنت	طراحی برنامه نویسی	1662	1
0302	رایانه و اینترنت	اینترنت	2525	1
0303	رایانه و اینترنت	اخبار آی تی و جامعه اطلاعاتی	1294	1
0304	رایانه و اینترنت	معرفی نرم افزار و بازیهای کامپیوتری	4292	1
0305	رایانه و اینترنت	سخت افزار	595	1
0306	رایانه و اینترنت	هک و امنیت شبکه	1779	1
04	علم و فن آوری	عمومی	19525	عمومی
0401	علم و فن آوری	بهداشت و سلامت	1370	1
0402	علم و فن آوری	پزشکی	1368	1
0403	علم و فن آوری	طبیعت و محیط زیست	1091	1
0404	علم و فن آوری	زبانهای خارجی	826	1
0405	علم و فن آوری	کشاورزی و بیو تکنولوژی	1088	1
0406	علم و فن آوری	برق و الکترونیک	1955	1
0407	علم و فن آوری	معماری و عمران	2032	1
0408	علم و فن آوری	نجوم	937	1

0409	علم و فن آوری	کتابداری	334	1
0410	علم و فن آوری	علوم پایه	2585	1
0411	علم و فن آوری	پرستاران	344	1
05	تجارت و اقتصاد	عمومی	7591	عمومی
0501	تجارت و اقتصاد	مقالات و مطالب	1202	1
0502	تجارت و اقتصاد	وبلاگ شرکتها و سازمانها	1628	1
0503	تجارت و اقتصاد	تجارت الکترونیک	1642	1
0504	تجارت و اقتصاد	کسب درآمد در اینترنت	1745	1
06	اندیشه و مذهب	عمومی	13657	عمومی
0601	اندیشه و مذهب	فلسفه و عرفان	1617	1
0602	اندیشه و مذهب	اسلام	6287	1
0603	اندیشه و مذهب	قرآن	957	1
0604	اندیشه و مذهب	مسیحیت	275	1
0605	اندیشه و مذهب	زرتشت	207	1
0606	اندیشه و مذهب	یهود	64	عدم کفایت تعداد
07	فوتو بلاگ	عمومی	2803	1
08	وبلاگ و وبلاگ نویسی	عمومی	4808	عمومی
0801	وبلاگ و وبلاگ نویسی	قالبهای وبلاگ	441	1
0802	وبلاگ و وبلاگ نویسی	مباحث وبلاگ و وبلاگ نویسی	248	1
09	فرهنگ و تاریخ	عمومی	2119	1
10	جامعه و سیاست	عمومی	7625	عمومی
1001	جامعه و سیاست	سیاست روز	2328	1
1002	جامعه و سیاست	زنان	396	1
1003	جامعه و سیاست	جامعه	1318	1
11	ورزش	عمومی	6634	1
12	سرگرمی و طنز	عمومی	13897	1
13	شخصی	عمومی	50766	1
14	خانواده و زندگی	عمومی	2963	1
15	سفر و توریسم	عمومی	1083	1
16	فارسی زبان در دیگر کشورها	عمومی	1052	یک دسته‌ی محتوایی نیستند
1601	فارسی زبان در دیگر کشورها	تاجیکستان	33	
1602	فارسی زبان در دیگر کشورها	افغانستان	560	

جدول ۱ : لیست کلاس‌های مورد بررسی

اکنون لیست کلاس‌های مورد بررسی جمعاً ۴۷ مورد هستند که در فایل `list/classes.txt` به همراه کد دسته نوشته شده‌اند.

۲.۴ جمع‌آوری داده‌ها

۲.۴.۱ پیدا کردن وبلاگ‌ها

در بخش قبلی گفتیم که ۴۷ دسته موضوعی (کلاس) از وبلاگ‌های فارسی بلاگفا را انتخاب کرده‌ایم. برای یک‌سان‌سازی حجم محتوای بررسی، از هر کدام از این ۴۷ دسته، ۲۰۰ وبلاگی که اخیراً به‌روز رسانی شده‌اند را پیدا می‌کنیم تا مجموعاً حدود ۱۰هزار وبلاگ را تشکیل بدهند. نام (آدرس) این ۱۰هزار وبلاگ در شاخه‌ی `list` پروژه قرار دارد (در فایل با کد-دسته‌ی کلاس مربوطه).

این کار با اجرای فایل `findLists.php` به‌صورت خودکار و با دریافت اطلاعات وبلاگ‌ها از آدرس‌های با فرمت http://www.blogfa.com/Members/UsersList.aspx?dir=$dirname$&page=$page$ انجام شده‌است که $dirname$ کد دسته و $page$ شماره صفحه وبلاگ‌های آن‌دسته (در هر دسته ۲۰ وبلاگ) را دارد.

به‌عنوان مثال لیست ۲۰۰ وبلاگ منتخب کلاس «سینما و تئاتر» (کد 0203) در جدول ۲: ۲۰۰ وبلاگ انتخاب‌شده برای کلاس «سینما و تئاتر» آمده است. تمامی این نام‌ها با آدرس blogfa.com. پایان می‌یابند.

۲.۴.۲ دریافت محتوای RSS

در گام بعدی RSS هر کدام از این ۹۴۰۰=۴۷×۲۰۰ وبلاگ ذخیره شده و در فایلی با نام خود وبلاگ و در شاخه `RSS/$class$/` ذخیره می‌شود که `$class$` شماره کلاس است. این کار با اجرای فایل `collectRSS.php` انجام می‌پذیرد که در آن شماره کلاس تنظیم شده و روی تمامی ۴۷ کلاس اجرا می‌شود.

اکنون در پایان مرحله جمع‌آوری داده‌ها ما هر وبلاگ را با یک فایل RSS می‌شناسیم. متأسفانه دسترسی به برخی از این وبلاگ‌ها توسّط مقامات قضایی مسدود شده‌است و در نتیجه فایل RSS آنان خالی یا حاوی یک پیام خطای دسترسی است؛ برخی نیز کاملاً پاک شده‌اند و تنها یک ردّ تغییر مسیر[۲۲] از آنان باقی مانده است.

[22] Redirect

500film	30namatarane	30namagarane-javan	1400film	09153621161
abdollahi-ha	aamirkhanbollywood	7thkind	60moj	5557520
alirezapoorsabagh	alimosleh	agdahak78	adamhayeroshan	absurddrama
ariaghoreishi	arfilm	amir-mor	amir-khajouie	aminetarokh
asrbekheir-hossein67	ashkiolabkhandi	artyekta	artiston2	arishtagroup
bazi-behdad	bandaregenaveh	bamn-f	bahmanabdollahi	azaadpic
cfj	bia2musicworld	bestload	behzad-joonam	beheshtidecor
cinemacafe	cinema57	cinema-cinemast	chehrepardaz	charpayetheater
comingsoon	clubninja	clipmob	cinemafilmha	cinemacallisto
dezdemona	dedalus	darabgerdtheater	dagvill	cutnews
eshghecinema	empori-sea-1372	eghlimaaaaaaa	dvdclassicfilm	donya2film
filmcomment	filmarchive	fathihassan	farzollahi	farid76
groupgolzar	gigabit	ghazalema	ghatreiazeshgh	gaw-sandogh
hanhayjin	hamedazizi	halgheh3	halfmoon	hadis591
hendi	heatfilm	hastiwaman	happytheatre	hanhyejin2
imdbfa	ibcpars	hosseinsoleimani	honar3	hollywood20
isfahan-theater	iribtv	iranshortfilm	iranianactors	iranboys20
kakashi	jumong4	javadroshan	jaskphoto	jahanpa
kianooshsemsari	khateratttt	khatam	kat	kamynetworkmovie
m-dogoharani	leila333	ky2	kouroshfilm	koochee
maji	mahnaficy	mahmoodnazeri	mahkhorshid	mahdis2011
mehdifilm2	mashghshab	manoochehr-khandan	maninasr	managara
mizansen	miladarang	migliore1362	meysamg	mehditabatabaei3d
mostafazamane	mostafarekabi	mojtaba000	mohsensarafi	mohamadireza
naghdefilm	myfreshmovies	my-call	msntheater	movie30ty
newavangard	nastaranbanou	nasimkocholo	namayesh0	najvan
okhovat1art	nnezami2001	niusha-zeighami0711	nimagroup	nezamabadi
perscnema	pakantheater	pafa	openfestival	onstage
pulpfiction94	postadam	pooyan-animation	pinkmax	persiapix
ramaneman	ram-art	raimand	rafei	qobbezarrin
saba-shop	rojinrahimi	rmvb-mkv-movie	rezacoron	rephotography
scarymovies	sahmema	saharfilm	sabzjan	sabrian
shabahpasha	sfxiran	seda-doorbin-sher	secret-7	se7enfilm
srttuteatr	soroushgudarzi	sinama-hl	simple	shanashir
tamashaclub	taktoop	t-t-aiene	sweeny1373	stalag17
tehran-movies	tazyeh-m	taravat-theater	taranehh-alidoosti	tarafdaran-glory
trannoom	time4me	thirdman	theater-theater	television-iran
vahiddarvishi	vahid-aks	v-film	twoshot	trix
yahage	wiseman	webnava	viista	vahmehonar
zohal777	zahrm75	z-movie	yeganehtheatergroup	yasna-bust

جدول ۲: ۲۰۰ وبلاگ انتخاب‌شده برای کلاس «سینما و تئاتر»

با این وصف، برای رفع این مشکل ما ۱۰ درصد فایل کم‌حجم‌تر (۲۰ عدد) از هر کدام از کلاس‌ها را حذف می‌کنیم تا دسته‌های ما به‌جای ۲۰۰ وبلاگ، عملاً ۱۸۰ وبلاگ داشته باشند. اکنون این فایل‌ها، که بین پنج تا ۵۰۰ کیلوبایت حجم داده‌ای دارند، مجموعه‌ی این ۸۴۶۰=۱۸۰×۴۷ وبلاگ، محک[۲۳] ما را با حجم تقریبی ۴۴۰ مگابایت تشکیل می‌دهند.

از آن‌جا که مراحل بعدی به تحلیل یک وبلاگ (یک RSS)، استخراج واژگان و اِعمال فیلترهای مختلف روی آن اختصاص دارد و ممکن‌ست این مراحل به نوع دیگری در کارهای آتی انجام شوند، آن مراحل را در یک فصل جداگانه ذکر می‌کنیم.

[23] Benchmark

۳ فصل سوم)

مناسب‌سازی و تحلیل داده‌های یک وبلاگ

در این فصل به تشریح گام‌های مختلف تبدیل داده‌های یک وبلاگ (یک فایل RSS) به فرمت مناسب برای ادامه‌ی کار و مقایسه‌ی وبلاگ‌ها با هم، پرداخته می‌شود. این گام‌ها شامل معرفی ۳ نوع فیلتر ساختاری (تجزیه ساختار RSS، مستقل از زبان فارسی)، زبانی (وابسته به ویژگی‌های خاص زبان فارسی) و نهایتاً پایانی (تبدیل داده‌ها به فرمت مطلوب ما) می‌باشند.

در پایان این فصل، داده‌های ما آماده برای اجرای الگوریتم‌های مورد نظر خواهند بود؛ چرا که هر وبلاگ تنها به‌صورت یک گنجینه‌ی واژگانی ساده با نرخ تکرار هر واژه در دسترس خواهد بود.

۳.۱ تحلیل یک RSS و معرفی فیلترها

یک فایل RSS تهیه شده از محک ما، یک فایل متنی با ساختار XML[۲۴] است. در بدنه‌ی این XML ابتدا آدرس، عنوان، معرفی و تاریخ آخرین ساخت RSS ذکر شده است، سپس ۱۰ پست اخیر هر کدام به‌صورت یک item آمده‌اند. هر پست شامل لینک مستقیم، محتوا (خلاصه‌ی محتوا در صورت داشتن بخش «ادامه مطلب»)، تاریخ، نام نویسنده و لینک به صفحه نظرات است.

در این بخش ما با اِعمال فیلترهای متعدد، این ساختار متنی XML را به فرمت مناسب برای تحلیل خودمان (گنجینه واژگانی با نرخ تکرار مفید) تبدیل می‌کنم. فیلترهایی که در این بخش استفاده می‌شوند به ۳ دسته‌ی ساختاری (با نماد S[۲۵])، زبانی (با نماد L[۲۶]) و نهایی (با نماد F[۲۷]) تقسیم می‌شوند.

[24] eXtensible Mark-up Language

[۲۵] مخفف Structural

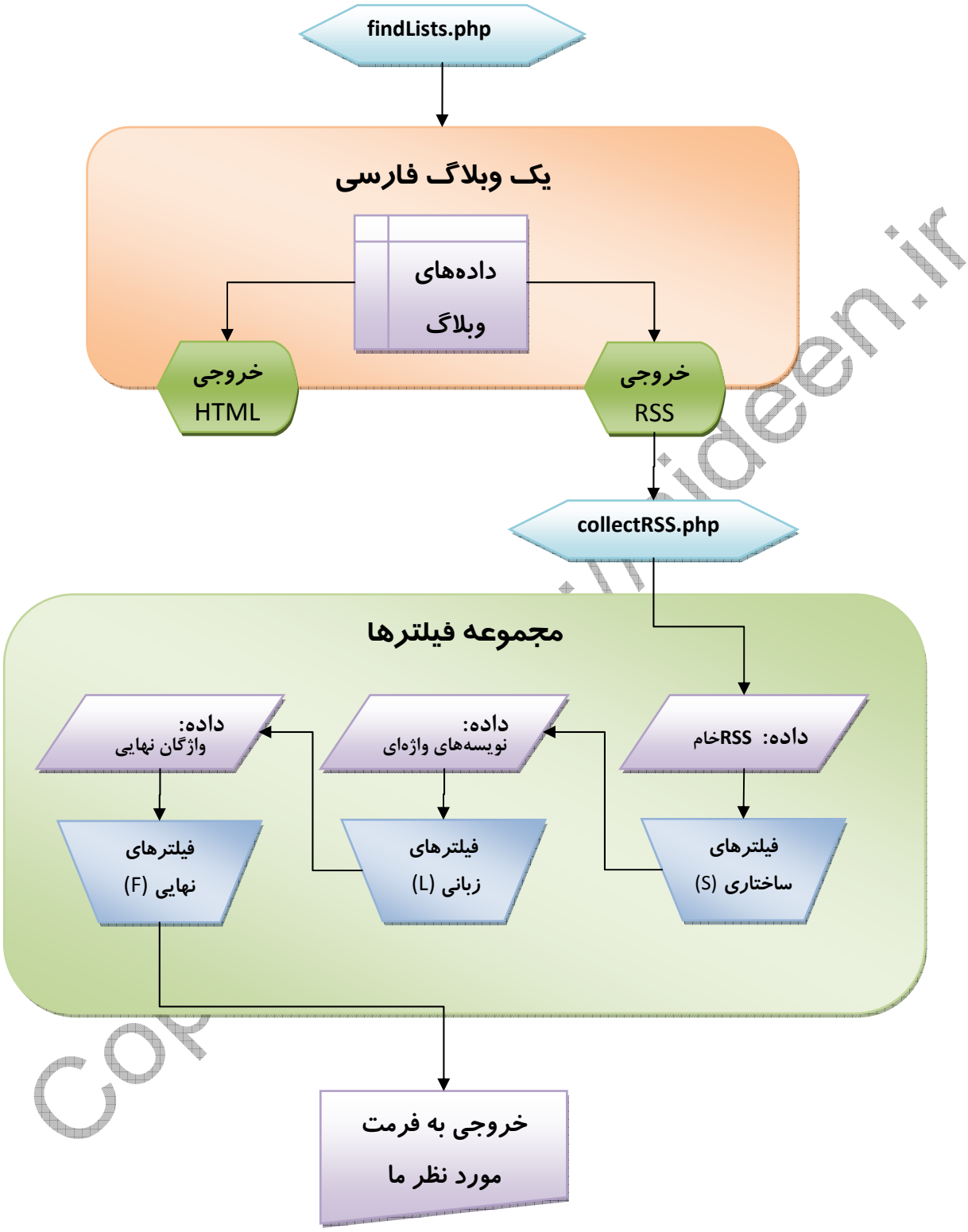
[۲۶] مخفف Linguistic

[۲۷] مخفف Finalizer

فیلترهای ساختاری، در تغییر ساختار و فرمت داده اثر دارند و مستقل از دانش خاص ما راجع به چگونگی شکل‌گیری واژگان فارسی قضاوت می‌کنند. به‌عبارت دیگر این فیلترها، به‌همین ترتیب و با تغییرات داخلی، برای پردازش هر زبان دیگری نیز باید استفاده شوند. نمونه‌ای از این فیلترها تبدیل «ي» عربی (نویسه نادرست) به «ی» فارسی (نویسه صحیح) است. خروجی این فیلترها تعدادی نویسه‌ی متوالی به‌مثابه‌ی واژه است.

فیلترهای زبانی، که پس از فیلترهای ساختاری اعمال می‌شوند، واژه‌های گرفته شده را با توجه به دانش‌ما درباره‌ی زبان فارسی نرمال‌سازی می‌کنند. حذف ایست‌واژه‌ها، یکسان‌سازی همزه‌ها و ریشه‌یابی سطحی، مثال‌هایی از فیلترهای زبانی ما هستند.

نهایتاً فیلترهای نهایی یا تثبیت‌کننده، خروجی فیلترهای زبانی را به فرمت ایده‌آل ما تبدیل می‌کنند. اعمالی نظیر متناسب سازی نرخ تکرار و نوشتن در فایل خروجی برای خواندن راحت‌تر داده‌ها در زبان PHP از وظایف این فیلترها هستند. نمودار ۲ مراحل مختلف پروسه و نحوه کارکرد این فیلترها را نمایش می‌دهد.



نمودار ۲: پروسه‌ی جمع‌آوری و مناسب‌سازی داده‌ها

۳.۲ فیلترهای ساختاری

چنان‌که گفته شد، فیلترهای ساختار وظیفه تبدیل ساختار RSS به دنباله‌ای از نویسه‌های متوالی (که واژه نامیده می‌شوند) را دارند. این فیلترها به متن به‌شکل دنباله‌ای از کلمات نگاه می‌کنند و وارد جزئیات داخلی متن (نظیر اعراب‌گذاری یا املای نادرست) نمی‌شوند. فیلترهای ساختاری به‌کار رفته شده در این بخش به شرح زیرند.

۳.۲.۱ فیلتر ساختاری یکم (S01): تجزیه‌کردن یک RSS

همان‌گونه که گفته‌شد، RSS یک فایل XML با فرمت و تگ‌های مربوط به خودش است. در این فیلتر، ما از بین تگ‌های مختلف یک RSS، تنها تگ‌های title و description موجود در itemها و هم‌چنین title و description اصلی را انتخاب کرده و محتوای این تگ‌ها را به‌صورت متنی ذخیره می‌کنیم.

این عملیات بر عهده‌ی تابع `parseRSS` در فایل `RssReader.php` است که یک XMLParser زبان PHP را پیاده‌سازی کرده است.

نکته قابل توجه در این فیلتر، اِعمال تأثیر عنوان و توضیح خود وبلاگ در متن خروجی است. از آن‌جا که نام و عنوان وبلاگ مهم‌ترین و دقیق‌ترین توصیف‌کننده‌های یک وبلاگ هستند، محتوای این دو بخش به‌ترتیب با ضریب تکرار ثابت ۱۰ (برابر تعداد پست‌های RSS) برای عنوان و ۵ برای توضیحات در تمام وبلاگ‌ها، در دنباله‌ی واژگانی اِعمالی می‌شوند. به‌عبارت واضح‌تر وجود یک کلمه‌ی خاص نظیر «زیستگاه» در عنوان یک وبلاگ، معادل تکرار این واژه در تمامی پست‌ها و وجود آن در توضیحات یک وبلاگ برابر تکرار آن در نیمی از پست‌ها به‌شمار می‌رود. انتظار می‌رود که این سیاست باعث شود تا وبلاگ‌هایی که عنوان‌های مشابه یا مرتبطی دارند، نزدیکی بیش‌تری بین‌شان به‌وجود بیاید.

۳.۲.۲ فیلتر ساختاری دوم (S02): حذف تگ‌های HTML

از آن‌جا که ما در این پروژه تنها با واژگان فارسی وبلاگ سر و کار داریم، در این فیلتر، با استفاده از تابع `strip_tags` زبان PHP اقدام به حذف تگ‌های HTML از RSS خوانده شده می‌کنیم.

با اِعمال این فیلتر، متن ما به‌صورت یک دنباله از کاراکترها خواهد بود که تشکیل متن ساده را می‌دهند.

۳.۲.۳ فیلتر ساختاری سوم (S03): تبدیل «ي» و «ك» عربی به «ی» و «ک» فارسی

ابتدایی‌ترین عمل بر روی داده‌های دریافت شده تبدیل حروف «ي» و «ك» عربی به «ی» و «ک» فارسی است. این مشکل از استاندارد نبودن صفحه‌کلیدهای رایج (مثلاً صفحه‌کلید پیش‌فرض سیستم‌عامل MS Windows XP) ناشی می‌شود. با اِعمال این فیلتر، تنها کافی‌ست یک‌بار واژه‌ی «مى» را در فیلترهای آتی بنویسیم و مطمئن باشیم این لفظ روی «مي» (با ي عربی) نیز کار خواهد کرد.

لیست این تبدیلات نویسه‌ای در جدول ۳ آمده‌است.

نویسه‌ی سابق	کد یونیکد قبلی	توضیحات	نویسه‌ی جدید	کد یونیکد جدید
ۀ	U+06C0	ه و همزه روی آن[۲۸]	ۀ	U+647U+0654
ھ	U+06BE	های دوچشم عربی	ه	U+0647
ي	U+064A	یای عربی	ی	U+06CC
ى	U+0649	الف مکسوره عربی[۲۹]	ی[۳۰]	U+06CC
ك	U+0643	کاف عربی	ک	U+06A9

جدول ۳ تبدیلات الزامی نگارشی، ناشی از اشتباه نگارنده

نکته قابل توجه در این فیلتر این است که ما فرض می‌کنیم کاربر هرگز به‌عمد از «ي» استفاده نمی‌کند و تمامی «ي»های موجود در واقع نسخه‌ی نادرستی از «ی» هستند. تبدیل‌های دیگری از این دست (تبدیل نویسه به نویسه، نظیر «آ» به «ا») که در راستای یکسان‌سازی نوشتار هستند، چون الزاماً تبدیل «نگارش نادرست به درست» نیستند، در فیلتر زبانی بخش ۳.۳.۲ اجرا می‌شوند.

[۲۸] باید دقت شود که این ترکیب تنها یک نویسه‌ی عربی یونیکد است و با «ه» + « ٔ » فارسی متفاوت است. استفاده از این نویسه باعث می‌شود که در ترکیب «خانهٔ من» عملاً واژه «خانه» موجود نباشد، حال آن‌که «خانهٔ من» تلفیقی از «خانه»+« ٔ »+فاصله+من است.

[۲۹] نام دقیق این نویسه «Arabic Letter Alef Maksura» است: http://www.fileformat.info/info/unicode/char/649

[۳۰] نام این نویسه «Arabic Letter Farsi Yeh» است: http://www.fileformat.info/info/unicode/char/6CC

۳.۲.۴ فیلتر ساختاری چهارم (S04): حائل واژگان

حائل[۳۱]ها نویسه‌هایی هستند که مرز بین کلمات را مشخص می‌کنند و در میانه‌ی یک کلمه به‌کار نمی‌روند. ساده‌ترین حائل‌های موجود فاصله خالی[۳۲] و کاراکتر خط جدید[۳۳] هستند؛ منتهی در این پروژه از آن‌جا که تنها خود واژگان به‌تنهایی و به‌عنوان دنباله‌ای از نویسه‌های فارسی مدّنظر ما هستند، ما تمامی کاراکترهای جدول ۴: لیست نویسه‌های حائل به‌کار گرفته شدهجدول ۴ را به‌عنوان حائل برای جداسازی واژگان به‌کار می‌بریم.

,	.	(	)	[	]	{	}	!	؟
\n	\r	\t	؟	،	[34];	_	;	\	
	'	%	:	-	>	<	=	|	/
#	@	$	%	^	&	*	`	'	~
1	2	3	4	5	6	7	8	9	0
۱	۲	۳	۴	۵	۶	۷	۸	۹	۰
a	b	c	d	e	f	g	h	i	j
k	l	m	n	o	p	q	r	s	t
u	v	w	x	y	z	è	é	à	á
A	B	C	D	E	F	G	H	I	J
K	L	M	N	O	P	Q	R	S	T
U	V	W	X	Y	Z				

جدول ۴: لیست نویسه‌های حائل به‌کار گرفته شده

تمامی این کاراکتر در فایل `delimiters.txt` با «و» فارسی جداشده‌اند (تو توسط تابع `explode` زبان PHP به‌راحتی تفکیک شوند).

۳.۲.۵ فیلتر ساختاری پنجم (S05): حذف ایست‌واژه‌ها

ایست‌واژه[۳۵]ها واژگانی هست که تقریباً در تمام متون (مستقل از محتوا و سبک) به‌وفور تکرار می‌شوند و حضورشان به‌تنهایی هیچ اطلاع خاصی به ما نمی‌دهد. در اکثر پژوهش‌های بازیابی اطلاعات، خصوصاً پژوهش‌های واژگانی مستقل، ایست‌واژه‌ها پیش یا پس از جمع‌آوری داده‌های خام حذف می‌شوند. حذف

---

[31] Delimiter

[۳۲] Blank Space. کاراکتر 0x20 هگزادسیمال جدول اسکی

[۳۳] New Line. کاراکتر 0x0A هگزادسیمال جدول اسکی

[۳۴] کاراکتر Non Breaking Space

[۳۵] Stop Words به نقل از ویکی‌پدیا: http://en.wikipedia.org/wiki/Stop_words

ایست‌واژه‌ها ممکن‌ست اثرات منفی برروی نتیجه‌ی کار داشته باشد؛ به‌ویژه در حالتی که عبارات چند کلمه‌ای (و نه تنها واژگان مستقل) قرارست مورد بررسی قرار بگیرند. اما در پروژه‌ی ما، از آن‌جا که استقلال واژگان پیش‌فرض عملیات قرار گرفته است (بخش ۲.۲.۲: پیچیدگی زبان فارسی و عبارات فارسی)، ایست‌واژه‌ها را بدون دغدغه حذف می‌کنیم.

در زبان فارسی متأسفانه از یک سو مرجعی برای ایست‌واژه‌ها به‌صورت رسمی وجود ندارد و از سوی دیگر، اشتقاق لغات (نظیر تمام صرف‌های فعل «کردن») باعث پیچیده شدن عملیات دستیابی و ساخت این لیست می‌شود. با این وجود تلاش‌های قابل قبولی در این زمینه بر روی ایست‌واژه‌های زبان فارسی شده است.

[Taghva, 03] لیستی شامل ۱۲ ایست‌واژه‌ی فعلی و حدود ۱۵۰ ایست‌واژه‌ی غیرفعلی، بر اساس ۱۸۵۰ سند فارسی‌زبان در داخل و خارج از ایران و در مقوله‌های مختلف ارائه کرده‌است که متأسفانه نرخ/نسبت تکرار این واژگان در مقاله مشخص نشده‌است. [Darrudi, 04] با استفاده از Hamshahri Corpus (متن روزنامه‌ی همشهری از سال‌های ۱۳۷۵ تا ۱۳۸۱) لیستی از ۱۰ واژه‌ی پرتکرار یک الی ۵ حرفی (جمعاً ۵۰ کلمه به‌همراه نرخ دقیق تکرار) ارائه کرده است که تلفیقی از افعال و اسامی هستند. [شیخ‌اسماعیلی، ۸۶] با ارجاع به هر دو منبع فوق، و با استفاده از مجموعه آزمون «محک»[۳۶]، مجموعه‌ای شامل ۲۰۰ ایست‌واژه‌ی فعلی و ۳۵ ایست‌واژه‌ی غیرفعلی تهیه کرده است. [نصیری‌شرق، ۸۷] با استفاده از بررسی بهبودیافته‌ی Hamshahri Corpus و کل داده‌های ویکی‌پدیای فارسی تا ماه سپتامبر ۲۰۰۸ و هم‌چنین متن ۱۰ هزار خبر از خبرگزاری مهر، لیستی از واژگان زبان فارسی (همراه با نسبت تکرار) ارائه کرده است. [Davarpanah, 09] نیز با استفاده از تحلیل Hamshahri Corpus و تعدادی مقاله‌ی مستقل به دامنه لیستی شامل ۱۰۰۰ ایست‌واژه‌ی زبان فارسی (آگاهانه و با بررسی انسانی) تولید کرده است.

به‌عنوان یک جمع‌بندی از این مقالات، ۵۰ واژه‌ی زیر (جدول ۵: ایست‌واژه‌های درنظرگرفته شده برای پردازش متون وبلاگ‌های فارسی) در این مرحله به‌عنوان ایست‌واژه‌ی کلّی تلقّی شده و به‌صورت خودکار در محتوای متن خوانده شده با فاصله‌خالی[۳۷] جایگزین می‌شوند. این واژگان در فایل `stopwords.txt` پروژه

[۳۶] http://ce.sharif.edu/~shesmail/Mahak/

[37] Space

هرکدام در یک سطر قرار دارند. کلماتی که در ادامه و پس از اِعمال فیلترهای مختلف به‌عنوان ایست‌واژه تشخیص داده خواهند شد، در بخش ۳.۳.۶ بررسی می‌شوند.

و	در	به	از	که	این
را	با	برای	آن	یک	خود
تا	بر	نیز	هم	یا	اما
باید	هر	است	مورد	آنها	دیگر
بین	پیش	پس	اگر	همه	یکی
می	ها	های	ای	شده	کرده
شود	دارد	کرد	شد	ی	داد
وی	اند	بود	نمی	همچنین	دهد
کنند	بی				

جدول ۵: ایست‌واژه‌های درنظرگرفته شده برای پردازش متون وبلاگ‌های فارسی

یک تصور بر این است که برخی واژگان که در حالت عادی ایست‌واژه در نظر گرفته می‌شوند، در دامنه‌ی خاص پروژه‌ی ما می‌توانند نقش تعیین‌کننده‌ای داشته باشند. برای مثال وجود واژگانی نظیر «من» و «تو» در صدر واژگان یک وبلاگ، می‌تواند از نظر مفهومی بیان‌گر رابطه‌ی نزدیک نویسنده با خواننده باشد که این نکته به‌نوعی قرارگیری در دسته‌ی وبلاگ‌های «شخصی» را القا می‌کند. اما ما این فرض را در نظر نمی‌گیریم چرا که لحن نویسنده‌ی یک وبلاگ الزاماً متناسب با عنوان وبلاگ نیست. مثلاً یک پزشک ممکن‌ست در وبلاگ تخصصی خود بارها از ضمایر «من» یا «ما» استفاده کند و در گزارش یک گفتگوی تخصصی، بارها لفظ «تو» را بیاورد.

۳.۳ فیلترهای زبانی

این فیلترها پس از فیلترهای ساختاری اِعمال می‌شوند و به تغییرات داخلی کلمه می‌پردازند. فیلترهای زبانی استفاده شده در این پروژه به شرح زیرند. این فیلترها در فایل `analyseBlog.php` اِعمال شده‌اند.

باید توجه داشت که برخی از این فیلترها ممکن‌ست عملکرد «تبدیل نادرست به درست» نداشته باشند (مانند تبدیل قرآن به قران)، منتهی عمدتاً باعث یک‌سان شدن فرم‌های مختلف نوشتار یک واژه می‌شوند.

مطالعه‌ی [استاندارد ۶۲۱۹] برای درک بهتر کاربرد و دلایل استفاده‌ی این فیلترها، خصوصاً نکات مربوط به یونیکد، توصیه می‌شود. هم‌چنین در [صدیقی، ۸۳] به گسترش‌های دیگری بر این نوع فیلترها اشاره شده‌است.

۳.۳.۱ فیلتر زبانی یکم (L01): حذف نویسه‌های نالازم

چنان که در مورد دوم بخش «۲.۲.۲: پیچیدگی زبان فارسی و عبارات فارسی» گفته شد، اِعراب‌گذاری واژگان، چون به‌صورت کامل اجرا نمی‌شود تنها باعث پیچیده شدن واژگان شده و یکسان بودن آن‌ها را زیرسؤال می‌برد. از این رو، در این فیلتر، اِعراب‌ها و نویسه‌های نالازم مرسوم در زبان فارسی را از درون واژگان حذف می‌کنیم. این نویسه‌ها تنها از داخل کلمه حذف می‌شوند و حذف آن‌ها کلمه را نمی‌شکند.

لازم به ذکر مجدّد است که، اگرچه قرار دادن اِعراب در اکثر اوقات باعث رفع ابهام می‌شود، اما این تنوع در نوشتار، لطمه بزرگی به تحلیل و اجرای پروژه‌ی ما خواهد زد. برای مثال نوشتن واژه شکر (شیرین مانند قند) به‌صورت‌های «شِکر (به کسر شین)»، «شکَر (به فتح کاف)»، «شِکَر (به کسر شین و فتح کاف)» و «شکر (بدون اعراب)» باعث می‌شود تا نرخ تکرار این واژه به‌درستی محاسبه نشود.

لیست اِعراب‌ها و نویسه‌هایی که در این فیلتر حذف می‌شوند در جدول ۶ ذکر شده است. این نویسه‌ها در فایل unuseds.txt هر کدام در یک سطر نوشته شده‌اند.

ـً	ـِ	ـُ	ـّ	ـٍ	ـٌ
ـٰ	ـْ	ـٔ	ـٕ	ـٖ	ـٓ

جدول ۶ نویسه‌های نالازم حذف شده توسط فیلتر زبانی یکم

لازم به ذکرست که از آن‌جا که (به‌استناد [فرهنگستان، ۸۴]) همزه‌ی بدون کرسی («ء») در اکثر قریب به اتفاق مواقع در انتهای کلمه می‌آید، آن را نیز حذف می‌کنیم تا نگارشی مانند «انشاء» به «انشا» تبدیل شود.

۳.۳.۲ فیلتر زبانی دوم (L02): حذف پیشوندها و پسوندهای رایج

در گنجینه‌ی واژگانی‌ای که ما پیدا کرده‌ایم، برخی کلمات به‌صورت جمع یا با پسوند و پیشوند استفاده شده و کلمه‌ی مجزایی به‌شمار می‌روند. برای مثال فرض کنید در کلاس «سینما»، واژگانی نظیر «بازیگر»، «بازیگرها»، «بازیگران»، «بازیگری» هر کدام ٪۵ کلّ واژگان را تشکیل دهند. در این حالت هیچ‌کدام از این واژگان نقش تعیین‌کننده و شاخصی ندارند؛ در حالی‌که اگر پسوندهای سه واژه‌ی آخر حذف شوند و همه‌ی تکرارهای آن‌ها به حساب «بازیگر» گذاشته شود، این واژه با نرخ تکرار ٪۲۰ عملاً شاخص خواهد بود.

در همین راستا، برخی پیشوندها و پسوندهای رایج زبان فارسی که با نیم‌فاصله توسط نگارنده متمایز شده‌اند در این فیلتر حذف می‌شوند.

۳.۳.۲.۱ معرفی فاصله‌ی مجازی یا نیم‌فاصله

کاراکتر موسوم به فاصله‌ی مجازی[۳۸] که در گفتار عام به نیم‌فاصله معروف است، همان کد U+200C یونیکد است که به‌نام Zero Width Non-Joiner یا ZWNJ ثبت شده است و هدف آن عدم اتّصال نویسه‌ی قبلی به بعدی‌اش است. برای مثال در واژه‌ی «نامه‌ها» بین دو نویسه‌ی ها، اگر فاصله‌ای نیاید به‌صورت «نامهها» نوشته می‌شود و اگر یک فاصله کامل (blank space) بیاید دو کلمه‌ی «نامه ها» می‌شود. از آن‌جا که در صفحه‌کلید استاندارد فارسی [استاندارد ۲۹۰۱]، این نویسه با نگه‌داری کلید تبدیل (shift) و فشاردادن کلید فاصله (space) قابل تایپ شدن است، بعضاً به آن «شیفت اسپیس» (Shift Space) هم گفته می‌شود.

[۳۸] این نام برای اوّلین بار در استاندارد ۳۳۴۲ ملّی (قابل مشاهده در آدرس: http://www.isiri.org/std/3342.htm) در سال ۱۳۷۲ هش مطرح شده‌است. علامت اختصاری آن نیز PSP (مخفف Pseduo-Space) قرار داده شده است.

۳.۳.۲.۲ پیشوندهای رایج واژگان زبان فارسی

در جدول ۷، لیست پیشوندهای مرسوم زبان فارسی آمده‌است. این پیشوندها تنها درصورتی از واژگان ما حذف می‌شوند که اوّلاً الزاماً در ابتدای کلمه آمده باشند و ثانیاً بلافاصله بعد از آن‌ها، در کلمه یک نویسه‌ی فاصله‌ی مجازی آمده باشد. این پیشوندها در فایل `prefixes.txt` هر کدام در یک سطر آمده‌اند.

پیشوند	مثال	پیشوند	مثال	پیشوند	مثال
می	می‌روم، می‌شود	**هر**	هرطور، هرگونه	**به**	به‌مدّت، به‌اندازه‌ی
نمی	نمی‌روم، نمی‌شود	**هیچ**	هیچ‌گونه	**نا**	ناشایست، نامرد
پر	پررنگ، پرآوازه	**هم**	هم‌میهن، هم‌صدا	**نه**	نه‌چندان
کم	کم‌رنگ، کم‌رو	**همین**	همین‌گونه	**چه**	چه‌طور، چه‌گونه[۳۹]
بی	بی‌رنگ، بی‌مزه	**همان**	همان‌گونه		
با	بامزه، باسابقه	**آن/ان/این**	آن‌طور، این‌که		

جدول ۷: پیشوندهای رایج واژگان زبان فارسی

لازم به ذکر مجدّد است که این پیشوندها تنها درصورتی که بلافاصله بعدشان فاصله‌ی مجازی بیاید حذف می‌شوند چرا که در غیر این‌صورت، حذف آن‌ها ممکن‌ست کلمه را از بین ببرد؛ نظیر حذف «با» در «باران» یا حذف «هر» در «هرمزگان». البته باید توجه داشت که برخی از این پیشوندها، نظیر «بی» یا «نا» عملکرد واژه‌ی بعدی را نقض می‌کنند (نظیر نامرد یا بی‌مزه)، اما همان‌طور که می‌دانیم مقصود ما در این پروژه درک مفهوم نیست، بلکه به‌دنبال ساده‌سازی واژگان هستیم.

۳.۳.۲.۳ پسوندهای رایج واژگان زبان فارسی

مشابه بخش قبل، این‌بار در جدول ۸، لیست پسوندهای مرسوم زبان فارسی آمده‌است. این پسوندها تنها درصورتی از واژگان ما حذف می‌شوند که اوّلاً الزاماً در انتها کلمه آمده باشند و ثانیاً بلافاصله قبل از آن‌ها، در کلمه یک نویسه‌ی فاصله‌ی مجازی آمده باشد. این پیشوندها در فایل `suffixes.txt` هر کدام در یک سطر آمده‌اند.

[۳۹] پیشنهاد [قدسی، ۸۳]

پسوند	مثال	پسوند	مثال	پسوند	مثال
م	کتابم[40]، جزوهم[41]	**ت**	کتابت، جزوهت	**ش**	کتابش، جزوهش
مان	خانه‌مان	**تان**	خانه‌تان	**شان**	خانه‌شان
ام	خانه‌ام	**ات**	خانه‌ات	**اش**	خانه‌اش
هایم	جزوه‌هایم	**هایت**	جزوه‌هایت	**هایش**	جزوه‌هایش
هایمان	نامه‌هایمان	**هایتان**	نامه‌هایتان	**هایشان**	نامه‌هایشان
ایم	رفته‌ایم	**اید**	رفته‌اید	**اند**	رفته‌اند
یم	صاحب‌اختیاریم	**ید**	صاحب‌اختیارید	**ند**	صاحب‌اختیارند
ی	خانه‌ی ما	**ای**	عجب خانه‌ای بود	**هایی**	خانه‌هایی مجلّل
ها	خانه‌ها	**های**	خانه‌های لوکس	**هائی**	خانه‌هائی لوکس
است	دیوانه‌است	**تر**	ساده‌تر	**که**	آن‌که، به‌طوری‌که
ست	مرده‌ست	**ترین**	ساده‌ترین	**چه**	کتاب‌چه، بازی‌چه
شدن	دیوانه‌شدن	**شده**	خوانده‌شده	**شونده**	جاری‌شونده
کردن	دیوانه‌کردن	**کرده**	جاری‌کرده	**کننده**	جاری‌کننده

جدول ۸: پسوندهای رایج واژگان زبان فارسی

لازم به‌ذکرست که صحّت رسم‌الخط‌ها و مثال‌های زده شده در جدول فوق بحث این پایان‌نامه نیست و مثال‌های فوق صرفاً از روی موارد مشاهده‌شده در وبلاگ‌ها گردآوری شده است.

۳.۳.۲.۴ حذف فاصله‌ها و اتصالات مجازی

در پایان این فیلتر، از آن‌جا که تلاش‌ها ما برای کمک‌گیری از وجود فاصله‌ی مجازی پایان یافته است، تمامی فواصل مجازی (ZWNJ یونیکد، PSP استاندارد فارسی) و تمامی اتصالات مجازی (ZWJ یونیکد، PCP استاندارد فارسی)، مانند اعراب‌های بخش ۳.۳.۱ از متن واژه حذف می‌شود.

[40] ضمایر ملکی متصل منفرد گاهی به اشتباه با فاصله‌ی مجازی نوشته می‌شوند. البته [قدسی، ۸۳] این فاصله‌ی مجازی را در مورد ضمایر ملکی جمع مجاز می‌داند.

[41]

۳.۳.۳ فیلتر زبانی سوم (L03): تبدیل نویسه‌های هم‌سان و یکسان‌سازی نوشتار

املای دوگانه یا چندگانه‌ی یک واژه از دیگر مشکلات تجزیه و تحلیل زبان فارسی است. در این فیلتر سعی می‌کنیم تا با تبدیل برخی نویسه‌ها به یک نویسه‌ی ثابت، مشکلات ناشی از دوگانه‌نویسی را برطرف کنیم.

لیست این تبدیلات در جدول ۹ آمده است.

نویسه‌ی سابق	کد یونیکد قبلی	توضیحات	نویسه‌ی جدید	کد یونیکد جدید	دلیل
آ	U+0622	آ با مدّ (کلاه)	ا	U+0627	ساده‌سازی
ة	U+0629	تای تأنیث	ت	U+062A	یکسان‌سازی
ؤ	U+0624	همزه با کرسی و	ئ	U+0626	یکسان‌سازی
أ	U+0623	همزه با کرسی الف	ئ	U+0626	یکسان‌سازی
إ	U+0625	همزه با کرسی الف زیرین	ئ	U+0626	یکسان‌سازی

جدول ۹: جدول تبدیلات زبانی در راستای رفع برخی مشکلات دوگانه‌نویسی

درست است که یکسان‌سازی کرسی همزه‌های متفاوت (ؤ، أ، إ، ئ) به کرسی «ی» (به‌صورت ئ) ممکن‌ست باعث املای نادرست برخی واژگان بشود (نظیر متأسفانه ← متئسفانه) اما این‌کار باعث یکسان‌سازی نگارش‌های مختلف مربوط به همزه می‌شود. مثال‌هایی از دوگانه‌نوشتن‌های مربوط به همزه عبارتند از: «مسئول» و «مسؤول»، «هیأت» و «هیئت»، «مسأله» و «مسئله».

۳.۳.۴ فیلتر زبانی چهارم (L04): حذف کلمات با محتوای غیرالفبایی

در نهایت کار، ممکن‌ست برخی از واژگان هم‌چنان نویسه‌ای غیر از نویسه‌های الفبای فارسی داشته باشند. دلیل این امر می‌تواند به استفاده از نویسه‌های غیرمتداول (نظیر نوشتن نام اصلی یک فیلم به زبان چینی)، اشکالات نوشتاری نویسنده (مثلاً استفاده از نویسه‌های نادرست زبان‌های نزدیک به فارسی، نظیر اردو به‌جای معادل فارسی) یا اشکالات فنّی تجزیه‌کننده (مثلاً خواندن ناقص یک نویسه‌ی دو بایتی یونیکد) برگردد. در این پروژه تماماً سعی شده که این مشکلات تا حد امکان رفع شوند؛ بااین‌حال، اگر موردی هم‌چنان با اشکال باقی‌مانده باشد، در این مرحله آن مورد (واژه) به‌کلّی حذف می‌شود.

۳.۳.۵ فیلتر زبانی پنجم (L05): ریشه‌یابی و اشتقاق احتمالی؟

در بخش فیلتر زبانی دوم (بخش ۳.۳.۲) دیدیم که حذف پیشوندها و پسوندها می‌توانند به ساده‌ترشدن کلمات ما کمک کنند. در آن بخش ما پیشوندها و پسوندها را الزاماً با وجود کاراکتر فاصله‌ی مجازی بررسی می‌کردیم؛ حال‌آن‌که در مورد کلماتی که پیشوندشان به‌خودی‌خود به نویسه‌ی آخرِ واژه نمی‌چسبند، لزومی به استفاده از فاصله‌ی مجازی نیست. برای مثال درمورد کلمه‌ی «کتابتان»، حذف پسوند نمی‌تواند به‌سادگی صورت بپذیرد.

ایده‌ای که در این مرحله مطرح می‌شود تجزیه‌ی واژه بر اساس نرخ تکرار اجزای احتمالی‌اش است. به‌عنوان مثال اگر «تان» را پسوند موجود در انتهای یک کلمه بگیریم، می‌توانیم بررسی کنیم که آیا کلمه‌ای که با حذف این پسوند به‌دست می‌آید از نرخ تکرار بالایی برخوردار است یا نه؟ با این‌کار، به‌احتمال قوی واژه‌ی «کتابتان» به «کتاب» تبدیل می‌شود و واژه‌ی «عربستان»، مطابق خواست ما، به «عربس» تبدیل نمی‌شود. همین ایده را می‌توان برای شکستن کلمه‌های طولانی نیز انجام داد؛ چراکه اکثر اشتباهات نگارشی، مربوط به عدم استفاده از فاصله است. به‌عنوان مثال عبارتی طولانی و کم‌تکرار مانند «یکصدموضوع» (که چون بدون فاصله نوشته شده است، یک واژه تشخیص داده شده است) را می‌توان از تمام فاصله‌های بین نویسه‌ای‌اش بُرید و برشی که دو کلمه‌ی با نرخ تکرار مناسب می‌سازند را انتخاب کرد. در این مثال، برش‌ها به‌صورت «ی+کصدموضوع»، «یک+صدموضوع»، «یکص+دموضوع»، ... و «یکصدموضو+ع» اند که از میان آن‌ها «یکصد+موضوع» چون دو کلمه‌ی پرتکرار را می‌سازند، برگزیده می‌شود.

اما باید درنظر داشت که این ایده هنوز خام است و احتیاج به بررسی دقیق‌تری دارد. برای مثال واژه‌ای نظیر «قهرمان» را درنظر بگیرید. وجود این واژه در متنی که در آن لفظ «قهر» مکرّر است، باعث تجزیه‌ی تک‌واژه‌ی «قهرمان» به «قهر» و «مان» می‌شود که کاملاً نادرست است. همین‌طور تجزیه‌ی ناخواسته‌ی «داستان» به «داس» و «تان». در صورتی که ایست‌واژه‌ها حذف نشده باشند، این تجزیه‌ی نادرست پررنگ‌تر هم می‌شود؛ مثلاً تجزیه‌ی «درمان» به «در» و «مان».

حتی اشکال ساده‌ی این تجزیه (نظیر «ی» پایانی» نیز می‌توانند تشخیص نادرستی را باعث شوند. برای مثال تجزیه‌ی «ماهی» (در آب) به «ماه» (در آسمان) + «ی».

از همین رو و به‌دلیل تأثیر منفی روی عملکرد کار ما بگذارد، در این پروژه این تجزیه را اِعمال نمی‌کنیم. با این حال معتقدیم که تأمل در این‌گونه تجزیه‌ها، می‌تواند از نکات مهم یک تجزیه‌گر کامل زبان فارسی باشد.

۳.۳.۶ فیلتر زبانی ششم (L06): حذف ایست‌واژه‌های جدید

به‌عنوان آخرین فیلتر زبانی، ایست‌واژه‌های جدیدی که پس از اِعمال فیلترهای قبلی کشف شده‌اند را در این مرحله شناسایی کرده و به لیست ایست‌واژه‌های شناخته شده (بخش ۳.۲.۵) اضافه می‌کنیم.

برای این منظور، ۲۰۰ واژه‌ی پرکاربرد هر کدام از ۴۷ کلاس‌مان را انتخاب کرده و آن‌ها را بررسی می‌کنیم. این واژه‌ها به‌همراه تعداد کلاس‌هایی که در ۲۰۰ واژه‌ی پرکاربردشان آمده‌اند (از ۴۷ کلاس؛ در حداقل نیمی آمده‌اند) در جدول ۱۰ ذکر شده‌اند.

از میان این واژگان ما ۶۲ واژه‌ی اوّل که هر کدام در بیش از ۳۷ کلاس (۸۰ درصد کلاس‌ها) ظاهر شده‌اند را برمی‌گزینیم و آن‌ها را ایست‌واژه‌های جدید می‌گیریم. این واژه‌ها در جدول ۱۱ لیست شده‌اند و در فایل `stopwords2.txt` نیز هر کدام در یک سطر آمده‌اند.

کلیه کارهای این فیلتر در فایل `findNewSWs.php` قابل مشاهده‌اند. نتایج این فیلتر (ایست‌واژه‌های جدید) در فایل `RssReader.php` استفاده شده است.

واژه	نرخ	واژه	نرخ	واژه	نرخ	واژه	نرخ	واژه	نرخ	واژه	نرخ
کردن	47	کنید	47	اول	44	یعنی	38	کسی	30	میان	26
دست	47	ما	47	همین	44	بدون	38	وبلاگ	30	عکس	26
زمان	47	ایران	47	کنیم	44	حتی	37	کشور	30	نشان	26
ولی	47	کند	47	دارند	44	برنامه	37	البته	30	کم	25
چه	47	نظر	47	توجه	43	جدید	37	سایت	29	امروز	25
حال	47	نام	47	هیچ	43	سر	36	اب	29	خوب	25
روی	47	اینکه	47	بار	43	وقتی	36	ایجاد	28	تواند	25
چند	47	شدن	46	داشت	42	بخش	35	کتاب	28	دوست	24
نیست	47	بیشتر	46	استفاده	42	تر	35	بالا	28	انتخاب	24
بسیار	47	زیر	46	ماه	42	کنم	35	مثل	27	حسین	24
بعد	47	تمام	46	فقط	42	توسط	34	خدا	27	مانند	24
وجود	47	تنها	46	عنوان	42	مردم	34	شب	27	شرکت	24
دو	47	انجام	46	زندگی	41	شود	34	جهت	27	امام	24
کار	47	راه	46	گرفته	41	خیلی	34	علی	27		
قرار	47	هستند	46	نه	41	هایی	33	مختلف	27		
شما	47	بوده	45	باز	41	توان	33	چرا	27		
من	47	سه	45	شوند	40	اید	33	دانلود	27		
روز	47	صورت	45	چون	40	جهان	32	باشند	26		
سال	47	ادامه	45	گفت	38	گروه	31	بودن	26		
او	47	تو	44	بزرگ	38	همان	31	ترین	26		

جدول ۱۰: لیست‌واژگانی که در میان ۲۰۰ واژه پرتکرار حداقل نیمی از ۴۷ کلاس ظاهر شده‌اند

کردن	دست	زمان	ولی	چه	حال	روی	چند	نیست
بسیار	بعد	وجود	دو	کار	قرار	شما	من	روز
سال	او	کنید	ما	ایران	کند	نظر	نام	اینکه
شدن	بیشتر	زیر	تمام	تنها	انجام	راه	هستند	بوده
سه	صورت	ادامه	تو	اول	همین	کنیم	دارند	توجه
هیچ	بار	داشت	استفاده	ماه	فقط	عنوان	زندگی	گرفته
نه	باز	شوند	چون	گفت	بزرگ	یعنی	بدون	

جدول ۱۱: لیست ایست‌واژه‌های جدید (سری دوم)

۳.۴ فیلترهای پایانی

۳.۴.۱ فیلتر پایانی یکم (F01): حذف واژگان با نرخ تکرار ۱؟

علی‌رغم اِعمال تمام فیلترهای قبلی، هم‌چنان تعدادی از واژگان به‌دلیل املای نادرست یا عدم رعایت فاصله‌گذاری در لیست واژگان هر وبلاگ یافت می‌شوند. از آن‌جا که این اشتباه‌های نوشتاری غالباً به یک‌صورت تکرار نمی‌شوند، معمولاً نرخ تکرار این واژگان در حدّ یک مورد باقی می‌ماند.

با علم به این موضوع، ایده‌ای که در این‌جا مطرح می‌شود این‌ست که واژه‌هایی که نرخ تکرارشان تنها یک مورد است را از گنجینه واژگانی حذف کنیم. منتهی با بررسی موردی در می‌یابیم که در هر وبلاگی حدود نیمی از واژگان نرخ تکراری برابر یک دارند. همچنین در گنجینه‌ی تمامی واژگان تمام وبلاگ‌ها که مشتمل بر ۲۸۰هزار واژه می‌شود[۴۲] نیمی از کلمات نرخ تکراری برابر یک دارند.

از این رو، ایده‌ی حذف واژگان کم‌تکرار منتفی می‌شود، چرا که این‌کار تأثیر منفی شایانی دارد و قسمت عمده‌ای از اطلاعات را از بین می‌برد. با این‌حال اگر حجم اسناد ما بالاتر می‌بود، مثلاً به‌جای ۱۰ پست آخر، شامل تمام وبلاگ می‌شد، احتمال داشت که تعداد واژگان با نرخ تکرار یک (به‌سبب اشتباهات نگارشی) کم بوده و قابل حذف باشند.

۳.۴.۲ فیلتر پایانی یکم (F02): نوشتن در خروجی به فرمت مناسب

به‌عنوان آخرین فیلتر، در این بخش تمام RSSهای تمام وبلاگ‌ها خوانده شده و با اِعمال فیلترهای پیشین در فایلی با نام وبلاگ در زیرشاخه‌ای با شماره کلاس وبلاگ از شاخه‌ی DATA نوشته می‌شوند. در این فایل‌ها هر کلمه در یک سطر می‌آید و در آن سطر بعد از خود کلمه، با یک فاصله‌ی tab، نرخ تکرار آن کلمه می‌آید. این شیوه کمک می‌کند تا با استفاده از تابع `parse_ini_file` زبان PHP بتوانیم در یک خط به این داده‌ها دسترسی پیدا کنیم.

[۴۲] فایل all.txt در شاخه words تمامی این ۲۸۰ هزار واژه را همراه با نرخ تکرارشان در بر دارد که بالطّبع تعداد کثیری از آن‌ها واژه‌ی ریشه‌ای نیستند و مشتقات، ترکیب‌های یا نگارش‌های نادرست واژه‌ها را در بر گرفته‌اند.

۴ فصل چهارم)

معرفی و مقایسه‌ی معیار پیشنهادی برای دسته‌بندی

در پایان فصل قبل، از هر وبلاگ یک لیست واژگان با نرخ تکرار به‌دست آمد. اکنون در این فصل شباهت این لیست‌ها با هم، برطبق معیار پیشنهادی ما و الگوریتم‌های شناخته‌شده‌ی دیگر به‌دست آورده می‌شود و از آن به‌عنوان معیاری برای دسته‌بندی[۴۳] وبلاگ‌ها استفاده می‌شود.

بخش ۱ این فصل به معرفی نمونه‌های آزمایشی و آموزشی و نحوه‌ی تقسیم ۹۰۰۰ وبلاگ جمع‌آوری شده به این دو قسمت اختصاص دارد. سپس با معرفی الگوریتم‌های Jaccard و Cosine Similarity با وزن‌دهی tf-idf، عملکرد آن‌ها بر روی نمونه‌های ما سنجیده خواهد شد. در نهایت الگوریتم پیشنهادی ما معرفی شده و عملکرد آن با عملکرد دو الگوریتم قبلی مقایسه می‌شود. نتیجه‌گیری و جمع‌بندی این روش‌ها، بخش پایانی این فصل را تشکیل می‌دهد.

۴.۱ معرفی نمونه‌های آزمایشی و آموزشی

داده‌های جمع‌آوری شده، چنان‌چه که در فصل دوم توضیح داده شدند، قریب ۸۵۰۰ وبلاگ به همراه دسته‌ی انتخاب‌شده توسط نویسنده‌ی وبلاگ هستند. ما انتخاب نویسنده‌ی وبلاگ را ملاک قرار داده و قصد داریم نمونه‌ها را بر طبق آن شناسایی کنیم. برای همین منظور، ما از هر کدام از ۴۷ کلاس معرفی شده، ۲۰ وبلاگ (حدود ۱۰ درصد) را به‌عنوان نمونه‌ی آزمایشی در نظر گرفته و سایر وبلاگ‌های دسته (۱۶۰=۱۸۰-۲۰ عدد) را نمونه‌ی آموزشی تلقّی می‌کنیم. از این رو، شیوه‌ی آماری[۴۴] ما به‌نوعی درستی‌سنجی ضربدری[۴۵] تک‌دوره[۴۶] است.

[43] Classification
[44] Statistical Method
[45] Cross-validation
[46] Single round

با کمک فایل پردازشی `makeTrainTest.php`، لیست ۹۴۰=۲۰×۴۷ وبلاگ آزمایشی (که می‌بایست دسته‌بندی شوند و نتایج الگوریتم‌های مختلف روی آن‌ها را مشاهده کنیم)، به همراه ۷۵۲۰=۱۶۰×۴۷ وبلاگ آموزشی (که دسته‌های تعیین‌شده‌ی آن‌ها، از ابتدا در اختیار الگوریتم دسته‌بندی قرار می‌گیرد) در فایل متنی `traintest.txt` گردآوری شده‌است.

در این فصل دو الگوریتم رایج را بر روی این نمونه‌های آزمایشی بررسی می‌کنیم و نهایتاً با اجرای الگوریتم خودمان، نتایج عملکرد این ۳ الگوریتم را مقایسه می‌کنیم.

لازم به ذکرست که الگوریتم‌های Jaccard و Cosine Similarity نیاز به مقایسه‌ی یک وبلاگ نمونه با یک کلاس از بلاگ‌ها داشته‌اند و از این رو ما نیز الگوریتم خود را در حالت «مقایسه وبلاگ با کلاس» اجرا کرده‌ایم.

۴.۲ معرفی و ارزیابی الگوریتم Jaccard

پول جکارد (Paul Jaccard)[۴۷] محقق فرانسوی/آلمانی اوایل قرن بیستم، رابطه‌ی ساده‌ی زیر را برای نزدیکی/شباهت دو مجموعه‌ی A و B پیشنهاد داده است:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$ فرمول (۱.۴)

برای هر کدام از ۹۴۰ وبلاگ نمونه‌ی آزمایشی (به‌عنوان مجموعه‌ی A)، ما این معیار را با تمامی گنجینه‌های واژگانی نمونه‌های آموزشی هر کدام از ۴۷ کلاس (به‌عنوان مجموعه‌ی B، جداگانه) محاسبه کردیم و از این بین کلاس B ای که $J(A,B)$ در آن بیش‌تر را به‌عنوان کلاس مناسب برای وبلاگ A برگزیدیم.

البته از آن‌جا که تعداد کلمات کلاس‌های مختلف بین ۱۷ تا ۴۰ هزار واژه متفاوت بوده است، ما تنها ۱۵هزار واژه‌ی پرتکرارتر هرکدام از دسته‌ها را انتخاب کرده‌ایم تا اندازه‌ی مجموعه‌های مرجع (B) ما در این معیار دخیل نباشد.

[۴۷] به نقل از http://en.wikipedia.org/wiki/Paul_Jaccard

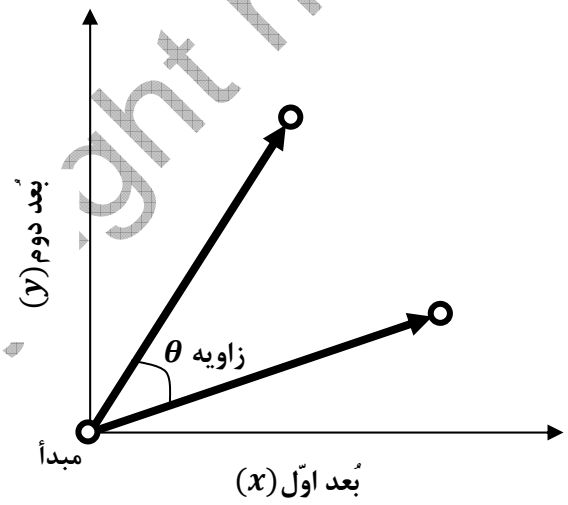
نتیجه‌ی نهایی عملکرد این الگوریتم صحّت ٪۳۶٫۲۷ در پیش‌بینی بوده‌است که به‌طور مفصّل در نمودار ۳ و نمودار ۴ نمایش داده است. این الگوریتم ساده، همان‌طور که انتظار می‌رفت عمل‌کرد ضعیف‌تری نسبت به ۲ الگوریتم دیگر دارد.

البته از مزایای دیگر این الگوریتم امکان مقایسه‌ی تک به تک وبلاگ‌ها در معیار تعریف‌شده است. منتهی اندازه‌گیری دقیق این معیار نیازمند تجزیه‌ی مناسب وبلاگ به واژگان است.

این الگوریتم در فایل `testing.php` پیاده‌سازی شده است.

۴.۳ معرفی و ارزیابی الگوریتم Cosine Similarity با وزن‌دهی tf-idf

الگوریتم دومی که برای دسته‌بندی مورد ارزیابی و مقایسه قرار گرفت، الگوریتم Cosine Similarity بود. در حالت این الگوریتم هر سند را که شامل n واژه باشد، یک بردار در فضای n بعدی تصوّر می‌کند و شباهت دو بردار را متناسب با زاویه‌ی بین دو بردار می‌داند. به‌عنوان مثال، اگر فرض کنیم فضای ما به‌صورت ساده تنها دو بعد دارد، شکل زیر میزان نزدیکی دو بردار را براساس زاویه‌ی بین‌شان نشان می‌دهد.



شکل ۱: نمایش دو سند در فضای نمونه‌ی دو بعدی و زاویه‌ی θ بین‌شان

با این تعریف، می‌توان نزدیکی دو نقطه (یا به‌عبارتی بردار، از مبدأ) نظیر A و B را با عنوان $Sim(A,B)$ چنین تعریف کرد:

$$Sim(A,B) = Cos(\theta) = \frac{A \cdot B}{|A||B|}$$ فرمول (۲.۴)

در عمل، اما به‌جای قرار دادن معیار نرخ تکرار برای هر واژه (هر بُعد)، از نرخ‌های ظریف‌تری و معنادارتری استفاده می‌شود. به‌عبارت دقیق، اگر بُعد i اُم سند (نقطه‌ی) A را (که متناسب با واژه‌ی i اُم کلّ گنجینه ات) $w_{i,A}$ بنامیم، در این‌صورت می‌توان $Sim(A,B)$ را به‌صورت زیر بسط داد:

$$Sim(A,B) = \frac{\sum_i (w_{i,A} \times w_{i,B})}{\sqrt{\sum_i (w_{i,A})^2}\sqrt{\sum_i (w_{i,B})^2}}$$ فرمول (۳.۴)

این رابطه اکنون شباهت کسینوسی را کاملاً منطبق بر وزن‌دهی w نشان می‌دهد. یکی از رایج‌ترین معیارهای وزن‌دهی، معیار وزن‌دهی tf-idf است که هم متناسب با نرخ تکرار واژه در سند است و هم متناسب با اهمیّت کلمه. منظور از اهمیّت کلمه این است که به‌عنوان مثال کلمه‌ای نظیر «است» یا «و» در زبان فارسی نرخ تکرار بالایی دارد و رخداد آن‌ها نمی‌تواند تأثیر دقیقی در ارزیابی داشته باشد.

۴.۳.۱ معرفی وزن‌دهی tf-idf

با این وصف، معیار ارزیابی w بر طبق روش tf-idf به‌صورت زیر قرار می‌گیرد:

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$ فرمول (۴.۴)

که $n_{i,j}$ تعداد دفعات تکرار عبارت i در سند j اُم است. $\mathrm{tf}_{i,j}$ که مخفف Term-Frequency است، به‌نوعی نرخ (درصد) رخداد یک واژه در عبارت را تعیین می‌کند.

$$\mathrm{idf}_i = \log \frac{|D|}{1 + |\{d: t_i \in d\}|}$$ فرمول (۵.۴)

که D مجموعه‌ی تمام سندهای ما در گنجینه است و $|\{d: t_i \in d\}|$ تعداد سندهایی را مشخص می‌کند که عبارت i در آن‌ها حداقل یک‌بار آمده باشد ($n_{i,j} \neq ۰$). به‌عبارتی idf_i بیان‌گر نرخ رایج‌بودن عبارت i اُم است که چون لگاریتم از آن گرفته می‌شود، شدت تکرار را به‌صورت عکس-نمایی نشان می‌دهد.

از آن‌جا که ممکن‌ست عبارت i مورد مقایسه‌ی ما کلاً در هیچ سندی (از اسناد گنجینه) وجود نداشته باشد، به مخرج یک واحد می‌افزاییم.

نهایتاً معیار وزن‌دهی ما به‌صورت زیر تعریف می‌شود که در فرمول (۳.۴) مورد استفاده قرار می‌گیرد.

$$w_{i,j} = (\text{tf-idf})_{i,j} = \text{tf}_{i,j} \times \text{idf}_i \qquad \text{فرمول (۶.۴)}$$

۴.۳.۲ ارزیابی و مقایسه عملکرد

نتیجه‌ی نهایی عملکرد این الگوریتم صحّت ۴۸٫۷۲٪ در پیش‌بینی بوده‌است که به تفصیل در نمودار ۳ و نمودار ۴ نمایش داده است. این الگوریتم، عمل‌کرد مناسبی نسبت به الگوریتم قبلی دارد و محاسبات پیچیده‌ای ندارد. ضمناً می‌توان با تعریف معیارهای دیگر وزن‌دهی (نظیر دادن توان‌های بالاتر از یک به tf یا idf) همچنان از نزدیکی کسینوسی استفاده کرد.

خوش‌بختانه در این الگوریتم تأثیر رایج بودن یا نبودن واژه‌ها با کمک معیار idf لحاظ شده است. به‌عنوان مثال دو واژه‌ی نمونه‌ی «کنند» و «نورپردازی» را در نظر می‌گیریم که در دو سند (وبلاگ) A و B ظاهر شده‌اند. تفاوت نرخ تکرار این دو واژه، بین دو سند به‌صورت اختلاف یک بُعد در نظر گرفته می‌شود که با ضریب idf واژه‌ی مورد نظر، این اختلاف ضریب می‌گیرد. به بیان ساده‌تر، از آن‌جا که idf واژه‌ی «کنند» تقریباً نزدیک به صفر است و idf واژه‌ی «نورپردازی» (به‌دلیل رخداد کمتر در بین کل اسناد) بالاتر است، وزن‌دهی w روی «نورپردازی» بزرگ‌تر و حساس‌تر از «کنند» است و این مسئله عیناً مطلوب ماست.

اما نکته‌ی قابل توجه در این الگوریتم که ما را به‌سوی الگوریتم سوم خودمان رهنمون می‌سازد، در نظر نگرفتن نرخ تکرار واژه‌ها در کلیه اسناد است. این مسئله از آن‌جا ناشی می‌شود که در مخرج رابطه‌ی محاسبه‌ی idf_i، ما تنها تعداد اسناد که عبارت i را دارند می‌شماریم، و کاری به نرخ تکرار i در آن اسناد نداریم. با این

تفسیر، idf برای دو واژه‌ی «عکس» و «دارند» نزدیک به هم شوند، چرا که هر دوی این واژه‌ها در اکثر وبلاگ‌ها حداقل یک‌بار آمده‌اند. حال‌آن‌که «عکس» واژه‌ای است که تکرّر آن در یک وبلاگ می‌تواند نشان‌دهنده‌ی اختصاص آن وبلاگ به دسته‌ی «عکاسی» باشد، درحالی‌که «دارند» به‌هیچ‌وجه اطلاعی راجع به دسته‌ی وبلاگ به ما نمی‌دهد.

لازم به ذکر است که اجرای این الگوریتم در فایل `testing.php` پیاده‌سازی شده است.

۴.۴ الگوریتم پیشنهادی ما

با بیان ایرادی که بر تعیین معیار idf وارد شد، در این بخش ما یک روش وزن‌دهی جدید پیشنهاد می‌دهیم و پارامترهای آن را برای این پروژه تعدیل می‌کنیم.

۴.۴.۱ رهیافت اوّلیه

چنان که گفته شد، مشکل اصلی ما با idf در عدم تأثیرگذاری نرخ تکرار واژه در اسناد مختلف است و تنها وجود یا عدم وجود یک واژه در یک سند، مخرج فرمول idf را مشخّص می‌کند. نتیجه‌ی این اهمال در تأثیر نرخ تکرار نیز باعث عدم تمایز بین یک واژه با فرکانس بسیار متغیّر (در برخی دسته‌ها زیاد، در برخی کم) با یک واژه با فرکانس ثابت - که عموماً فاقد محتوای مفید برای دسته‌بندی است- می‌شود.

از این رو، ما معیار تصمیم‌گیری برای هر واژه را بر حسب تفاوت نرخ تکرارش، با نرخ تکرار عمومی (در کلیه اسناد گنجینه) انتخاب می‌کنیم. به عبارت دقیق‌تر ما α_i را نرخ متوسط و مورد انتظار واژه‌ی i می‌گیریم و آن را بدین‌گونه محاسبه می‌کنیم:

$$\alpha_i = \frac{\sum_j n_{i,j}}{\sum_i \sum_j n_{i,j}}$$ فرمول (۷.۴)

که $n_{i,j}$ تعداد دفعات تکرار عبارت i در سند j اُم است. با این تعریف ما می‌توانیم اختلاف نرخ تکرار هر واژه‌ی خاص در یک سند را با این نرخ متوسط مقایسه کنیم و به شدّت غیرعادی بودن آن نرخ پی ببریم. سپس با مقایسه‌ی همین شدّت میان دو سند، می‌توانیم یک معیار برای نزدیکی دو سند داشته باشیم.

به‌عنوان مثال اگر نرخ تکرار متوسط واژه‌ای نظیر «بازیگر» برابر ۰٫۰۰۰۱۰۴= $\alpha_{\text{بازیگر}}$ باشد و این واژه در دو سند A و B نرخ تکراری نزدیک به ۰٫۰۰۱ (حدود ۱۰ برابر نرخ تکرار متوسّط) داشته باشد، آن‌گاه می‌توان انتظار داشت که دو سند A و B از نظر محتوایی شباهت خاصی با هم داشته باشند.

اکنون با کمک معیار α ما می‌توانیم معیار برای غیرعادی بودن رخدادن یک واژه در یک سند تعریف کنیم و این معیار را برای مقایسه‌ی دو سند ملاک قرار بدهیم. این معیار به‌صورت دقیق چنین تعریف می‌شود:

$$f_{i,j} = \frac{\mathrm{tf}_{i,j} - \alpha_i}{\alpha_i}$$ فرمول (۸.۴)

که $\mathrm{tf}_{i,j}$ طبق فرمول (۴.۴) برابر $\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ است. مقدار $f_{i,j}$ عددی بین منفی یک تا مثبت بی‌نهایت است. میزان صفر برای این مقدار نشان‌دهنده‌ی رخداد کاملاً عادی (برابر میانگین) است؛ هر چه میزان از صفر کمتر باشد، بیانگر رخداد کمتر از میانگین و هر چه میزان بیش‌تر از صفر باشد برابر رخداد بیش از میانگین (α) است. برای مثال میزان $f_{i,j} = ۱$ به معنای رخداد دوبرابر میانگین واژه‌ی i در سند j است.

۴.۴.۲ معرفی فرمول میزان شباهت پیشنهادی

حال با استفاده از میزان f تعریفی برای شباهت دو سند A و B (که آن را با $S(A,B)$ نشان می‌دهیم) ارائه می‌کنیم. اگر واژه‌های متفاوت سند X را C_X بنامیم، در این‌صورت

$$S(A,B) = \frac{\sum_{i \in C_A \cup C_B} \mathrm{Sign}(f_{i,A}, f_{i,B}) \times f_{i,A} \times f_{i,B} \times \alpha_i}{|C_A \cup C_B| \times \log(\sum_{i \in C_A} \mathrm{tf}_{i,A}) \times \log(\sum_{i \in C_B} \mathrm{tf}_{i,B})}$$ فرمول (۹.۴)

در فرمول فوق تابع Sign معرّف اهمیّت ما به یکسو بودن $f_{i,A}$ و $f_{i,B}$ است. به‌صورت تجربی و با آزمایش‌های متعدّد مقادیر زیر برای تابع چند مقداره‌ی Sign درنظرگرفته شده است.

$$\text{Sign(X,Y)} = \begin{cases} ۱۰, & X \geq ۱ \text{ و } Y \geq ۱ \\ ۱, & X < ۱ \text{ و } Y < ۱ \\ -۰/۱, & \text{else} \end{cases}$$

فرمول (۱۰.۴)

به‌صورت مفهومی، معیار ما چنین می‌گوید که اگر یک واژه در هر دو سند A و B بیش از مقدار میانگین باشد، ضریب تأثیر برابر ۱۰ بوده و ما حاصل‌ضرب نسبت نرخ تکرار واژه به نرخ میانگین (f) را برای هر دو سند محاسبه کرده و در آن ضریب تأثیر ضرب می‌کنیم. در صورتی که واژه در هر دو سند نرخ تکراری کمتر از میانگین داشته باشد، در آن‌صورت ضریب تأثیر ۱ بوده و در صورتی که واژه در یکی از اسناد بیش از میانگین و در دیگری کمتر از میانگین باشد، ضریب تأثیر منفی ۰/۱ است، چراکه نرخ تکرار ناهم‌گون این واژه، عدم شباهت دو سند را می‌رساند.

نهایتاً میانگین محاسبه‌ی فوق به‌ازای تمام واژگان را بر حاصلضرب لگاریتم تعداد کلمات دو سند تقسیم می‌کنیم تا میزانی برای شباهت به‌دست بیاید.

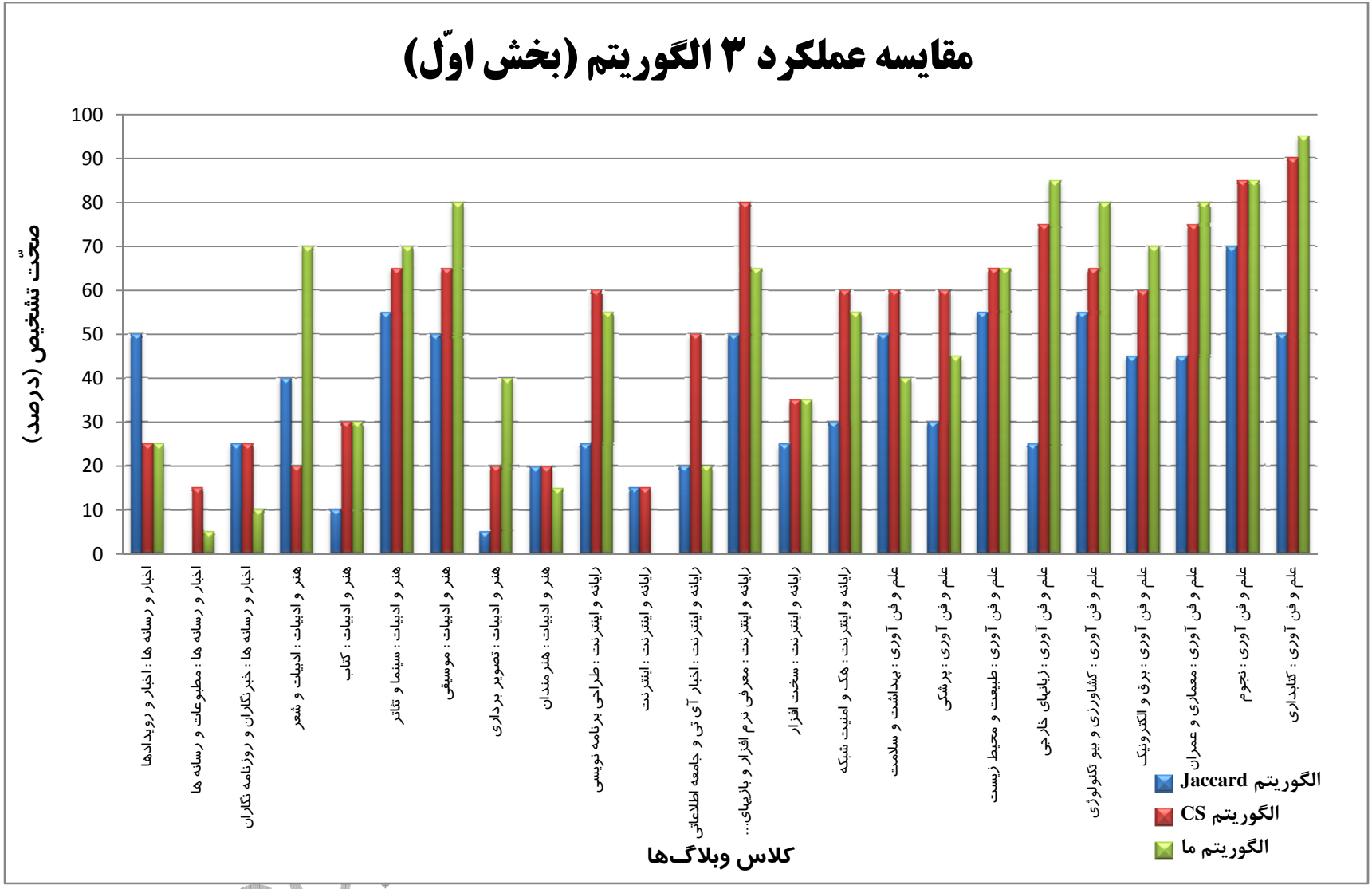
۴.۴.۳ نحوه‌ی اجرا و محاسبه

از آن‌جا که در الگوریتم‌های قبلی، مقایسه را بین یک وبلاگ (نمونه آزمایشی) و ۴۷ گروه دیگر انجام داده بودیم، در این الگوریتم نیز همین رویه را در پیش گرفتیم. منتهی اولاً تنها ۱۰۰۰ واژه‌ی برتر هر دسته را مورد مقایسه قرار دادیم (تا نرخ‌های f نزدیک‌تر به معیار واقعی و متناسب با یک تک وبلاگ بشود). ثانیاً واژه‌های مورد بررسی ما در هر محاسبه‌ی شباهت علاوه بر مجموعه واژه‌های وبلاگ (که به‌فرض k واژه بوده‌اند)، مجموعه‌ی k واژه‌ی پرتکرار دسته نیز بوده‌اند. اجرای این الگوریتم در فایل `gensim.php` قرار داده شده است.

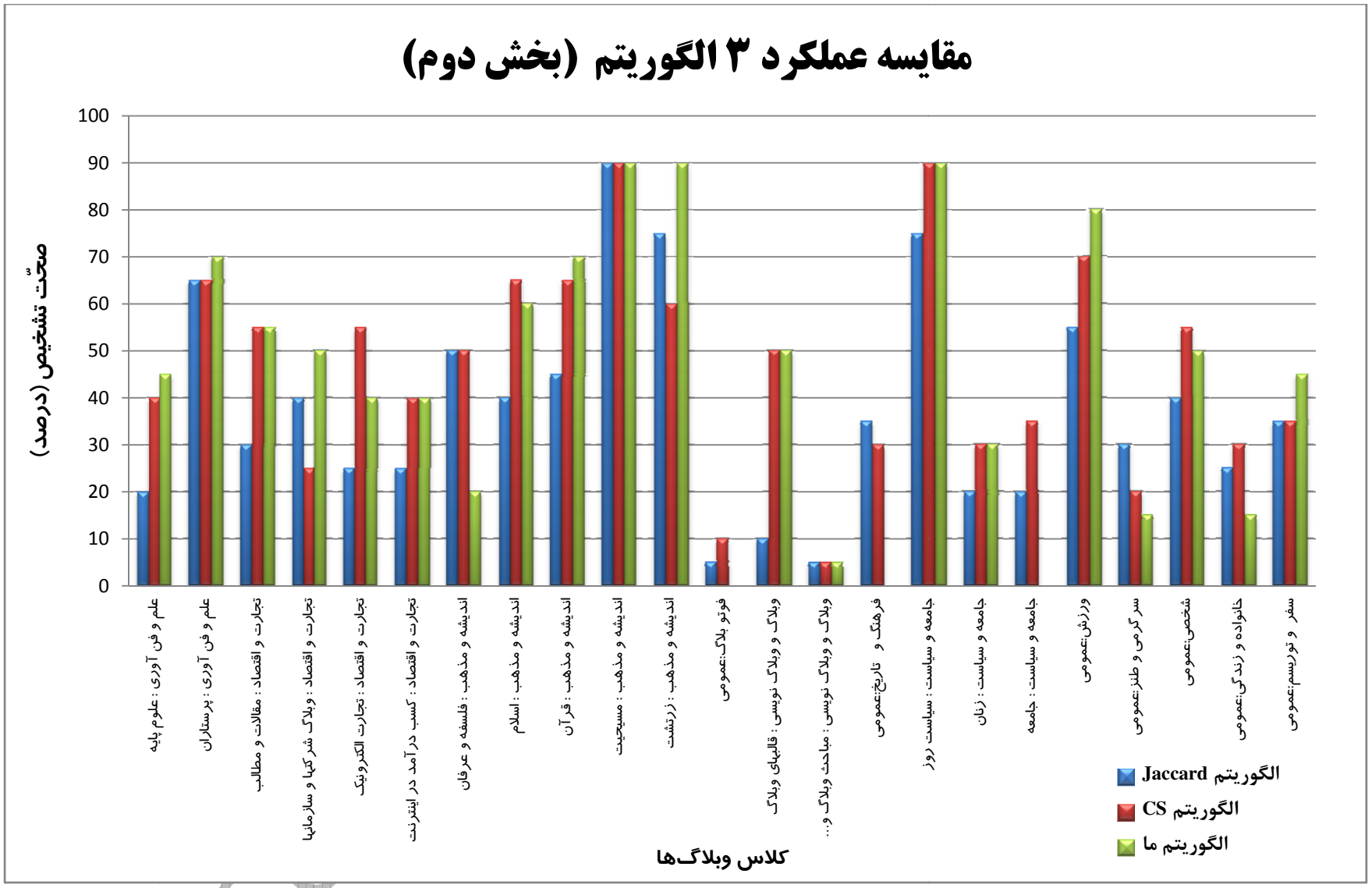
۴.۴.۴ ارزیابی و مقایسه عملکرد الگوریتم ما

نتیجه‌ی نهایی عملکرد این الگوریتم صحّت ۴۷٫۴۴٪ در پیش‌بینی بوده‌است که به‌طور مفصّل در نمودار ۳ و نمودار ۴ نمایش داده است. علی‌رغم این‌که الگوریتم ما عملکردی نزدیک به الگوریتم Cosine Similarity (یا مختصراً CS) دارد (۴۷ درصد در مقایسه با ۴۸ درصد) اما بررسی موردی دسته‌ها نشان می‌دهد که عملکرد این دو الگوریتم در دسته‌های متفاوت، الزاماً یکسان نبوده است.

برای مثال موفقیت الگوریتم ما در دسته‌ی «هنر و ادبیات: ادبیات و شعر» برابر ٪۷۰ بوده است در حالیکه الگوریتم CS در این دسته تنها توانسته ٪۲۰ نمونه‌های آزمایشی را به‌درستی دسته‌بندی کند. اختلاف ٪۴۰ به ٪۲۰ به نفع الگوریتم ما در دسته‌ی «تصویر برداری» نیز از مصادیق دیگر قوّت الگوریتم ما نسبت به CS بوده است. در سمت دیگر، الگوریتم CS در دسته‌ی «اخبار IT» و مشابهاً «فلسفه و عرفان» با برتری ٪۵۰ به ٪۲۰ توانسته بهتر از الگوریتم ما عمل بکند.



نمودار ۳ مقایسه عملکرد ۳ الگوریتم (بخش اوّل)



نمودار ۴ مقایسه عملکرد ۳ الگوریتم (بخش دوم)

۴.۵ جمع‌بندی

در این فصل ضمن معرفی نمونه‌های آموزشی و آزمایشی، دو الگوریتم رایج برای سنجش نزدیکی را معرفی و مقایسه کردیم. سپس روش خودمان را معرفی کردیم و دیدیم که عملکرد آن در حالت میانگین شبیه به عملکرد الگوریتم Cosine Similarity است؛ منتهی این دو الگوریتم در دسته‌های مختلف، عملکرد متفاوتی نشان می‌دهند.

باید توجه داشت که سنجش و مقایسه‌های ما عاری از مشکلات نبوده است. برخی از این مشکلات عبارتند از:

۱. علی‌رغم این‌که در فصل سوم، راهکارها و فیلترهایی برای رفع اشکالات نگارشی و یکسازی نوشتار پیشنهاد شده است، اما همچنان قسمت عمده‌ای از واژه‌های تفکیک شده توسط فیلترهای ما، از دقت لازم برخوردار نیستند. به‌عنوان نمونه، واژگانی نظیر «سینما»، «سینمای»، «سینمائی» (۳ کلمه مجزا که عملاً یک واژه‌اند) باعث می‌شوند تا الگوریتم‌های پیش‌نهاد شده، خصوصاً الگوریتم ما که به تفاوت نرخ تکرار با نرخ تکرار متوسط بها می‌دهد، عملکرد مناسب خود را نشان ندهند.

۲. دسته‌بندی معیار ما، دسته‌بندی خود وبلاگ‌نویسان است که در برخی از موارد دسته‌بندی درستی را بیان نمی‌کند. برای مثال در میان وبلاگ‌ها، وبلاگی دیده می‌شود که در دسته‌ی «عکاسی» ثبت شده است، حال آن‌که محتوای آن دانلود ترانه است. دلیل چنین اشتباهاتی می‌تواند به تغییر محتوای وبلاگ با گذشت زمان یا اشتباهات فنی (مثلاً در هنگام پر کردن فرم) برگردد.

۳. دسته‌های معرفی شده الزاماً فاصله‌ی مناسب از هم را ندارند. برای مثال سه دسته‌ی « رایانه و اینترنت: اینترنت»، «رایانه و اینترنت: اخبار آی تی و جامعه اطلاعاتی» و «تجارت و اقتصاد: کسب درآمد در اینترنت»، هر سه می‌توانند معرف یک وب‌سایت اینترنتی جدید در رابطه با شبکه Forex (نقل و انتقالات پول و کسب درآمد) باشند.

۴. خروجی‌های الگوریتم‌های معرفی شده به‌صورت تنها یک وبلاگ (یک نظر نهایی) مورد سنجش قرار گرفت و امکان ارائه‌ی دسته‌های دوم و سوم (و امتیازدهی به آن‌ها) میسّر نشد. برای مثال اگر یک الگوریتم، یک وبلاگ عکاسی را چه در دسته‌ی تصویربرداری (مرتبط نسبی) قرار می‌داد و چه در دسته‌ی «پرستاران» (بی‌ربط)، امتیاز صفر می‌گرفت.

۵. عمومی‌نویسی هم‌چنان در بین وبلاگ‌ها شایع است. از آن‌جا که داده‌های این پروژه در ایّام ماه محرم جمع‌آوری شده‌است، اکثر وبلاگ‌ها حداقل یک پست از خود را به یادواره‌ی حماسه‌ی کربلا و یا اتفاقات

سیاسی (بی‌حرمتی برخی عزاداران در عاشورای ۱۳۸۸[۴۸]) اختصاص داده‌اند که این مسئله، نویز زیادی در تحلیل ما ایجاد کرد.

با تمام این اوصاف، به‌نظر می‌رسد که الگوریتم ما راه جدیدی برای ارزیابی شباهت اسناد و خصوصاً وبلاگ‌ها پیش‌نهاد کرده است و پارامترهای آن (نظیر ضرایب تابع چند مقداره‌ی Sign یا لگاریتم‌های مخرج) می‌توانند بسته به گنجینه و دامنه‌ی مورد استفاده تعدیل شوند.

۴.۵.۱ ایجاد گراف شباهت

در این پروژه، معیار بررسی ما در ارزیابی این ۳ الگوریتم تنها مقایسه‌ی وبلاگ با دسته‌ها بوده‌است. با این حال، یکی از ایده‌های قابل اجرا در همین زمینه، محاسبه‌ی شباهت وبلاگ-به-وبلاگ و ساخت یک گراف کامل وزن‌دار دوبخشی (بخش اوّل: نمونه‌های آزمایشی، بخش دوم: نمونه‌های آموزشی) است. در صورت ساخت این گراف و وزن‌دهی به یال‌های آن با هر یک از ۳ الگوریتم گفته شده، امکان اجرای الگوریتم‌های نظیر KNN (انتخاب دسته‌ای که بیش‌ترین رأس را در مجموعه‌ی K همسایه‌ی نزدیک یک رأس خاص دارد، به‌عنوان دسته‌ی آن) و یا الگوریتم میانگین فاصله (و انتخاب دسته‌ای که کمترین میانگین فاصله را دارد) فراهم خواهد شد.

از آن‌جا که ساخت این گراف‌ها نیازمند محاسبه و نگه‌داری حدود ۷ میلیون یال (حدود ۱۰۰۰ نمونه آزمایشی ضرب‌در ۷۰۰۰ نمونه‌ی آموزشی) در محک ما خواهد بود، این ایده در پروژه‌ی ما بررسی نشده و تنها در راهکارهای آتی پیشنهاد شده است.

البته لازم به ذکر است که انتخاب تنها بخشی از نمونه‌های آموزشی به‌عنوان شاخص دسته (مثلاً ۱۰درصد از ۱۶۰ وبلاگ آموزشی هر دسته) و سپس اجرای الگوریتم KNN در کنار ۳ الگوریتم معرفی شده، انجام شد؛ منتهی از آن‌جا که وبلاگ‌های شاخص، بعضاً محتوای نادقیق و بی‌ربطی داشتند، این روش به‌دلیل کیفیت نامناسب کنار گذاشته شد.

[48] http://fa.wikipedia.org/wiki/(تظاهرات_هواداران_جنبش_سبز_ایران_در_تاسوعا_و_عاشورا_(۵_و_۶_دی_۱۳۸۸)

۵ نتیجه‌گیری و راهکارهای آتی

۵.۱ نتیجه‌گیری و جمع‌بندی

در این پروژه ضمن معرفی وبلاگ‌ها و مشکلات خط و زبان فارسی (در فصل اول)، تلاش‌هایی برای تحلیل خودکار محتوای وبلاگ‌ها انجام و گزارش آن‌ها ارائه شد. این تلاش‌ها در فصل دوم در راستای دستیابی به یک بانک داده‌ی مناسب بود؛ سپس در فصل سوم مناسب‌سازی داده‌های جمع‌آوری شده و در فصل چهارم معیارهایی برای سنجش و قیاس این محتواها بود.

به‌عنوان یک معیار برای ارزیابی کیفیت تلاش‌های انجام شده، مسئله‌ی دسته بندی وبلاگ‌ها به دسته‌های محتوایی (نظیر سیاسی، ادبی، اقتصادی، علمی، ...) مطرح شد. سپس الگوریتم پیش‌نهادی ما برای ارزیابی شباهت وبلاگ‌ها با دو الگوریتم دیگر مقایسه شد که در این میان، عملکرد و صحّت الگوریتم ما تقریباً مساوی با الگوریتم شناخته شده و رایج «Cosine Similary با وزن‌دهی tf-idf» و بهتر از الگوریتم «Jaccard» بود.

البته از آن‌جا که عملکرد الگوریتم ما در برخی دسته‌ها به‌شدت بهتر از الگوریتم CS بود، احتمال آن می‌رود که با تغییر و بهبود الگوریتم ما، بتوان بر الگوریتم CS نیز غلبه کرد. تکمیل رویه‌های جمع‌آوری و مناسب‌سازی داده‌ها نیز از جمله عملیاتی است که می‌تواند کیفیت ارزیابی الگوریتم ما را بهبود ببخشد.

۵.۲ پیشنهاد برای کارهای آتی

با در نظر گرفتن مشکلات، ساده‌سازی‌ها و عملیات انجام‌نشده در این پروژه، موارد زیر به‌عنوان پیش‌نهاد برای ادامه و بهبود این پروژه و موارد مشابه ارائه می‌شود.

۱. **کار با عبارات مرکّب**[۴۹]. ویکی‌پدیا معرفی مختصری از این روش دارد[۵۰]. یک مثال از اجرای این روش روی متد اجرایی ما می‌تواند چنین باشد که به‌جای در نظر گرفتن هر تک واژه، واژه‌های

[49] Compound Terms

[۵۰] قابل مطالعه در http://en.wikipedia.org/wiki/Compound_term_processing

متوالی (برای مثال، هر دو واژه‌ی متوالی) یک عبارت (term) در نظر گرفته شوند. با این فرض، تعداد کلّ عبارات در یک متن با n واژه‌ی تکی، برابر $n - ۱$ عدد می‌شوند، منتهی تنوع عبارات ما به‌صورت نمایی رشد می‌کند. بدیهی‌ست که ایست‌واژه‌های چنین پردازشی متفاوت از کار ما خواهد بود و برخی فیلترهای زبانی معرفی شده در این پروژه نیز می‌بایست اصلاح اساسی پیدا کنند.

۲. **تلفیق شباهت محتوایی و ساختاری.** [جمالی، ۸۶] وبلاگ‌های فارسی را بر اساس ساختار پیوندی (لینک‌ها) مورد بررسی قرار داده‌است و اجتماعات وبلاگی را در آن شناسایی کرده‌است. در این پایان‌نامه نیز وبلاگ‌ها بر اساس محتوای واژگانی‌شان بررسی شده‌اند تا شباهت بین‌شان اندازه‌گیری شود. تلفیق این دو فاکتور می‌تواند نتیجه‌ی مناسب‌تر و دقیق‌تری را ارائه دهد. در همین راستا [Herring, 09] در مقاله‌اش با بررسی روش‌های مختلف آنالیزکردن محتوای وب (نظیر وبلاگ‌ها) اشاره کرده است که یک روش تلفیقی به احتمال زیاد نتیجه‌ی دقیق‌تری نسبت به روش‌های منفرد (فقط ساختار لینک‌ها، فقط محتوا) می‌دهد.

۳. **تعدیل پارامترهای فرمول شباهت ارائه شده.** فرمولی که در این پروژه برای شباهت وبلاگ‌ها ارائه شده‌است، تجربی و شهودی بوده است. با افزودن پارامترهای دیگری به این فرمول و تعدیل پارامترهای موجود، می‌توان شاهد نتایج جدیدی شد. برای مثال اجرای یک الگوریتم ژنتیک برای پیداکردن کاراترین رابطه‌ی شباهت (در مسئله‌ی دسته‌بندی) می‌تواند نتایج مفیدی در پی داشته باشد.

۴. **بررسی الگوریتم‌های دیگر در مسئله‌ی دسته‌بندی روی یک گراف وزن‌دار.** در فصل سوم ما داده‌های متنی (RSS) را از فضای برداری[۵۱] به یک مقدار عددی بین هر دو وبلاگ تبدیل کردیم تا دامنه‌ی کار ما از k عدد نقطه‌ی پراکنده در فضای n-بُعدی به یک گراف کامل وزن‌دار k رأسی تبدیل شود. پس از آن الگوریتم‌هایی نظیر KNN به‌راحتی روی مسئله کاربست‌پذیر[۵۲] خواهند بود. با اجرای این تبدیل، اکنون راه برای اجرا و مقایسه‌ی سایر الگوریتم‌ها که تنها نیاز به یک گراف وزن‌دار دارند، باز شده است.

[51] http://en.wikipedia.org/wiki/Vector_space_model
[52] Applicable

۵. **اجرای الگوریتم روی زبان‌های دیگر.** تفکیک فیلترهای ما به سه دسته‌ی ساختاری، زبانی، پایانی، باعث شده‌است تا برای اجرای این الگوریتم روی زبان‌ها دیگر (نظیر عربی، اردو و ...) تنها نیاز به اطلاعات بسیار مختصری از آن زبان داشته باشیم. بررسی این‌که آیا اجرای راهکار ما در ارزیابی شباهت، بر روی زبان‌های دیگر بدون تسلط مناسب نیز میسّر است یا نه، می‌تواند به کشفیات جالب زبانی بیانجامد.

۶ مراجع

آدرس‌های اینترنتی مراجع، همگی در تاریخ ۳۰ آذر ۱۳۸۸ قابل بازدید و دسترسی بوده‌اند. بدیهی است که ممکن‌ست برخی از این آدرس‌ها با گذشتن زمان دستخوش حذف یا تغییر شوند.

۶.۱ مراجع فارسی

۱. **[استاندارد ۲۹۰۱]** استاندارد ملّی شماره‌ی ۲۹۰۱. «طرز قرار گرفتن حروف و علائم زبان فارسی بر روی صفحه کلید کامپیوتر»، تجدید نظر اوّل، چاپ سوم. مؤسسه‌ی استاندارد و تحقیقات صنعتی ایران. قابل مشاهده از طریق آدرس: http://www.isiri.org/std/2901.htm

۲. **[استاندارد ۶۲۱۹]** استاندارد ملّی شماره‌ی ۶۲۱۹. «فنّاوری اطلاعات - تبادل و شیوه‌ی نمایش اطلاعات فارسی براساس یونی‌کد». نسخه نهایی. قابل دریافت از طریق آدرس: http://sourceforge.net/projects/farsitools/files/Persian%20Standard/Final%20Draft/finalversion.pdf/download

۳. **[اشرف‌زاده، ۸۲]** اشرف‌زاده، بهرام. **«زبان فارسی در وبلاگ‌های فارسی»**. ماهنامه‌ی دنیای کامپیوتر و ارتباطات، سال سوم، ویژه‌نامه وبلاگ، زمستان ۱۳۸۲. قابل مشاهده از طریق آدرس: http://atalebi.com/articles/show.asp?ID=381

۴. **[بهبهانی، ۸۳]** بهبهانی، رضا. **«پدیده وبلاگ»**. مجله ره‌آورد نور، شماره ۹. زمستان ۱۳۸۳. قابل دریافت از آدرس: http://www.noormags.com/view/Magazine/ViewPages.aspx?ArticleId=24349

۵. **[جمالی، ۸۶]** جمالی، سید محسن. **«کاوش در شبکه احتماعی وبلاگ‌های فارسی»**. پایان‌نامه کارشناسی ارشد گرایش مهندسی نرم‌افزار، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف. بهار ۱۳۸۶.

۶. **[رسولی، ۸۵]** رسولی، سید محمدتقی. **«دنیای مجازی وبلاگ»**. روزنامه‌ی همشهری، یکشنبه ۸ بهمن ۱۳۸۵. قابل دریافت از آدرس: http://www.hamshahrionline.ir/HAMNEWS/1385/851108/world/sciew.htm

۷. **[زارعیان، ۸۳]** زارعیان، کاظم. **«ارائه‌ی راه‌حل برای برخی مسائل اتوماسیون و نگارش فارسی»**. فصل‌نامه اطلاع‌رسانی، دوره ۱۹، شماره ۳ و ۴، بهار و تابستان ۱۳۸۳. قابل دریافت از آدرس: http://www.irandoc.ac.ir/etela-art/19/19_3_4_1.htm

۸. **[شورا، ۸۵]** شورای عالی انقلاب فرهنگی. **«فراگیر کردن دستور خط فارسی مصوب فرهنگستان زبان و ادب فارسی با بهره‌گیری از رایانه»**. مصوب پانصد و هشتاد و ششمین جلسه مورخ ۶/۴/۱۳۸۵. قابل دریافت از آدرس:
http://tarh.majlis.ir/?ShowRule&Rid=D725EAD2-8B7E-437E-84F1-FB6B79B8BEE8

۹. **[شوقی، ۸۵]** شوقی، رضا. **«فراوانی موضوعات وبلاگستان فارسی»**. وب‌سایت رادیو زمانه. ۴ دی ۱۳۸۵. قابل دسترسی از آدرس:
http://www.radiozamaneh.org/radioblog/2006/12/post_54.html

۱۰. **[شیخ‌اسماعیلی، ۸۶]** شیخ‌اسماعیلی، کیومرث؛ رستمی، ئه‌سرین. «لیست کلمات ایست فارسی». آزمایشگاه وب‌معنایی، دانشگاه صنعتی‌شریف. ۱۳۸۶. قابل‌دسترسی از آدرس:
http://www.iranianlinguistics.org/papers/Stopwords-EsmailiRostami.pdf

۱۱. **[صابری، ۸۶]** صابری، محمدکریم؛ صدیقی، حسن. **«مروری بر وب ۱ با نگاهی به وب ۲»**. مجله الکترونیکی پژوهشگاه اطلاعات و مدارک علمی ایران، دوره هفتم، شماره سوم. دی ۱۳۸۶. قابل دریافت از آدرس: http://www.irandoc.ac.ir/data/E_J/vol7/saberi_sedighi.pdf

۱۲. **[صدیقی، ۸۳]** صدیقی، محسن؛ زمانی‌فر، کامران؛ شهیدی، سیّد مجتبی. **«روشی برای رفع چالش‌های محتواکاوی وب‌های فارسی‌زبان»**. مجله الکترونیکی پژوهشگاه اطلاعات و مدارک علمی ایران، دوره چهارم، شماره دوم. اسفند ۱۳۸۳. قابل دریافت از آدرس:

۱۳. **[ضیایی‌پرور، ۸۶]** ضیایی‌پرور، حمیدرضا. **«وبلاگ‌های فارسی در کانون توجه رسانه‌های جهان»**. سایت خبری ebtekarnews، شماره ۱۱۵۳. اسفندماه ۱۳۸۶. قابل مشاهده از آدرس:
http://www.reporter.ir/archives/86/12/005462.php

۱۴. **[ضیایی‌پرور، ۸۷]** ضیایی‌پرور، حمیدرضا. **«وبلاگستان فارسی، بررسی وضعیت وبلاگ‌های فارسی در سال ۱۳۸۶»**. دفتر مطالعات و توسعه رسانه‌ها، معاونت امور مطبوعاتی و اطلاع‌رسانی وزارت فرهنگ و ارشاد اسلامی. اردی‌بهشت ۱۳۸۷. قابل دریافت از آدرس:
http://www.rasaneh.org/pics/attach-content-296-100.pdf

۱۵. **[فرهنگستان، ۸۴]** فرهنگستان زبان و ادب فارسی. **«دستور خطّ فارسی»**. فرهنگستان زبان و ادب فارسی، نشر آثار، چاپ چهارم، ۱۳۸۴. قابل دریافت از آدرس:
http://www.persianacademy.ir/fa/dastoorpdf.aspx

۱۶. **[قدسی، ۸۳]** قدسی، محمّد؛ ضرابی زاده، حمید. **«برخی نکته‌های نگارش فارسی به سبک جدانویسی»**. قابل دریافت از آدرس: http://sharif.edu/~ghodsi/typing-rules.pdf

۱۷. **[کرمی، ۸۴]** کرمی، طاهره. **«وبلاگ: دریچه‌ای نو در دنیای کتابداری و اطلاع‌رسانی»**. فصل‌نامه علوم و فناوری اطلاعات، دوره ۲۱، شماره ۱. پاییز ۱۳۸۴. قابل دریافت از آدرس: http://www.irandoc.ac.ir/etela-art/21/karami.htm

۱۸. **[مرتضایی، ۸۰]** مرتضایی، لیلا. **«مسائل زبان و خط فارسی در ذخیره‌سازی و بازیابی اطلاعات»**. فصل‌نامه اطلاع‌رسانی، دوره ۱۷، شماره ۱ و ۲. پاییز و زمستان ۱۳۸۰. قابل دریافت از آدرس: http://unipers.com/articles/mortezayi.pdf

۱۹. **[نصیری‌شرق، ۸۷]** نصیری‌شرق، آیدین. **«تصحیح و بهبود غلط‌یاب املایی aspell برای زبان فارسی»**. گزارش دوره کارآموزی، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی‌شریف. پاییز ۱۳۸۷. قابل دریافت از آدرس: http://aideen.org/InternshipReport-NasiriShargh83113804.pdf

۶.۲ مراجع انگلیسی

20. **[Darrudi, 04]** Darrudi, E., Hejazi, M. R, Oroumchian, F. **"Assessment of a Modern Farsi Corpus".** In Proceedings of the 2nd Workshop on Information Technology & its Disciplines (WITID) 2004, ITRC, Kish Island, Iran. Available at: http://ece.ut.ac.ir/dbrg/Hamshahri/Papers/WITID.pdf

21. **[Davarpanah, 09]** Davarpanah, M.R; Sanji, M; Aramideh, M. **"Farsi Lexical Analysis and Stop Words List".** Library Hi Tech Emerald Journal, Vol 27, Issue 3, 2009. Preprint available at: http://www.emeraldinsight.com/Insight/ViewContentServlet?contentType=Article&Filename=/published/emeraldfulltextarticle/pdf/lht-03-2009-0031_rtc_cl_final.pdf

22. **[Herring, 04]** Herring, Susan C., et al. **"Bridging the Gap: A Genre Analysis of Weblogs"** hicss, vol. 4, pp.40101b, Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 4, 2004. Available at: http://www.mediensprache.net/archiv/pubs/4061.pdf

23. **[Herring, 05]** Herring, Susan C., et al. **"A longitudinal content analysis of weblogs: 2003-2004"**. School of Library and Information Science Indiana University, Bloomingtoon. 2005. Available at: http://www.blogninja.com/brog-tremayne-06.pdf

24. **[Herring, 09]** Herring, Susan. C. (In press, 2009). **"Web content analysis: Expanding the paradigm"**. In J. Hunsinger, M. Allen, & L. Klastrup (Eds.), The International Handbook of Internet Research. Springer Verlag. Preprint: http://ella.slis.indiana.edu/~herring/webca.preprint.pdf

25. **[Oreilly, 05]** O'Reilly, T., **"What is Web 2.0. Design Patterns and Business Models for the Next Generation of Software"**, Communications & Strategies, No. 1, p. 17, First Quarter 2007. Available at: http://facweb.cti.depaul.edu/jnowotarski/se425/What%20Is%20Web%202%20point%200.pdf

26. **[Taghva, 03]** Taghva, K., et al. **"A List of Farsi Stopwords"**, Technical Report, Information Science Research Institute, University of Nevada, Las Vegas, July 2003. Available at: http://www.isri.unlv.edu/publications/isripub/Taghva2003-01.ps

۷ واژه‌نامه انگلیسی به فارسی

Applicable	کاربست‌پذیر
Benchmark	محک
Blog	بلاگ
Blogger	بلاگر
Blogging	بلاگ کردن، وبلاگ نوشتن
Character	نویسه
Classification	دسته‌بندی
Comment	نظر
Compound Terms	عبارات مرکّب
Corpus	گنجینه
Cross Validation	درستی‌سنجی ضربدری
Delimeter	حائل
Entry	مدخل
Filter	فیلتر
Format	قالب، فرمت
Information Retrieval	بازیابی اطلاعات
Knowledge Base	پایگاه دانش
Link	پیوند
Noise	نویز

Normalization	نرمال‌سازی
Provider	سرویس‌دهنده، تأمین‌کننده
Post	پست
Stop Word	ایست‌واژه
Tagging	تگ‌گذاری
Term Frequency (TF)	نرخ تکرار عبارت
Training Data	داده آموزشی
Weblog	وبلاگ، بلاگ
Web	وب
Website	وب‌سایت، وبسایت

Abstract

Nowadays, Weblogs are one of the newest phenomenons in generating content and increasing social connectivity in the World Wide Web. It is gussed that a significant portion of the contents of the WWW, especially in Web 2.0, are produced by normal users and through weblogs ([Saberi, 86] and [Oreilly, 05]).

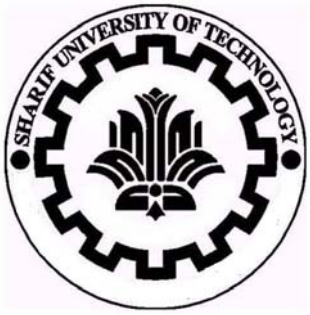
Most of performed analyses on weblogs, in engineering areas, are based on the structure of and hyperlinks (inter-connection) between Weblogs; as they are a great benchmark for statistical analyses in a Graph Theory model or as a reliable instance of a Social Network [Jamali, 86]. Other analyses are mostly in human science fields and performed via a manual reading by human (e.g. [ZiaeiParvar, 86] on Persian Weblogs and [Hurring, 05] on English Weblogs).

One of the toughest difficulties in analyzing the content of Persian Weblogs backs to the structructure of Persian language and writing, and the variety in styles (which is quite visible and mentionable in compare to other languages like English). Difference between speaking and writing language from one side and lack of usage of accent (which causes two words to be written exactly the same and read in different ways) from other side made Pesian Weblogs, a rugged field to explore.

In this project, first an attempt to move toward the decomposing a Weblog into a set of words is proposed in details. Then a measure to evaluate similarity between Persian Weblog based on their contents will be suggested. After that, as an evaluation of that measure, it is going to be compared to two other known algorithms (Jaccard method and Cosine Similarity with tf-idf weighting). Finally, mentioning challenges and difficulties of our path, some ideas for future works in this rarely-explored field will be presented.

Keywords

Weblogs, Persian Weblogs, Content Analysis, Weblog Content, Similarity between Weblogs, Weblog Classification



Sharif University of Technology
Department of Computer Engineering

B. Sc. Thesis
In Software Engineering

Title:

Suggesting and Evaluating a New Content-Based Measure to Find Similarity between Persian Weblogs

By:
Aidin NasiriShargh

Supervisor:
Dr. Hassan Abolhassani

December 2009