نخستین کنفرانس بین المللی پردازش خط و زبان فارسی
۱۵ و ۱۶ شهریور ۱۳۹۱
دانشگاه سمنان - دانشکده مهندسی برق و کامپیوتر

بررسی ریشه‌یاب‌های واژگان زبان فارسی و تاثیر آنها در کارایی سیستم‌های بازیابی اطلاعات متنی

محمد صادق زاهدی[۱]، ارسطو بزرگی[۲] و کاوان فاتحی[۳]
s.zahedi@ece.ut.ac.ir[۱]
a.bozorgi@Mail.sbu.ac.ir [۲]
kavanfatehi@gmail.com [۳]

چکیده

ریشه‌یابی یکی از مهمترین مباحث مطرح شده در پردازش زبان‌های طبیعی و بازیابی اطلاعات متنی است. در این مقاله سعی شده است در ابتدا یک دسته‌بندی کلی از روش‌هایی که اخیرا برای ریشه‌یابی واژگان فارسی انجام شده است، ارائه دهیم و سپس به بررسی این روش‌ها پرداخته و کارایی این روش‌ها را در سیستم‌های بازیابی اطلاعات متنی با پیاده‌سازی مدل بازیابی اطلاعات "Okapi BM25" مورد ارزیابی قرار داده و بر اساس مقادیر معیارهای کارایی سیستم‌های بازیابی اطلاعات متنی ، مقایسه کرده‌ایم. روش‌های بررسی شده در این مقاله عبارتند از: ریشه‌یاب آماری، ریشه‌یاب‌های مبتنی بر ساختار شامل ریشه‌یاب بن، ریشه‌یاب کاظم تقوی، ریشه‌یاب پایین به بالا، ریشه یاب چندفازه.

کلمات کلیدی

ریشه‌یابی، پردازش زبان فارسی، بازیابی اطلاعات متنی.

۱- مقدمه

ریشه‌یابی به عملیاتی اطلاق می‌شود که در آن با حذف پیشوندها و پسوندها، نهایتا ریشه واژه مشخص می‌شود. روش‌های ریشه‌یابی به دو گروه کلی ساختاری و غیرساختاری تقسیم می‌شوند که در روش‌های ساختاری از ساختار کلمات و قواعد زبان برای ریشه‌یابی استفاده می‌شود. حذف وندها و جدول مراجع دو روش مهم در ریشه‌یابی هستند که در گروه روش‌های ساختاری قرار می‌گیرند. در روش حذف وندها، آنقدر پیشوندها و پسوندهایی را از هر دو طرف کلمه حذف می‌کنیم تا اینکه به ریشه کلمه برسیم. در روش جدول مراجع، کلمه و ریشه آن در یک جدول نگه‌داری می‌شوند که برای ریشه‌یابی یک کلمه، آن‌را در جدول جستجو کرده و ریشه متناظر آن برگردانده می‌شود. روش‌هایی نیز وجود دارند که بصورت ترکیبی از دو روش حذف وندها و جدول مراجع هستند. روش‌های غیر ساختاری با استفاده از اطلاعات محدودی از یک زبان سعی می‌کنند به ریشه‌یابی لغات بپردازند. این روش‌ها وابسته به زبان نیستند و روش‌های آماری در این گروه قرار می‌گیرند.

ادامه مقاله به صورت زیر سازماندهی شده است:

الگوریتم‌هایی که اخیرا برای ریشه‌یابی واژگان فارسی ارائه شده را در بخش دوم به طور کامل شرح داده و در بخش سوم به ارزیابی و مقایسه این روش‌ها پرداخته و آنها را بر اساس معیارهای کارایی و دقت در بازیابی اطلاعات متنی مورد بررسی قرار می‌دهیم. نهایتا در بخش چهارم به نتیجه‌گیری می‌پردازیم.

۲- الگوریتم‌های ریشه‌یابی برای واژگان فارسی

در یک دسته‌بندی کلی می‌توان روش‌های ریشه یابی برای زبان فارسی را در دو دسته‌ی ساختاری و غیرساختاری قرار داد که در ادامه این روش‌ها شرح داده می‌شود.

۲-۱- الگوریتم‌های غیرساختاری

تنها ریشه‌یاب پیاده‌سازی شده برای کلمات فارسی که در این گروه قرار می‌گیرد، ریشه‌یاب آماری است که در آن از روشی مبتنی بر گراف و مدل آماری استفاده شده است[۱]. در این روش یک مجموعه از کلمات زبان در نظر گرفته شده و هر کلمه به دو زیررشته تقسیم می‌شود. زیررشته اول پیشوند و زیررشته دوم پسوند نامیده می‌شود. سپس هر زیررشته به عنوان یک گره از گراف در نظر گرفته شده و یک یال بین دو گره نشان‌دهنده این است که از ترکیب این دو زیررشته، یک کلمه از مجموعه لغات بدست می‌آید. برای نشان دادن این تاثیر متقابل، از یک نمادگذاری استفاده شده است تا پیشوند بهینه که همان ریشه