به نام خدا

وزارت علوم، تحقیقات، و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران

(ایرانداک)

خلاصه گزارش برای پژوهشگران

طراحی یک الگوریتم همانندجو برای تشخیص متون بازنویسی شده در زبان فارسی

۱۳۹۶

نصراله پاک‌نیت

معرفی پژوهش

پیشرفت تکنولوژی موجب آسان شدن انتشار و دسترسی به مدارک علمی و در نتیجه سهولت انجام سرقت علمی شده است. با توجه به این موضوع، بررسی میزان مشابهت یک مدرک جدید با مدارک موجود مبدل به مساله‌ای مهم تحت عنوان مساله همانندجویی گشته است. متاسفانه، روش‌های ارائه شده برای این مساله در زبان فارسی تنها متون دقیقا کپی شده را تشخیص داده و قادر به تشخیص متون بازنویسی شده نمی‌باشند. در این پژوهش:

- روش‌های همانندجویی ارائه شده برای سایر زبان‌ها را بررسی می‌کنیم.
- ابزارهای پردازش زبان طبیعی به کار رفته در این روش‌ها را مشخص کرده و کیفیت نمونه‌های موجود از این ابزارها برای پردازش زبان فارسی را بررسی می‌کنیم.
- الگوریتم‌های جدیدی برای همانندجویی در زبان فارسی با هدف تشخیص متون فارسی بازنویسی شده ارائه می‌کنیم.
- با پیاده‌سازی و انجام آزمایشات، کیفیت الگوریتم‌های پیشنهادی را بررسی می‌کنیم.

روش پژوهش

در فاز اول این طرح از رویکرد کتابخانه‌ای استفاده شده و الگوریتم‌های همانندجوی ارائه شده برای سایر زبان‌ها مورد بررسی قرار گرفته است. به عنوان یکی از نتایج این فاز، لیستی از ابزارهای پردازش زبان طبیعی به کار رفته در الگوریتم‌های بررسی شده تهیه گردیده که در فاز دوم این طرح کیفیت نمونه‌های ارائه شده از این ابزارها برای زبان فارسی با رویکردی کتابخانه‌ای مورد بررسی واقع شده است. در فاز سوم این طرح، دو الگوریتم همانندجوی جدید برای زبان فارسی پیشنهاد شده و کیفیت این الگوریتم‌ها با پیاده‌سازی و انجام آزمایشاتی بر روی آن‌ها مورد بررسی واقع شده است.

یافته‌ها

در این طرح پژوهشی، با توجه به ابزارهای موجود، دو الگوریتم برای همانندجویی در متون فارسی بازنویسی شده طراحی شده است. الگوریتم اول طراحی شده در دسته روش‌های همانندجویی معنایی قرار گرفته و برای بررسی همانندی دو واژه از لغت‌نامه استفاده می‌کند. دیگر الگوریتم پیشنهادی، الگوریتمی فازی بوده و از ماتریس هم‌رخدادی لغات برای بررسی همانندی دو واژه استفاده می‌کند. نتایج تجربی آزمایشات صورت گرفته بیانگر کیفیت مناسب روش‌های پیشنهادی در تشخیص همانندی‌های حاصل از بازنویسی متون فارسی می‌باشد.

نتیجه گیری و پیشنهادها

روش‌های همانندجویی ارائه شده برای سایر زبان‌ها را می‌توان به شش دسته ۱) روش‌های مبتنی بر کاراکتر، ۲) روش‌های مبتنی بر بردار، ۳) روش‌های مبتنی بر گرامر، ۴) روش‌های معنایی، ۵) روش‌های فازی و ۶) روش‌های ترکیبی تقسیم‌بندی کرد. با بررسی‌های صورت گرفته در زمینه ابزارهای موجود برای پردازش زبان فارسی، مشخص گردید که دسته روش‌های معنایی و فازی برای تشخیص متون فارسی بازنویسی شده مناسب‌تر می‌باشند. با توجه به این مهم، در ادامه این

طرح پژوهشی، دو الگوریتم جدید برای همانندجویی در متون فارسی بازنویسی شده ارائه گردید. اولین الگوریتم پیشنهادی، الگوریتمی معنایی بوده که در آن همانندی دو واحد از متن با توجه به یک لغت‌نامه بررسی و تعیین می‌شود. دومین الگوریتم ارائه شده الگوریتمی فازی بوده و همانندی واحدهای مختلف متن در آن با توجه به احتمال حضور آن‌ها در کنار یکدیگر با توجه به پیکره‌ای بزرگ از متون فارسی محاسبه می‌شود. نتایج آزمایشات صورت گرفته بر روی دو الگوریتم نشان می‌دهد الگوریتم معنایی پیشنهادی کیفیت بهتری را ارائه می‌نماید. لازم به ذکر است که در الگوریتم‌های پیشنهادی، همانندی هر دو جمله به صورت مستقل بررسی شده است. اما، با توجه به این نکته که قسمت‌های همانند غالبا در متن مشکوک و متن منبع احتمالی نزدیک یکدیگر به کار رفته‌اند، به نظر می‌رسد بتوان با در نظر گرفتن همانندی هم‌زمان جملات همسایه کیفیت الگوریتم‌های پیشنهادی را بهبود داد.