دومین کنفرانس ملی
محاسبات توزیعی و پردازش داده‌های بزرگ
2nd National Conference on Distributed Computing and Big Data Processing

دانشگاه شهید مدنی آذربایجان

25 آبان‌ماه 1395

استخراج باهم‌آیی‌های دوتایی و سه‌تایی از پایگاه داده بزرگ بایگانی روزنامه همشهری

دانیال ابراهیم‌زاده[1]، محمد ملااحمدی[2]، احمد یوسفان[3]

[1] دانش آموخته دوره کارشناسی، گروه مهندسی کامپیوتر، دانشگاه کاشان، کاشان،
danial199472@gmail.com

[2] دانش آموخته دوره کارشناسی، گروه مهندسی کامپیوتر، دانشگاه کاشان، کاشان،
mollaahmadimohammad@gmail.com

[3] عضو هیأت علمی، گروه مهندسی کامپیوتر، دانشگاه کاشان، کاشان،
yoosofan@kashanu.ac.ir

چکیده

در بحث متن‌کاوی و پردازش متن بحث استخراج کلمات و رابطه‌ی بین آن‌ها وجود دارد. استخراج و بررسی کلمات باید به نحوی باشد که اطلاعات مفیدی در پی داشته باشد در نتیجه پژوهشگران به بررسی روی گروهی از کلمات متوالی و پی‌درپی پرداختند تا به این وسیله نتایج بهتری از پردازش متون به دست آورند و این گروه، کلمات باهم‌آیی یا همایند نام گرفتند.

در این مقاله که کار بر روی پایگاه داده‌ی اخبار فارسی روزنامه همشهری انجام شده است کوشش بر آن بوده که به وسیله‌ی روش‌های مناسب باهم‌آیی‌ها استخراج شوند. در اینجا ما به استخراج باهم‌آیی‌های دوتایی مانند «جمهوری اسلامی» و باهم‌آیی‌های سه‌تایی مانند «جمهوری اسلامی ایران» به کمک جداول دست‌آوری فراوانی این کلمات پرداخته‌ایم. بر اثر کارهای انجام شده و با توجه به روش مورد نظر در طول این تحقیق با زمان مناسبی باهم‌آیی‌ها به دست آمدند و می‌توان این کار را در پایگاه داده‌های دیگر نیز استفاده کرد و نتایج را بررسی کرد.

کلمات کلیدی

باهم‌آیی، پردازش متن، کلمات متوالی.

1- مقدمه

بررسی باهم‌آیی در هر زبان از اهمیت ویژه‌ای نزد زبان‌شناسان آن زبان برخوردار است. نخستین بار اصطلاح باهم‌آیی را دانشمند زبان‌شناس فرانسوی جی.آر.فرث (J.R.Firth) در نظریه معنایی خود مطرح کرده است. او اساسا این پدیده زبانی را معنا بنیاد فرض کرد نه دستوری، و آن را برای نامیدن و مشخص کردن ترکیبات، بر اساس رابطه‌ی معنایی_اصطلاحی و بسامد وقوع آن‌ها در زبان به کار برد. به نظر او، هم‌نشینی یکی از شیوه‌های بیان معنی است[1]. در زبان فارسی مفهوم باهم‌آیی را نخستین بار محمدرضا باطنی در توصیف ساختمان دستوری زبان فارسی بر اساس نظریه‌ی «مقوله و میزان» هلیدی (Halliday) شناساند و چنین تعریف کرد: سازگاری افراد طبقات (مقولات دستوری) را با یکدیگر در روی زنجیر گفتار، همنشینی (collocation) می‌گوییم[2].

باهم‌آیی یک کلمه‌بندی از متن است که شامل دو یا بیشتر از دو کلمه می‌باشد و در این باهم‌آیی بین کلمات آن شباهت و ارتباط خاصی وجود دارد. در تعریف قدیمی‌تر، باهم‌آیی به این صورت بیان شده است که باهم‌آیی شامل مکان‌های مرسوم و همیشگی یک سری کلمات در کنار هم می‌باشند[7]. به عنوان مثال باهم‌آیی «جمهوری اسلامی» یا «جمهوری اسلامی ایران» که از نتایج به دست آمده در این تحقیق می‌باشند.

در زمینه باهم‌آیی، بررسی‌ها و مقالات و نرم‌افزارهای زیادی به خصوص در زمینه‌هایی همچون پردازش زبان‌های طبیعی (Natural Language Processing) و بازیابی اطلاعات (Information Retrieval) و متن‌کاوی (Text Mining) ارائه شده است که اکثر این کارها روی متن‌های بزرگ صورت گرفته است[8].

کاربردهای عملی تجزیه و تحلیل باهم‌آیی‌ها در زمینه‌های گوناگون:

بازیابی متن‌های ناقص (پردازش صفحات کتاب و یا مجلات)، ترجمه متون (لغتنامه‌های تخصصی که برای ترکیبات کلمات، معانی متناسب و بر اساس گفتگوی زبانی مردم دارند)، متن‌کاوی، تولید زبان طبیعی ربات‌ها و ساخت ربات‌های سخنور (ربات‌هایی که به جای استفاده و ترکیب کلمات واحد، از مجموعه‌های باهم‌آیی‌ها جهت ارتقای سطح گفتگوی خود استفاده کنند).

2- باهم‌آیی از منظرهای متفاوت

باهم‌آیی کلمات در چند زمینه مورد بررسی قرار می‌گیرد و در ادامه به سه زمینه اشاره می‌شود.

2-1- حوزه‌های علمی

از نظر حوزه‌های علمی باهم‌آیی را می‌توان از سه دیدگاه زیر بررسی کرد:

حوزه نحوی باهم‌آیی‌ها: در این حوزه بیشتر به نقش کلمات در جمله و معنای کلی جمله توجه می‌شود[9]. باهم‌آیی به لحاظ صرفی نتیجه‌ی فرآیند ترکیبی واژه‌سازی محسوب می‌شود که طول آن نقش نحوی کلمه نیز مورد نظر قرار می‌گیرد[5].

حوزه معنایی باهم‌آیی‌ها: این حوزه به معنای باهم‌آیی ارتباط دارد به نحوی که این معنی از کلمات باهم‌آیی به طور جداگانه قابل استنتاج نیست. برای مثال باهم‌آیی red tape در اولین نگاه اگر کلمات را به صورت جداگانه معنی کنیم معنای «نوار قرمز» را می‌توان نتیجه گرفت اما این باهم‌آیی معانی کلی‌تر و واضح‌تری مانند «مقررات دست و پا گیر» یا «فرمالیته اداری» را به همراه دارد که این معانی فقط با بررسی خواص معنایی یک زبان و باهم‌آیی‌های آن قابل استنتاج است[10].

حوزه آماری باهم‌آیی‌ها: این حوزه به احتمال وقوع هر کلمه در یک باهم‌آیی مورد نظر می‌پردازد. بر اساس احتمال وقوع و مقدار فراوانی‌های به دست آمده و سایر پارامترهای آماری همچون انحراف معیار و غیره، برای هر باهم‌آیی می‌توان مقادیر و پارامترهایی به دست آورد که بتوان باهم‌آیی‌های هر متن را مورد بررسی و ارزیابی قرار داد[11].

در این تحقیق توجه ما به حوزه‌ی آماری باهم‌آیی‌ها است.

یکی از مهم‌ترین مجموعه فرمول‌ها برای مقایسه باهم‌آیی‌های دوکلمه‌ای (Bigram)، مجموعه ضرایب تخمین وابستگی (Association Measures) است. ساسا پترویچ (sasa petrovic) یکی از کسانی بوده که با استفاده از این ضرایب و تعمیم آن‌ها و هیورستیک‌های مناسب موجود برای این ضرایب، توانست فرمول‌های مناسب و قابل قبولی برای باهم‌آیی‌های بیش از دو کلمه نیز به دست آورد[8].

2-2- تعداد کلمات

باهم‌آیی‌ها در این کار بر اساس تعداد کلمات تشکیل دهنده‌ی آن شامل سه دسته می‌شوند:

باهم‌آیی‌های دوتایی: ترکیبی از دو کلمه مانند «جمهوری اسلامی» که به آن‌ها Bigram گفته می‌شود.

باهم‌آیی‌های سه‌تایی: ترکیبی از سه کلمه مانند «جمهوری اسلامی ایران» که به آن‌ها Trigram گفته می‌شود.

باهم‌آیی‌های با تعداد کلمات بیشتر: این باهم‌آیی‌ها شامل زنجیره‌ای N کلمه‌ای می‌باشند و به آن‌ها ngram گفته می‌شود.

در این مقاله ما فقط به باهم‌آیی‌های دوتایی و سه‌تایی می‌پردازیم

و آن‌ها را استخراج می‌کنیم و باهم‌آیی‌های مراتب بالاتر را در نظر نمی‌گیریم.

2-3- ریشه‌یابی

اصطلاح ریشه‌یابی به معنای حذف پسوندها، پیشوندها و میانوندها و به طور کلی قسمت‌های اضافی کلمه برای به دست آوردن ریشه کلمه است. از جمله اهداف این کار در بازیابی اطلاعات، جستجوی کلمه بر اساس ریشه‌ی آن می‌باشد. برای مثال با استفاده از ریشه‌یابی کلمات «درمان» و «درمانگاه» هر دو به یک شکل و به صورت «درمان» در نظر گرفته می‌شود و هرگاه کلمه درمان لازم باشد اسناد مربوط به کلمه درمانگاه نیز می‌توانند نتیجه این کار باشند[6].

از رایج‌ترین الگوریتم‌های ریشه‌یابی در زبان انگلیسی الگوریتم پورتر (porter) است. در روش پورتر با حذف پسوندهای کلمات، از تعداد واژه‌های منحصر به فرد در بازیابی اطلاعات کاسته می‌شود. در نتیجه موجب بالا رفتن کارایی سیستم خواهد شد[12]. از روش‌های دیگر در زبان انگلیسی الگوریتم لوینس (lovins) است که این الگوریتم شامل 250 پسوند است که بزرگ‌ترین پسوند متصل به کلمه را از آن حذف می‌کند، به شرطی که نتیجه کلمه باقیمانده حداقل سه نویسه طول داشته باشد[13]. روش‌های دیگری نیز برای ریشه‌یابی در زبان انگلیسی مانند الگوریتم bacchin ارائه شده است[14].

در زبان فارسی نیز روش‌هایی ارائه شده است مانند کار شریف‌لو و شمس‌فرد که در مقابل روش حذف پسوندها روش پایین به بالا را به کار برده‌اند به این شکل که ابتدا کل ریشه‌ها را استخراج کرده و بعد بررسی می‌شود که کدام ریشه با قواعد موجود ارتباط دارد[15]. روش دیگر به وسیله تشکری بر اساس حذف حروف ارائه شده است بدین صورت که حروف را تا حد امکان از انتهای کلمه حذف می‌کند[16]. روش‌های دیگری نیز در زبان فارسی وجود دارد مانند: رحیم‌طرقی و همکاران و ...[17].

با توجه به ریشه‌یابی استخراج باهم‌آیی‌ها در دو گروه انجام می‌شود:

استخراج باهم‌آیی‌ها از متن اصلی ریشه‌یابی نشده.

استخراج باهم‌آیی‌ها از متن اصلی ریشه‌یابی شده.

در این کار ما به استخراج باهم‌آیی‌ها بدون ریشه‌یابی متن پرداخته‌ایم.

3- پیش فعالیت‌ها

این مقاله بر روی پایگاه داده‌ی اخبار فارسی روزنامه همشهری انجام گرفته است که این اخبار در چند مرحله توسط چندین نفر از دانشجویان دانشگاه کاشان استخراج شده است و متن آن هر یک به شکل مناسبی در پایگاه داده ذخیره شده است. ملکیان این اخبار را از سایت روزنامه همشهری استخراج کرده و این اخبار را به صورت مناسب و ساخت‌یافته در قالب تاریخ خبر، عنوان خبر و متن خبر ایجاد کرده است[3]. این پایگاه داده بزرگ حدودا دارای 41 هزار خبر و در حدود 2 میلیون خط داده و ده‌ها میلیون کلمه است که ما باهم‌آیی‌ها را از این مقدار داده استخراج کرده‌ایم. ما در ادامه این ویرایش‌ها تگ‌های html و قسمت‌های اضافه را حذف کرده و کار را بر روی متن کاملا ویرایش شده انجام داده‌ایم.

در این تحقیق از زبان پایتون که از زبان‌های بسیار مناسب برای کار با متن است، استفاده شده است.

چندین کار به صورت پیش‌پردازش انجام شده است تا استخراج باهم‌آیی بتواند انجام شود که از جمله آن‌ها می‌توان به «انتقال اخبار به پایگاه داده»، «پیش‌پردازش متن» و «انجام اصلاحات روی پایگاه داده» اشاره کرد.

4- استخراج باهم‌آیی‌ها

همانطور که گفته شد از نظر تعداد کلمات باهم‌آیی به باهم‌آیی‌های دوتایی و سه‌تایی و از نظر ریشه‌یابی به استخراج باهم‌آیی‌های ریشه‌یابی نشده پرداخته‌ایم. قوانینی که در این تحقیق برای استخراج باهم‌آیی استفاده می‌شود به شرح زیر است.

استخراج نکردن باهم‌آیی‌هایی که تعداد تکرار کم دارند. در این مورد هیچ اتفاق نظری بین پژوهشگران نیست که این مقدار کم چقدر است و چقدر باید باشد که چه مقدار از باهم‌آیی‌ها را پوشش بدهد؟ نام این اصل «قانون حد آستانه» گذاشته شده است[4]. در اکثر متون، تعداد باهم‌آیی‌های با تکرار 1 یا 2 و یا 3 حدود شامل 80 درصد کل باهم‌آیی‌ها می‌شود. به همین خاطر زمانی که این حد آستانه انتخاب شود حدود 80 درصد باهم‌آیی‌ها رد می‌شود و سرعت اجرای برنامه بیشتر می‌شود. به همین دلیل است که در اکثر مقالات حد آستانه‌ای را برای استخراج باهم‌آیی‌ها قرار می‌دهند تا هم سرعت اجرای برنامه بالا رود و هم اینکه باهم‌آیی‌های با تعداد کم که ارزش پردازشی ندارند در نتایج دخیل نشوند.

قانون بعدی این است که باهم‌آیی نباید شامل کلمه‌ای با POS(X) باشد، یعنی اجزای باهم‌آیی‌ها فقط باید شامل یکی از نقش‌های اسم، صفت و ایست واژه باشند. البته ما در این تحقیق به علت ریشه‌یابی نکردن متن نقش کلمات را تعیین نکرده‌ایم و تمامی کلماتی که در متن با هم آمده‌اند را به عنوان باهم‌آیی در نظر گرفته می‌شوند.

حساب کردن پارامترهای مورد نیاز هر باهم‌آیی که در ادامه باید برای هر باهم‌آیی بر اساس یک سری از فرمول‌ها عددی نسبت بدهیم تا بتوانیم آن‌ها را دسته‌بندی و باهم‌دیگر مقایسه کنیم.

نحوه محاسبه آن‌ها در ادامه نشان داده شده است. با این پارامترها می‌توان به راحتی «ضرایب تخمین وابستگی» را برای هر باهم‌آیی

محاسبه کرد.

همان گونه که جدول 1 نشـــان می‌دهد برای هر باهم‌آیی دوتایی UV چهار پارامتر Oij می‌توان به دســـت آورد که هر کدام از آن‌ها به صورت زیر است:

O11 به این معنی اســـت که کلمه اول باهم‌آیی دقیقا U و کلمه دوم باهم‌آیی دقیقا V باشند.

O12 به این معنی اســـت که کلمه اول باهم‌آیی دقیقا U و کلمه دوم باهم‌آیی هر چیزی به جز V باشند.

O21 به این معنی اســـت که کلمه اول باهم‌آیی هر چیزی به جز U و کلمه دوم باهم‌آیی دقیقا V باشند.

O22 به این معنی است که کلمه اول و دوم باهم‌آیی هر چیزی به جز U و V باشند.

جدول (1) : نحوه محاسبه پارامترهای مربوط به باهم‌آیی‌های دوتایی

	V=v	V≠v
U=u	O۱۱	O۱۲
U≠u	O۲۱	O۲۲

حـال اگر تمـام این چهـار پـارامتر را با هم جمع کنیم تعداد کل کلمات متن به دست می‌آید که آن را با N نشان می‌دهیم.

$$N=011+012+021+022 \quad (1)$$

علاوه بر این پارامتر R1 به معنای تعداد تکرار کلمه اول باهم‌آیی (U) در کل متن به صورت زیر محاسبه می‌شود.

$$R1 = O11 + O12 \quad (2)$$

R2 به معنای تعداد کلمات متن به جز کلمه اول باهم‌آیی (U) است.

$$R2 = O21 + O22 \quad (3)$$

اکنون اگر R1 بـه معنای تعداد تکرار کلمه اول باهم‌آیی و R2 به معنـای تعداد تکرار هر کلمه‌ای به جز کلمه اول باهم‌آیی را با هم جمع کنیم برابر با N یعنی تعداد کل کلمات است.

$$N = R1 + R2 \quad (4)$$

پارامترهای دیگر C1 و C2 هســتند که مانند R1 و R2 هستند با این تفاوت که برای کلمه دوم باهم‌آیی (V) می‌باشند.

در باهم‌آیی‌های ســـه‌تایی UVZ ماهیت پارامترها همچنان همان اســت، ولی به دلیل وجود ســه کلمه تا حدی نوع به دست آوردن آن‌ها متفاوت خواهد بود. پارامتر Oijk به این صـورت است که زمانی که هر کدام از i، j و k برابر مقدار 1 باشـــد یعنی حضـــور آن کلمه، ولی اگر مقدار 2 باشد یعنی عدم حضور کلمه مورد نظر. برای مثال:

O112: یعنی کلمـه اول باهم‌آیی دقیقا U باشـــد و کلمه دوم نیز دقیقا همان V باشد و کلمه سوم هر چیزی به جز Z باشد.

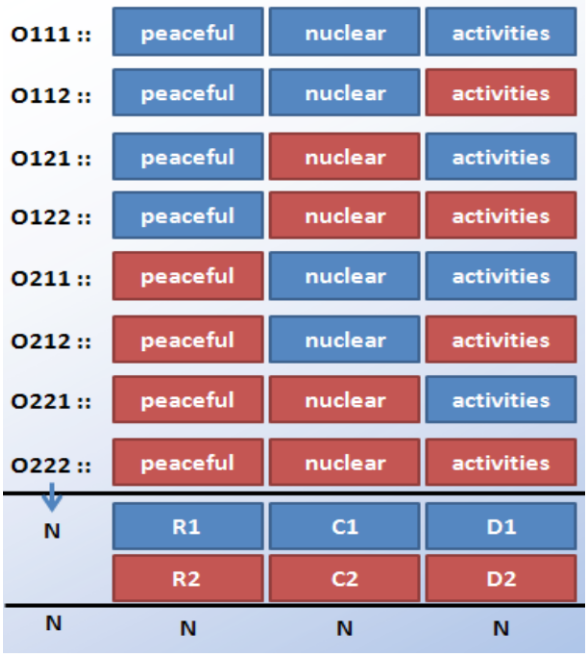
O111: یعنی هر ســـه کلمـه بـاهم‌آیی دقیقـا برابر با U، V و Z هستند.

O222: یعنی کلمـه اول بـاهم‌آیی هرچیزی به جز U و کلمه دوم باهم‌آیی هر چیزی به جز V و کلمه ســـوم باهم‌آیی هر چیزی به جز Z باشند.

شکل 1 نحوه محاسبه پارامترها را روی یک مثال نشان می‌دهد و همچون دوتایی‌ها که پارامترهای R و C داشتند در سه‌تایی‌ها علاوه بر آن‌ها پارامتر دیگری به نام D وجود دارد که مربوط به تعداد تکرار کلمه سوم باهم‌آیی است که به شکل زیر هستند:

R1: تعداد تکرار کلمه اول باهم‌آیی (U) در کل متن

R2: تعداد کلمه‌های کل متن به جز کلمه اول باهم‌آیی



شکل (1) : نحوه محاسبه Oijkها در باهم‌آیی‌های سه‌تایی

C1: تعداد تکرار کلمه دوم باهم‌آیی (V) در کل متن

C2: تعداد کلمه‌های کل متن به جز کلمه دوم باهم‌آیی

D1: تعداد تکرار کلمه سوم باهم‌آیی (Z) در کل متن

D2: تعداد کلمه‌های کل متن به جز کلمه سوم باهم‌آیی

همانند باهم‌آیی‌های دوتایی در این قســمت هم اگر هر Oijk را با هم جمع کنیم برابر با کل تعداد کلمات متن می‌شود.

برای اســتخراج باهم‌آیی‌ها در کل متن کلمات باهم‌آیی را جستجو می‌کنیم و هرگاه به نتیجه رســـیدیم مقدار O11 را برای باهم‌آیی‌های دوتـایی یا O111 را برای باهم‌آیی‌های ســـه‌تایی را یک واحد افزایش می‌دهیم و این کار را تا انتهای متن انجام می‌دهیم. ســپس در آخر اگر مقدار O11 یا O111 بیشتر از حد آستانه بود بقیه پارامترها را نیز برای این باهم‌آیی محاسبه می‌کنیم.

بعد از استخراج باهم‌آیی‌ها تمامی اطلاعات مربوط به آن از جمله پارامترهای مربوطه و کلمات تشکیل دهنده‌ی آن را در جداولی در پایگاه داده ذخیره می‌کنیم. برای سه‌تایی‌ها به ذخیره یازده پارامتر نیاز داریم که هشت مورد از آن‌ها مربوط به Oijk ها و سه مورد دیگر هم برای نمایش کلمات تشکیل دهنده باهم‌آیی است.

5- نتایج کار

در این کار با استفاده از قواعدی همچون قاعده‌ی حد آستانه و اجرای چندین روزه آن به نتایج خوبی از پایگاه داده موجود رسیدیم.

در جدول 2 چند نتیجه‌ی ابتدایی استخراج باهم‌آیی دوتایی را مشاهده می‌کنید. این نتایج فراوانی بیشتری از نظر تعداد تکرار در متن اخبار روزنامه همشهری نسبت به دیگر باهم‌آیی‌های دوتایی برخوردار بودند. در جدول 2 مشاهده می‌کنید که باهم‌آیی‌های «جمهوری اسلامی» و «گزارش خبرگزاری» بیشترین تعداد تکرار را داشته و فقط فعل «شده است» تعداد تکرار بیشتری نسبت به این دو دارد.

در جدول 3 نیز مشاهده می‌کنید که باهم‌آیی سه‌تایی مانند «به گزارش خبرگزاری» و «جمهوری اسلامی ایران» از پر کاربردترین باهم‌آیی‌های سه‌تایی هستند.

6- نتیجه‌گیری و پیشنهادات

در طول این مقاله کوشش بر این بوده است که از مناسب‌ترین روش‌ها، نرم‌افزارها و بسته‌ها در حوزه پردازش متن استفاده شود و بر اساس روش اتخاذ شده و استفاده از قواعد موجود به نتایج خوبی در استخراج باهم‌آیی‌ها رسیدیم.

از این نتایج می‌توان در مباحثی همچون پردازش زبان‌های طبیعی و ایجاد فرهنگ لغت‌ها استفاده کرد و یا این مراحل کار را بر روی پایگاه داده‌های دیگر انجام داد و نتایج به دست آمده را مقایسه کرد. همچنین در اثر این کارها که بر روی بایگانی روزنامه همشهری انجام شده است می‌توان فهمید که چه باهم‌آیی در این سال‌ها بیشتر استفاده شده که همانطور که در نتایج مشاهده شد اسم یک شخصیت یا اصطلاح خبرنگاری یا سیاسی و سایر باهم‌آیی‌ها می‌توانند باشند. سپس آن را با بایگانی‌های سال‌های بعد مقایسه کرد و ببینیم آن باهم‌آیی هنوز جایگاهی دارد یا خیر. مثلا بعضی باهم‌آیی‌های مربوط به مسائل اقتصادی و دینی مشاهده می‌شود که هنوز کاربرد دارند اما بعضی اسم شخصیت‌ها مانند رئیس‌جمهورها در طی چندین سال استفاده شده و بعد تقریبا بدون کاربرد می‌شوند.

از جمله کارهایی که می‌توان در آینده در رابطه با این کار انجام داد می‌توان ریشه یابی را به این کار اضافه کرد و همچنین از دستور

جدول (2) : نتایج حاصل از استخراج باهم‌آیی‌های دوتایی

word1	word2	O11	O12	O21	O22
شده	است	8124	15490	79708	6033020
جمهوری	اسلامی	3134	5152	8876	6119180
گزارش	خبرگزاری	2584	6484	1680	6125594
تیم	ملی	2346	7532	3618	6122846
اعلام	کرد	2306	4526	19074	6110436

جدول (3) : نتایج حاصل از استخراج باهم‌آیی‌های سه‌تایی

word1	word2	word3	O111	O112	O121	O122	O211	O212	O221	O222
به	گزارش	خبرگذاری	2488	3488	46	162768	96	2996	1634	5962826
با	توجه	به	1658	228	2808	66316	834	2770	163490	5898238
جمهوری	اسلامی	ایران	1136	1998	24	5128	430	8446	13755	6105425
آقای	هاشمی	رفسنجانی	1094	282	0	3114	208	416	254	6130974

دوم یعنی همان POS(X) اســتفاده کرد و نتایج به دست آمده را مورد بررسی قرار داد و با نتایج این تحقیق مقایسه کرد.

مراجع

[1] پالمر، فرانک، **نگاهی تازه به معنی شناسی**، ترجمهٔ صفوی، کوروش، چاپ دوم، تهران، نشر مرکز، 1374.

[2] باطنی، محمدرضا، **توصیف ساختمان دستوری زبان فارسی**، چاپ نهم، تهران، انتشارات امیرکبیر، 1348.

[3] ملکیان رجبی، هاجر، **پیاده‌سازی نرم‌افزاری برای استخراج اطلاعات غیرساخت‌یافته و تبدیل آن به قالب ساخت‌یافته‌ی xml به همراه امکان جست و جوی ساده و ساخت xHTML از نتیجه‌ی جست و جو**، پایان‌نامه کارشناسی، دانشگاه کاشان، کاشان، 1385.

[4] عباسیان، مسعود، **استخراج و مقایسه و بررسی باهم‌آیی‌ها از سایت خبری انگلیسی صدا و سیمای جمهوری اسلامی ایران با روش‌های Word Clustering و Yarowshy**، پایان‌نامه کارشناسی، دانشگاه کاشان، کاشان، 1391.

[5] پناهی، ثریا، **فرآیند باهم‌آیی و ترکیبات باهمایند در زبان فارسی**، نامه فرهنگستان، شماره 3، دوره 5، صفحه‌های 199-211، 1381.

[6] احسان، نوا، فیلی، هشام، **بررسی تاثیرات ریشه‌یابی در بازیابی اطلاعات در زبان فارسی**، نشریه پردازش علائم و داده‌ها، شماره 1، صفحه‌های 17-24، 1390.

[7] Manning, Christopher, Schutze, Hinrich, *"Foundations of Statistical Natural Language Processing"*, *MITPress*, 1999, Chapter 1.

[8] Petrovic, Sasa, Snajder, jan, *"Extending Lexical Association Measures for Collocation Extraction"*, *Published In Journal of Computer Speech and Language*, 2009, Abstract Chapter.

[9] Smadja, Frank, McKeown, Kathleen, *"Translating Collocations for Bilingual"*, *Published In Journal Computational Linguistics*, *MITPress*, Cambridje, MA, USA, March 1996, Chapter 3, Collocations and Machine Translation.

[10] Wermter, Joachim, Hahn, Udo, *"Collocation Extraction Based on Modifiability Statistics"*, *Published In Proceeding COLING '04 Proceedings of the 20th International Conference on Computational Linguistics*, Association for Computational Linguistics Stroudsburg, PA, USA, 2004, Chapter 1, Introduction.

[11] Petrovic, Sasa, Snajder, Jan, Bojana, Kolar *"Comparison of Collocation Extraction Measures for Document Indexing"*, *28th International Conference on Information Technology Interfaces*, *Published In IEEE*, 2006.

[12] Porter, M.F, *"An Algorithm for Suffix Stripping"*, *Program*, Vol. 14, No. 3, pp. 130-137, July 1980.

[13] Lovins, Julie, *"Development of A Stemming Algorithm"*, *Mechanical Translation and Computation Linguistics*, pp. 23-32, 1968.

[14] Bacchin, Michela, Ferro, Nicola, Melucci, Massimo, *"A Probabilistic Model for Stemmer Generation"*, *Information Processing & Management*, Vol. 41, No. 1, pp. 121-137, 2005.

[15] Sharifloo, Amirazim, Shamsfard, Mehrnoosh, *"A Bottom Up Approach to Persian Stemming"*, *Published In Proceedings of Third International Joint Conference on Natural Language Processing*, 2008.

[16] Tashakori, Masoud, Meybodi, Mohammadreza, Oroumchian, Farhad, *"Bon: The Persian Stemmer"*, *Published In Proc 1th EurAsian Conference on Information*, 2003.

[17] Rahimtoroghi, Elaheh, Faili, Hesham, Shakery, Azadeh, *"A Structural Rule-Based Stemmer for Persian"*, *International Symposium on Telecommunications*, 2010.