

بسم الله الرحمن الرحيم



وزارت علوم، تحقیقات و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

طرح پژوهشی

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

مجری

نصراله پاک‌نیت

تابستان ۱۳۹۶

حقوق: پژوهشگاه علوم و فناوری اطلاعات ایران

طرح پژوهشی "طراحی یک الگوریتم همانندجو برای تشخیص متون بازنویسی شده در زبان فارسی" پیرو قرارداد شماره ۳۳/۳۶۴۲ مورخ ۱۳۹۵/۴/۱ میان پژوهشگاه علوم و فناوری اطلاعات ایران (کارفرما) و آقای نصراله پاک‌نیت اجرا شده است. گزارش حاضر یک جلد از مستندات این طرح است.

این گزارش و تمامی حقوق مادی آن بر اساس قانون حمایت حقوق مولفان و مصنفان و هنرمندان، مصوب ۱۳۴۸ و اصلاحیه‌های بعدی آن و همچنین آیین‌نامه‌های اجرایی این قانون متعلق به پژوهشگاه علوم و فناوری اطلاعات ایران و هرگونه استفاده از تمامی یا پاره‌ای از آن، شامل: نقل قول، تکثیر، انتشار، کاربرد نتایج، تکمیل و مانند آنها به صورت چاپی، الکترونیکی یا وسایل دیگر فقط با اجازه کتبی پژوهشگاه امکان‌پذیر است. نقل قول در حد هزار واژه در انتشارات علمی مانند کتاب و مقاله با درج اطلاعات کامل کتابشناختی، نیازی به مجوز پژوهشگاه ندارد.

صحت مندرجات گزارش بر عهده مجری طرح پژوهشی است.



برنام خدا

وزارت علوم، تحقیقات و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران

ایرانداک

شماره: ۳۳/۸۴۸۴

تاریخ: ۱۳۹۶/۰۷/۰۵

پیوست: ندارد

پژوهشی، مدیریت دانش، آموزش

گواهی

انجام طرح پژوهشی:

◇ عنوان: طراحی یک الگوریتم همانندجو برای تشخیص متون بازنویسی شده در زبان

فارسی

◇ مجری: دکتر نصراله پاک‌نیت

◇ همکار(ان): -

◇ شماره قرارداد: ۳۳/۳۶۴۲:۵۱

◇ تاریخ قرارداد: ۱۳۹۵/۴/۱

◇ تاریخ شروع: ۱۳۹۵/۴/۱

◇ تاریخ پایان: ۹۶/۶/۲۵

◇ ناظران: دکتر سیدمهدی سمایی، دکتر جلال‌الدین نصیری

در پژوهشگاه علوم و فناوری اطلاعات ایران گواهی می‌شود. گزارش پایانی این طرح، بر پایه داوری در نشست ۲۶۰ (تاریخ ۹۶/۷/۴) شورای پژوهشی پژوهشگاه به تصویب رسیده است. لازم می‌دانم مراتب قدردانی خود را به مجری و ناظران این طرح تقدیم دارم.

با سپاس

رویا پورنقی

سرپرست معاونت پژوهشی و آموزش

قدردانی

اکنون که به یاری خدا طرح طراحی یک الگوریتم همانندجو برای تشخیص متون بازنویسی شده در زبان فارسی را به پایان رسانده‌ام، از جناب آقای دکتر سیروس علیدوستی، ریاست محترم پژوهشگاه که در انتخاب طرح جاری کمک شایانی به اینجانب نمودند، تشکر نموده و از جناب آقای دکتر جلال‌الدین نصیری و جناب آقای دکتر سید مهدی سمائی که با نظرات ارزشمندشان بر غنای این طرح پژوهشی افزودند، قدردانی ویژه دارم. در خاتمه، از همکاران پژوهشکده مدیریت فناوری اطلاعات به پاس پشتیبانی‌هایشان تشکر می‌نمایم.

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

مدیر طرح پژوهشی: نصراله پاک‌نیت، پژوهشگاه علوم و فناوری اطلاعات ایران.

نشانی: تهران، خیابان انقلاب، چهارراه فلسطین، شماره ۱۰۹۰، طبقه چهارم، اتاق ۴۰۳

صندوق پستی: ۱۳۱۵۷۷۳۳۱۴ تلفن: ۰۲۱-۶۶۴۹۴۹۸۰ (داخلی ۴۲۲)

وبگاه: <http://irandoc.ac.ir/u/n-pakniat>

رایانامه: pakniat@irandoc.ac.ir

چکیده

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، مراکز علمی متمایل به آگاهی از میزان همانندی مدارک علمی جدید اعضا خود به مدارک علمی موجود می‌باشند. همانندی می‌تواند به شکل‌های همانندی آشکار (به صورت رونوشت تحت‌اللفظی از یک منبع بدون تغییر) و همانندی مستتر (به شکل‌های بازنویسی متن موجود در یک منبع، ترجمه متن موجود در یک منبع از یک زبان دیگر و سرقت ایده از منبعی دیگر) باشد. روش‌ها و نرم‌افزارهای زیادی برای همانندجویی ارائه شده اما از یک طرف، با توجه به قواعد و ویژگی‌های خاص هر زبان، استفاده از روش‌ها و نرم‌افزارهای همانندجوی موجود برای سایر زبان‌ها به منظور همانندجویی در زبان فارسی ناممکن بوده و یا منجر به ارائه نتایجی نادقیق خواهد شد. از طرف دیگر، تحقیقات قابل دسترس انجام شده در زمینه همانندجویی در متون فارسی، تنها همانندی آشکار را در نظر گرفته‌اند. با توجه به اهمیت موضوع، در این طرح پژوهشی، به بررسی متداول‌ترین همانندی مستتر یعنی بازنویسی متن موجود در یک منبع دیگر در متون فارسی پرداخته خواهد شد. در این راستا، در ابتدا روش‌های همانندجوی موجود برای سایر زبان‌ها و ابزارهای موجود برای پردازش زبان فارسی مورد بررسی واقع می‌شوند. در ادامه، با توجه به ابزارهای موجود، دو الگوریتم برای همانندجویی در متون فارسی بازنویسی شده طراحی خواهد شد. الگوریتم اول طراحی شده در دسته روش‌های همانندجویی معنایی و الگوریتم دوم در دسته روش‌های همانندجویی فازی قرار می‌گیرند. در الگوریتم همانندجوی معنایی پیشنهادی برای بررسی همانندی دو واژه از لغت‌نامه و در الگوریتم فازی پیشنهادی از ماتریس هم‌رخدادی لغات استفاده شده است. در خاتمه، روش‌های پیشنهادی پیاده‌سازی گشته و با انجام آزمایشاتی کیفیت الگوریتم‌های ارائه شده بررسی شده است.

کلیدواژه‌ها: سرقت علمی، همانندجویی، تطبیق متن، متون بازنویسی شده

فهرست مطالب

شماره صفحه

عنوان

.....أ	چکیده.....
..... ۱	۱- فصل اول: مقدمه و اهداف.....
..... ۲	۱-۱- مقدمه.....
..... ۴	۲-۱- اهداف طرح.....
..... ۵	۳-۱- پیوند با اساس نامه و برنامه استراتژیک.....
..... ۷	۲- فصل دوم: بررسی روش های همانندجویی موجود.....
..... ۹	۲-۱- روش های همانندجویی مبتنی بر کاراکتر.....
..... ۱۰	۲-۱-۱- روش همانندجویی ارائه شده توسط گروزا و همکاران.....
..... ۱۱	۲-۱-۲- روش همانندجویی ارائه شده توسط باسیله و همکاران.....
..... ۱۳	۲-۱-۳- روش ارائه شده توسط کاسپرزاک و همکاران.....
..... ۱۵	۲-۱-۴- روش ارائه شده توسط مینایی و نیک نام.....
..... ۱۶	۲-۱-۵- روش ارائه شده توسط یرا و ان جی.....
..... ۱۶	۲-۱-۶- روش ارائه شده توسط ولاسکوئر و همکاران.....
..... ۱۸	۲-۱-۷- روش ارائه شده توسط شرینین و بوتاکف.....
..... ۱۸	۲-۱-۸- روش ارائه شده توسط استاماتاتوس.....
..... ۲۰	۲-۱-۹- روش ارائه شده توسط گیپ و ماسچه.....
..... ۲۱	۲-۲- روش های همانندجویی مبتنی بر بردار.....
..... ۲۲	۲-۲-۱- روش ارائه شده توسط بارون-سدنو و همکاران.....
..... ۲۲	۲-۲-۲- روش ارائه شده توسط موروگیسان و همکاران.....
..... ۲۳	۲-۲-۳- روش ارائه شده توسط بارون-سدنو و همکاران.....
..... ۲۴	۲-۲-۴- روش ارائه شده توسط وایت و جوی.....
..... ۲۵	۲-۲-۵- روش های ارائه شده توسط کنگ و همکاران.....

- ۲-۲-۶-روش ارائه شده توسط ممتاز و همکاران.....۲۶
- ۲-۲-۷-روش ارائه شده توسط قروی و همکاران.....۲۷
- ۲-۳-۳-روش‌های همانندجویی مبتنی بر گرامر.....۲۸
- ۲-۳-۱-روش‌های ارائه شده توسط الهادی و ال-توبی.....۲۸
- ۲-۳-۲-روش ارائه شده توسط عثمان و همکاران.....۲۹
- ۲-۳-۳-روش ارائه شده توسط رام و همکاران.....۳۰
- ۲-۴-۴-روش‌های همانندجویی معنایی.....۳۱
- ۲-۴-۱-روش ارائه شده توسط لی و همکاران.....۳۱
- ۲-۴-۲-روش ارائه شده توسط مشهدی رجب و شمس فرد.....۳۳
- ۲-۵-۵-روش‌های همانندجویی فازی.....۳۵
- ۲-۵-۱-روش ارائه شده توسط یرا و ان جی.....۳۵
- ۲-۵-۲-روش ارائه شده توسط الظهرانی و سلیم.....۳۷
- ۲-۵-۳-روش ارائه شده توسط الظهرانی و سلیم.....۳۷
- ۲-۵-۴-روش ارائه شده توسط کویراشتاین و ان جی.....۳۷
- ۲-۵-۵-روش ارائه شده توسط آنگر و همکاران.....۳۹
- ۲-۵-۶-روش ارائه شده توسط سانچز-پرز.....۳۹
- ۲-۶-۶-روش‌های همانندجویی ترکیبی.....۴۰
- ۲-۶-۱-روش ارائه شده توسط آدام.....۴۱
- ۲-۶-۲-روش ارائه شده توسط الزهرانی و همکاران.....۴۱
- ۲-۶-۳-روش ارائه شده توسط گیپ و ماسچه.....۴۳
- ۳-فصل سوم: ابزارهای پردازش زبان‌های طبیعی به کار رفته برای همانندجویی: بررسی نمونه‌های ارائه شده از این ابزارها برای زبان فارسی.....۴۴
- ۳-۱-ریشه‌یابی لغوی.....۴۶
- ۳-۱-۱-روش‌های ساختاری.....۴۷
- ۳-۱-۲-روش‌های مبتنی بر جدول جستجو.....۴۸
- ۳-۱-۳-روش‌های آماری.....۴۹

۲-۳-برچسب‌گذاری اجزای واژگانی کلام.....	۵۰
۳-۳-شبکه واژگانی.....	۵۲
۴-۳-برچسب‌زنی نقش معنایی.....	۵۳
۵-۳-ایست‌واژه‌ها.....	۵۵
۴-فصل چهارم: الگوریتم‌های همانندجوی پیشنهادی و ارزیابی آن‌ها.....	۵۷
۴-۱-الگوریتم‌های همانندجوی ارائه شده.....	۶۰
۴-۱-۱-الگوریتم همانندجوی معنایی پیشنهادی.....	۶۰
۴-۱-۲-الگوریتم همانندجوی فازی پیشنهادی.....	۶۴
۴-۲-ارزیابی الگوریتم‌های همانندجوی پیشنهادی.....	۶۶
۴-۲-۱-معیارهای ارزیابی یک الگوریتم همانندجو.....	۶۶
۴-۲-۲-نتایج تجربی ارزیابی الگوریتم‌های همانندجوی پیشنهادی.....	۶۸
(۱)نحوه پیاده سازی.....	۶۹
(۲)ارزیابی.....	۷۰
(۳)تحلیل کارایی.....	۸۲
۵-نتیجه‌گیری.....	۸۱
۶-واژه‌نامه فارسی به انگلیسی.....	۸۲
۷-واژه‌نامه انگلیسی به فارسی.....	۸۴
۸-مراجع.....	۸۶

فهرست جداول

شماره صفحه

عنوان

۵۵.....	جدول ۱: ایست‌واژه‌های فعلی زبان فارسی.....
۵۶.....	جدول ۲: ایست‌واژه‌های غیر فعلی زبان فارسی.....
۷۲.....	جدول ۳: برخی نتایج حاصل از اجرای الگوریتم‌های ارائه شده در دسته Simulated obfuscation.....
۷۴.....	جدول ۴: برخی نتایج حاصل از اجرای الگوریتم ارائه شده بر روی دسته سرقت علمی Random obfuscation.....
۷۸.....	جدول ۵: برخی نتایج حاصل از اجرای الگوریتم ارائه شده بر روی دسته سرقت علمی No obfuscation.....

فهرست شکل ها

شماره صفحه

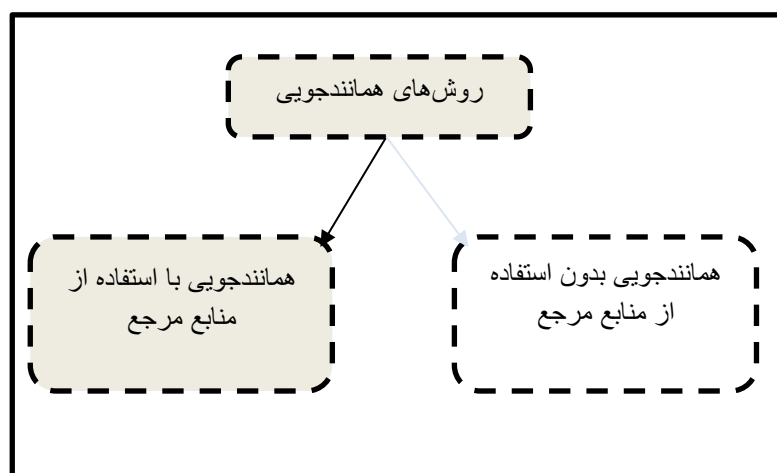
عنوان

-
- شکل ۱: تقسیم‌بندی روش‌های همانندجویی بر اساس استفاده یا عدم استفاده از منابع مرجع در همانندجویی..... ۲
- شکل ۲: دسته‌بندی روش‌های همانندجویی با توجه به قابلیت آن‌ها همانندجویی بین زبانی یا تک‌زبانی..... ۳
- شکل ۳: دسته‌بندی روش‌های همانندجویی بر اساس ابزارهای به کار رفته در آن‌ها..... ۹

فصل اول: مقدمه و اهداف

۱-۱- مقدمه

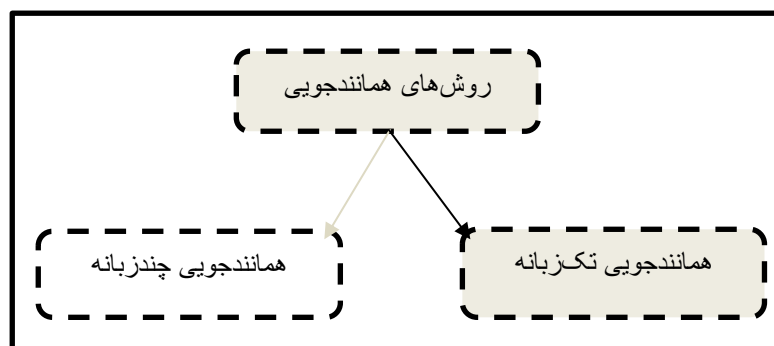
همانندجویی ابزاری است که برای تشخیص سرقت علمی به کار گرفته می‌شود. سرقت علمی عبارت است از استفاده غیرمجاز یا تقلید نزدیک از ایده‌ها و نگارش شخصی دیگر و ارائه آن به عنوان کار خود (Hannabuss, 2001). پیشرفت تکنولوژی موجب آسان شدن انتشار و دسترسی به مدارک علمی و در نتیجه سهولت انجام سرقت علمی شده است. این مساله به وضوح در روزنامه‌نگاری (روزنامه و وبلاگ‌های اینترنتی) و جامعه علمی (مقالات علمی چاپ شده، پایان‌نامه‌ها و رساله‌های دانشگاهی، طرح‌های پژوهشی و ...) قابل رویت است (Stamatatos, 2011). سرقت علمی موجب ضایع شدن حقوق دیگران و کاهش انگیزه پژوهشگران برای کارهای پژوهشی می‌شود. با توجه به این موضوع، ارائه راهکارهایی برای جلوگیری از سرقت علمی ضروری است. یکی از راهکارهای مبارزه با سرقت علمی برخورد با متخلفین است که این مهم نیازمند شناسایی متخلفان است. با توجه به این موضوع، مراکز علمی متمایل به آگاهی از میزان همانندی مدارک علمی جدید اعضا خود به مدارک علمی موجود می‌باشند. روش‌ها و نرم‌افزارهای زیادی برای تشخیص همانندی (همانندجویی) ارائه شده که هر یک قادر به تشخیص درجه‌ای از همانندی‌ها شامل همانندی آشکار (به صورت رونوشت تحت‌اللفظی از منبعی دیگر) و همانندی مستتر (به شکل‌های بازنویسی متن موجود در یک منبع، ترجمه متن موجود در یک منبع از یک زبان دیگر و سرقت ایده از منبعی دیگر) می‌باشند. همانطور که در شکل ۱ نشان داده شده، الگوریتم‌های ارائه شده برای همانندجویی را می‌توان به دسته‌های "همانندجویی بدون استفاده از منابع مرجع" و "همانندجویی با استفاده از منابع مرجع" دسته‌بندی کرد.



شکل ۱: تقسیم‌بندی روش‌های همانندجویی بر اساس استفاده یا عدم استفاده از منابع مرجع در همانندجویی.

در همانندجویی بدون استفاده از منابع مرجع، ورودی الگوریتم تنها یک متن علمی مشکوک بوده و الگوریتم با استفاده از سبک(های) نگارشی به کار رفته در این مدرک علمی، یکتایی نویسنده و یا قسمت‌های نوشته شده توسط نویسندگان متمایز را مشخص می‌کند (Eissen et al., 2006). به عبارتی دیگر، در همانندجویی بدون منابع مرجع، به جای همانندی، ناهماهنگی‌های موجود در سبک نگارشی به کار رفته در مدرک علمی مشکوک بررسی و مشخص می‌شود. در همانندجویی با استفاده از منابع مرجع، ورودی الگوریتم یک متن علمی مشکوک و یک مجموعه از منابع مرجع می‌باشد. هدف الگوریتم، مشخص کردن همانندی‌های موجود مابین متن علمی مشکوک و متون علمی موجود در منابع مرجع می‌باشد (Potthast et al., 2009). بررسی‌های انجام شده نشان می‌دهد که الگوریتم‌های ارائه شده برای همانندجویی بدون استفاده از منابع مرجع در مقایسه با الگوریتم‌های موجود برای همانندجویی با استفاده از منابع مرجع از کارایی پایینی برخوردار هستند (Potthast et al., 2009).

شکل ۲، با توجه به توانایی روش‌های همانندجویی در پشتیبانی از زبان‌های متفاوت، دسته‌بندی مهم دیگری از این روش‌ها را بیان می‌کند: روش‌های تک‌زبانه (Alzahrani et al., 2012) و روش‌های چند زبانه (Barrón-Cedeño et al., 2013). همان‌طور که از نام‌گذاری این دو دسته مشخص است، مزیت الگوریتم‌های ارائه شده در دسته دوم نسبت به دسته اول تشخیص همانندی مستتر از نوع ترجمه شده می‌باشد.



شکل ۲: دسته‌بندی روش‌های همانندجویی با توجه به قابلیت آن‌ها
همانندجویی بین زبانی یا تک‌زبانی

در این طرح پژوهشی به همانندجویی تک‌زبانه و با استفاده از منابع مرجع پرداخته خواهد شد. یک روش همانندجویی با استفاده از منابع مرجع از ۳ فرایند تشکیل شده است: ۱- بازیابی منبع، ۲- تطبیق متن و ۳- پردازش نهایی. در فرایند بازیابی منبع، با فرض در اختیار داشتن یک متن علمی مشکوک، هدف بازیابی تمام منابعی است که از متن آن‌ها در نگارش متن مدرک علمی مشکوک استفاده شده است به طوری که هزینه‌های بازیابی کمینه شود. در فرایند تطبیق متن، با ورودی یک متن علمی مشکوک و یک متن علمی منبع، هدف مشخص نمودن همه قسمت‌هایی از متن مدرک علمی مشکوک می‌باشد که با استفاده از متن مدرک علمی منبع بازنویسی شده است. در فرایند پردازش نهایی، با ورودی مدرک علمی مشکوک، هدف انجام بررسی‌های بیشتر برای ویرایشات نهایی در قسمت‌های مشابه تشخیص داده شده می‌باشد (Potthast et al., 2013).

۲-۱- اهداف طرح

روش‌های همانندجویی موجود برای سایر زبان‌ها قابل استفاده در همانندجویی در زبان فارسی نبوده و اغلب روش‌های همانندجویی ارائه شده برای زبان فارسی تنها همانندی از نوع آشکار را در نظر گرفته و قادر به تشخیص همانندجویی حاصل از بازنویسی متون نیستند. تنها روش‌های همانندجویی موجود که قادر به تشخیص متون فارسی بازنویسی شده هستند نیز در اواسط اجرای این طرح و جهت شرکت در مسابقاتی با این هدف ارائه شده و جزئیات دقیق آن‌ها مشخص نبوده یا دارای نقاط ضعفی جهت استفاده در دنیای واقعی هستند. هدف این طرح پژوهشی، با توجه به نیاز پژوهشگاه^۱ به داشتن سامانه همانندجویی قوی که قادر به تشخیص متون بازنویسی شده نیز باشد، ارائه روش (روش‌هایی) جهت تشخیص همانندی‌های حاصل از بازنویسی در متون فارسی می‌باشد. در ادامه، مراحل انجام شده در راستای دستیابی به این هدف کلی بیان می‌شود:

۱. بررسی روش‌های همانندجویی ارائه شده برای سایر زبان‌ها.

۲. ارائه لیستی از ابزارهای پردازش زبان‌های طبیعی به کار رفته در همانندجویی و بررسی نمونه‌های ارائه شده از این ابزارها برای پردازش زبان فارسی.

^۱ پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

۳. تلفیق نتایج دو مرحله قبل و ارائه الگوریتم(های) همانندجویی که قادر به تشخیص همانندی حاصل از بازنویسی متون فارسی باشد.

۴. انجام آزمایشات مورد نیاز جهت تعیین دقت و کیفیت الگوریتم(های) ارائه شده.

پرسش‌ها/ فرضیه‌های پژوهش را در زیر می‌توان برشمرد:

۱. در روش‌های همانندجوی ارائه شده برای سایر زبان‌ها از چه الگوریتم‌هایی استفاده شده است؟

۲. در روش‌های همانندجویی موجود برای سایر زبان‌ها از کدام یک از ابزارهای موجود برای پردازش زبان‌های طبیعی استفاده شده است؟

۳. کیفیت ابزارهای پردازش زبان طبیعی مورد نیاز برای همانندجویی در زبان فارسی چگونه است؟

۴. الگوریتم(های) طراحی شده برای همانندجویی در متون فارسی بازنویسی شده از چه مولفه‌های تشکیل شده‌اند؟

۵. کیفیت الگوریتم(های) ارائه شده چگونه است؟

۱-۳- پیوند با اساس نامه و برنامه استراتژیک

با توجه به عزم پژوهشگاه بر انجام تحقیقات مرتبط با فعالیت‌ها و اهداف پژوهشگاه به منظور ایجاد یکپارچگی و اثربخشی تحقیقات، همسویی تحقیق جاری با اساس نامه و اسناد بالادستی پژوهشگاه در زیر آمده است:

۱. افزایش کیفیت و کاربرمداری خدمات اطلاعات.

۲. تولید ارزش افزوده از اطلاعات علمی و فناورانه پژوهشگاه.

۳. طراحی و ساخت سامانه تشابه‌یاب.

۴. زمینه‌سازی مناسب برای ارتقا فعالیت‌های پژوهشی مرتبط.

طراحی يك الگوریتم همانندجو برای تشخیص متون بازنویسی شده در زبان فارسی □ ۶

۵. ماموریت گرایی در توسعه و گسترش پژوهش در زمینه علوم و فناوری اطلاعات و دستیابی به علوم و فناوری های نوین در زمینه یاد شده به نحوی که بتواند سرآمدی این زمینه در کشور را تامین کند.

فصل دوم: بررسی روش‌های همانندجویی موجود

در این فصل به بررسی روش‌های همانندجویی موجود پرداخته می‌شود. روش‌های همانندجویی^۱ را بر اساس ابزارهای مورد استفاده در آن‌ها می‌توان به صورت زیر دسته‌بندی کرد:

- روش‌های همانندجویی مبتنی بر کاراکتر
- روش‌های همانندجویی مبتنی بر بردار
- روش‌های همانندجویی مبتنی بر گرامر
- روش‌های همانندجویی معنایی
- روش‌های همانندجویی فازی
- روش‌های همانندجویی ساختاری
- روش‌های همانندجویی مبتنی بر سبک نگارش

علاوه بر دسته‌های بیان شده در (Alzahrani et al., 2012) و با توجه به اینکه برخی از روش‌های همانندجویی از ترکیبی از روش‌های فوق استفاده می‌کنند، در این طرح دسته‌ای دیگر به عنوان دسته

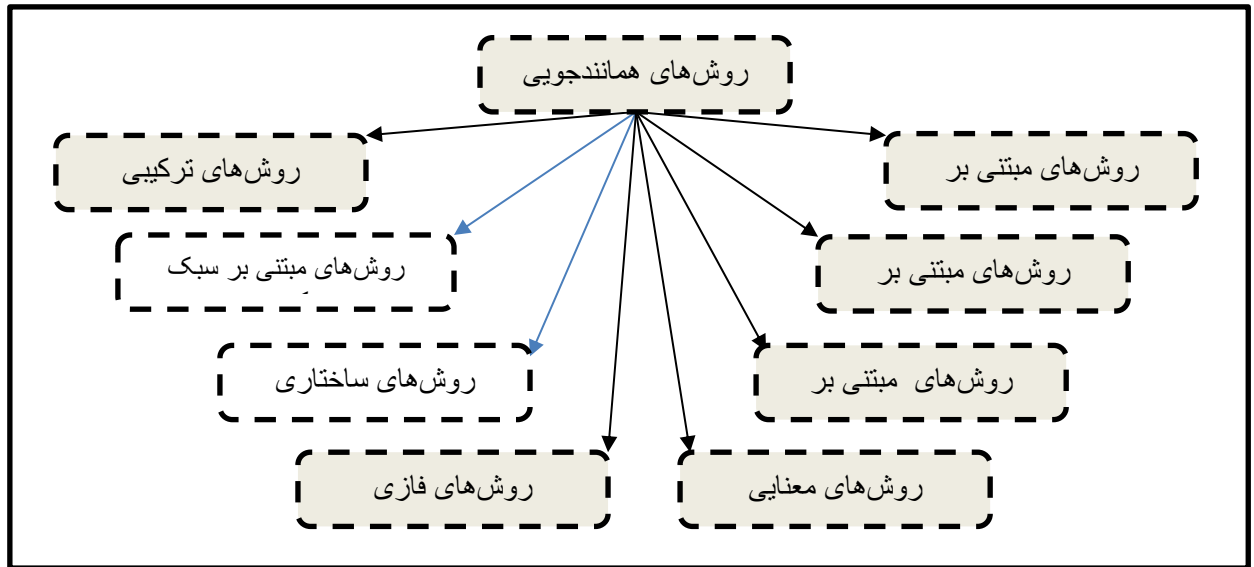
- روش‌های همانندجویی ترکیبی

در نظر گرفته می‌شود.

در ادامه به توصیف ویژگی‌های مهم هر یک از این دسته‌ها پرداخته و مهم‌ترین روش‌های موجود در هر دسته را بررسی خواهیم کرد.

لازم به ذکر است که در این گزارش روش‌های همانندجویی ساختاری و روش‌های همانندجویی مبتنی بر سبک نگارش مورد بررسی قرار نخواهند گرفت. دلیل این امر این است که روش‌های همانندجویی ساختاری در تشخیص سرقت علمی از نوع ایده کاربرد داشته تا تشخیص متون بازنویسی شده و روش‌های همانندجویی مبتنی بر سبک نگارش همان روش‌های همانندجویی بدون منابع مرجع بوده و از حیثه این طرح خارج می‌باشند. شکل ۳، بیان‌گر دسته‌بندی ارائه شده در مورد روش‌های همانندجویی بر مبنای ابزار مورد استفاده و دسته‌های پوشش‌یافته توسط این گزارش می‌باشد.

^۱ در ادامه، منظور از همانندجویی، روش‌های همانندجویی با استفاده از منابع مرجع و تک زبانه می‌باشد.



شکل ۳: دسته‌بندی روش‌های همانندجویی بر اساس ابزارهای به کار رفته در آن‌ها: قسمت‌های پر رنگ شده بیان‌گر دسته‌های پوشش‌یافته در این طرح می‌باشد

روش‌های همانندجویی مبتنی بر کاراکتر ۱-۲

اغلب روش‌های همانندجویی ارائه شده تا به امروز در این دسته قرار می‌گیرند. این دسته از روش‌های همانندجویی به دلیل کارایی و سرعت انجام بالا مقبول بوده اما تنها قادر به تشخیص متون دقیقاً یکسان یا اندکی تغییر یافته می‌باشند. در این روش‌ها، در ابتدا متن مشکوک و متون منبع به مجموعه‌ای از واحدها تقسیم می‌شوند. تقسیم‌بندی می‌تواند به صورت "رشته‌هایی چند کاراکتری"، "رشته‌هایی چند کلمه‌ای"، "جملات" یا حتی واحدهایی با اندازه بزرگ‌تر باشد. در ادامه میزان همانندی هر واحد متن مشکوک با هر واحد از متون منبع با استفاده از الگوریتم‌های "تطبیق رشته" مورد بررسی قرار می‌گیرد. روش‌های مختلف ارائه شده و موجود در این دسته را می‌توان بر حسب استفاده از الگوریتم‌های تطبیق رشته دقیق^۱ یا تقریبی به دو زیر دسته تقسیم کرد. روش‌های استفاده کننده از الگوریتم‌های تطبیق رشته دقیق از کارایی بهتری برخوردار بوده اما تنها قادر به تشخیص قسمت‌های دقیقاً کپی شده می‌باشند. اغلب الگوریتم‌های موجود در این دسته از توابع درهم‌ساز^۲ استفاده می‌کنند.

^۱ String matching

^۲ Hash function

در سوی دیگر، روش‌های مبتنی بر الگوریتم‌های تطبیق رشته تقریبی^۱ قادر به تشخیص قسمت‌های اندکی تغییر یافته بوده اما سرعت اجرای کمتری دارند. از مهم‌ترین معیارهای به کار برده شده در این زیر دسته برای محاسبه همانندی تقریبی می‌توان به فاصله همینگ^۲، فاصله لون‌اشتاین^۳ و بزرگ‌ترین زیر دنباله مشترک اشاره کرد. نکته مهم دیگر در هنگام استفاده از این روش‌ها، طول واحد در نظر گرفته شده می‌باشد. معمولاً تقسیم متن به واحدهایی با طول بیشتر سرعت همانندجویی را بالا برده اما دقت همانندجویی را پایین می‌آورد. در ادامه این بخش، برخی از الگوریتم‌های ارائه شده در این دسته به صورت جزئی‌تر مورد بازبینی قرار می‌گیرند.

۲-۱-۱- روش همانندجویی ارائه شده توسط گروزا^۴ و همکاران

در روش ارائه شده در (Grozea et al., 2009)، پس از تعیین مداخلک منبع احتمالی:

۱- مدرک مشکوک و مدارک منبع احتمالی به رشته‌های n -کاراکتری تقسیم می‌شوند. در این مقاله، ۱۶ به عنوان بهترین مقدار n گزارش شده است. دلیل این انتخاب، همان‌طور که نویسندگان بیان کرده‌اند، این است که با در نظر گرفتن ۵ کاراکتر به عنوان طول متوسط هر کلمه، رشته‌های ۱۶ کاراکتری تقریباً معادل دنباله‌های ۳ کلمه‌ای می‌شوند و در نتیجه این انتخاب با مقدار بهینه گزارش شده در سایر روش‌ها، یعنی ۳ کلمه، سازگار است و دارای مزایایی به شرح زیر است. انتخاب کاراکتر به جای کلمه برای به دست آوردن واحدها تاثیر کلمات طولانی را بیشتر می‌کند. همچنین، این انتخاب منجر به کمتر نمودن تاثیر اصطلاحات رایج در زبان مانند "as far as" می‌شود، زیرا از آن‌جا که طول این عبارت کمتر از ۱۶ کاراکتر است، روش نهایی چنین عباراتی به تنهایی را همانند تشخیص نمی‌دهد. در حالی که روش‌های همانندجویی با واحدهایی بر حسب کلمه ممکن است این عبارت را همانند تشخیص داده و در نتیجه نتایج ضعیف‌تری به دست آورند. همچنین، لازم به ذکر است که تقسیم متون به رشته‌های ۱۶ کاراکتری در راستای افزایش سرعت اجرای روش بوده و

¹ Approximate string matching

² Hamming distance

³ Levenshtein

⁴ Grozea

نویسندگان بیان کرده‌اند که با تقسیم متون به رشته‌های ۸ کاراکتری نتایج بهتری به دست آورده‌اند، اما با هدف دستیابی به سرعت اجرای بهتر، اندکی کیفیت روش را قربانی کرده‌اند.

۲- دنباله‌های n -تایی به دست آمده در هر دو مدرک مرتب می‌شوند.

۳- با الگوریتمی به دست آمده از الگوریتم ترکیب دو مجموعه مرتب دو مجموعه مرتب حاصل از مرتب‌سازی عناصر هر یک از مجموعه‌ها را با هم مقایسه می‌کند. الگوریتم به دست آمده به این صورت کار می‌کند که در هر مرحله، کوچک‌ترین عضو دو مجموعه را در نظر می‌گیرد. در صورتی که کوچک‌ترین عضو در مجموعه دیگر نباشد، این عضو حذف می‌شود. در صورت وجود این عنصر در مجموعه دیگر، عناصر حذف شده و محل وقوع آن‌ها در مدارک مشکوک و منبع احتمالی علامت گذاری می‌شود.

۴- بعد از مشخص شدن تمام زوج‌های یکسان از n -تایی‌ها در مدارک مشکوک و مدرک منبع، این زوج‌ها بر اساس محل وقوعشان در اصل مدارک مرتب شده و پیوستگی قسمت‌های همانند تشخیص داده شده بررسی می‌شود. در ادامه، از الگوریتم مونت کارلو^۱ در راستای ترکیب گروه‌هایی از دنباله‌های n -تایی برای تشخیص کامل قسمت‌های همانند و همچنین حذف قسمت‌های کوتاه استفاده شده است.

۲-۱-۲- روش همانندجویی ارائه شده توسط باسیله^۲ و همکاران

در روش ارائه شده در (Basile et al., 2009)، با هدف افزایش کارایی سرعت اجرای روش و همچنین کاهش فضای مورد نیاز برای ذخیره‌سازی مدارک پردازش شده موجود در پایگاه داده، قبل از اضافه کردن هر مدرکی به پایگاه داده، پیش‌پردازشی اولیه بر روی آن اعمال می‌شود. این پیش‌پردازش به صورت جایگزینی طول یک لغت با خود آن لغت می‌باشد. با انجام این پیش‌پردازش، همانندجویی در متون با حروف الفبای زبان طبیعی به همانندجویی در زبانی جدید و با الفبای {۱، ۲، ...، ۹} تبدیل می‌شود. نویسندگان در نتیجه انجام این پیش‌پردازش، توانسته‌اند که ۸۲٫۵٪ در فضای مورد نیاز برای ذخیره‌سازی داده‌های اضافه مورد نیاز برای همانندجویی صرفه‌جویی کنند. مراحل دقیق‌تر این روش به شرح زیر می‌باشد:

^۱ Monte Carlo

^۲ Basile

۱- در هر یک از مدارک منبع، طول (تعداد کاراکتر) هر کلمه از متن محاسبه شده و جایگزین آن کلمه می‌شود. همچنین، در صورتی که طول یک کلمه بیش از ۹ کاراکتر باشد، عدد ۹ به عنوان طول آن کلمه در نظر گرفته می‌شود.

۱- مدارک جدید به دست آمده به رشته‌هایی n -تایی (n بین ۴ تا ۸) از اعداد تقسیم می‌شوند (توجه شود که n -تایی‌ها هم‌پوشان در نظر گرفته می‌شوند، برای مثال در صورتی که $d = 236542351$ یک مدرک و $n = 5$ باشد، ۵-تایی‌های متناظر با d برابر خواهند بود با ۲۳۶۵۴، ۳۶۵۴۲، ۶۵۴۲۳، ۵۴۲۳۵ و ۴۲۳۵۱).

۲- فراوانی n -تایی‌ها در هر یک از مدارک منبع محاسبه می‌شود.

۳- برای بررسی همانندی یک متن مشکوک با مجموعه متون منبع، فرایند ذکر شده (مراحل ۱ تا ۳) بر روی متن مشکوک نیز اعمال شده و دنباله n -تایی‌های متناظر با آن محاسبه می‌شود.

۴- با استفاده از معیار زیر میزان همانندی مدرک مشکوک y با هر یک از متون منبع x مورد بررسی قرار می‌گیرد.

$$d_n(x, y) = \frac{\sum_{w \in D_n(x) \cup D_n(y)} \frac{f_y(w) - f_x(w)}{f_y(w) + f_x(w)}}{|D_n(x) + D_n(y)|} \quad (1)$$

که $f_x(w)$ فراوانی w در متن x و $D_n(w)$ لیستی شامل تمام n -تایی‌های موجود در x می‌باشد.

۵- با استفاده از معیار فوق، مدارک موجود در پایگاه داده مرتب شده و ۱۰ متنی که با معیار فوق بیشترین مقدار برای آن‌ها به دست آمده به عنوان مدارک منبع احتمالی مدرک مشکوک انتخاب می‌شوند.

۶- برای بررسی بیشتر همانندی بین متن شکوک و هر یک از این ۱۰ منبع احتمالی، از روش کدگذاری جدیدی شبیه به کدگذاری $T9$ استفاده شده و متون موجود در اصل مدارک مشکوک و ۱۰ منبع احتمالی با این روش کدگذاری می‌شوند. این روش کدگذاری نیز باعث کوتاه شدن طول متون اما از بین رفتن مقدار خیلی کمتری از اطلاعات در فرایند کدگذاری نسبت به کدگذاری قبل می‌شود. در این روش، ۲ یا بیش از دو حرف به یک کاراکتر کدگذاری می‌شوند و در نهایت متون جدیدی با الفبای $\{0, \dots, 9\}$ به دست می‌آید.

۷- برای مقایسه مدرک مشکوک با یک مدرک منبع احتمالی، از هر نقطه شروع ممکن در متن مدرک مشکوک شروع کرده و بزرگ‌ترین تطابق در مدرک منبع را به دست می‌آورد و در صورتی که زیررشته‌ای از رشته‌های یافته شده قبلی نبوده و طول آن از مقداری آستانه‌ای بیشتر باشد، آن را ذخیره می‌کند.

۸- فرض کنیم از محور X دستگاه مختصات دکارتی برای نمایش واحدهای موجود در مدرک مشکوک و از محور Y دستگاه مختصات دکارتی برای نمایش واحدهای موجود در مدرک منبع احتمالی استفاده شود. در این صورت، اگر نمودار برخورد واحدهای مشترک رسم شود، در حالتی که وابستگی بین متن منبع احتمالی و متن مشکوک وجود نداشته باشد، نمودار رسم شده دارای نقاطی به صورت یکنواخت پراکنده خواهد بود؛ در غیر این صورت، نمودار رسم شده شامل قسمت‌هایی به صورت خط یا مربع بوده که نمایانگر یکسانی قسمتی با اندازه قابل قبول از متن در هر دو مدرک می‌باشد.

۹- در نهایت، برای متصل کردن قسمت‌های همانند تشخیص داده شده نزدیک به هم، در صورتی که تمام شرایط زیر برقرار باشد، این قسمت‌ها به یکدیگر متصل می‌شوند:

- a. اگر تطابق‌ها دارای درایه X متوالی باشند،
- b. درایه X تطابق‌ها فاصله‌ای بزرگ‌تر از صفر داشته اما این فاصله کمتر از ضریبی از طول بزرگ‌ترین این دو تطبیق باشد،
- c. درایه Y تطابق‌ها فاصله‌ای بزرگ‌تر از صفر داشته اما این فاصله کمتر از ضریبی از طول بزرگ‌ترین این دو تطبیق باشد.

۱۰- مرحله شماره ۹ با ضرایبی کمتر از ضرایب استفاده شده در آن مرحله تکرار می‌شود. انجام این کار موجب بهتر شدن نتایج همانندجویی می‌شود.

لازم به ذکر است که مقادیر بهینه پارامترها، ضرایب و مقادیر آستانه‌ای مورد استفاده در این روش با انجام آزمایشات به دست خواهد آمد.

۲-۱-۳- روش ارائه شده توسط کاسپرزاک^۱ و همکاران

در روش ارائه شده در (Kasprzak et al., 2009):

¹ Kasperzak

- ۱- با هدف حذف فرایند ریشه‌یابی لغوی^۱، در ابتدا اعراب‌گذاری‌ها^۲ و کلمات کوچک از متون حذف می‌شوند. بنا به گفته نویسندگان، حذف فرایند ریشه‌یابی لغوی با هدف غیر وابسته کردن همانندجویی به زبانی خاص انجام شده است.
- ۲- در ادامه، مدارک به دنباله‌هایی n -تایی از لغات تقسیم می‌شوند. بر حسب آزمایشات انجام گرفته نویسندگان به این نتیجه رسیده‌اند که برای دستیابی به نتایج بهتر، مطلوب است که n مقداری کم (بین ۴ تا ۶) داشته باشد.
- ۳- با استفاده از الگوریتمی درهم‌ساز با خروجی ۲۸ تا ۳۲ بیت، مقدار چکیده متناظر با هر واحد محاسبه می‌شود.
- ۴- با استفاده از مقادیر چکیده به دست آمده شاخص وارون متناظر با پایگاه داده مدارک تشکیل داده می‌شود (در شاخص وارون، لیست متناظر با یک واحد n -تایی از لغات، در برگیرنده اندیس متون شامل این واحد و در صورت لزوم محل وقوع این واحدهاست).
- ۵- برای بررسی همانندی یک مدرک مشکوک با مدارک موجود در پایگاه داده، تعداد واحدهای مشترک بین آن‌ها محاسبه می‌شود.
- ۶- در صورتی که تعداد واحدهای همانند تشخیص داده شده از مقداری آستانه‌ای (در این مقاله ۶ واحد معادل ۲۴ کلمه) بیشتر باشد، دو مدرک دارای همانندی تشخیص داده می‌شوند.
- ۷- با توجه به محل وقوع هر واحد در هر مدرک، قسمت‌های همانند بین هر مدرک منبع همانند تشخیص داده شده و مدرک مشکوک مشخص می‌شود.
- ۸- فرض کنیم D_1 و D_2 مدرک مشکوک و مدرک منبع باشند. در این صورت، تنها بازه‌هایی از واحدها از D_1 را که شرایط زیر برای آن‌ها برقرار باشد، به عنوان قسمت‌های نهایی همانند تشخیص داده شده نگه داشته و مابقی حذف می‌شوند:
 - a. اولین و آخرین واحد از این بازه در D_2 موجود باشد.
 - b. بازه دارای حداقل ۲۰ واحد (احتمالاً هم‌پوشان) باشد که در D_2 نیز موجود باشند.

¹ Stemming

² Diacritic

c. فاصله بین هر دو واحد همسایه در این بازه در D_1 ، در مدرک D_2 از مقداری آستانه‌ای بیشتر نباشد (در این مقاله، این فاصله ۴۹ واحد در نظر گرفته شده است).

۲-۱-۴- روش ارائه شده توسط مینایی^۱ و نیک‌نام^۲ در روش ارائه شده در (Minaei and Niknam, 2016):

- ۱- در ابتدا متون به واحدهایی n -تایی از کلمات تقسیم‌بندی شده و به هر n -تایی به ترتیب یک شماره اختصاص می‌یابد.
- ۲- برای مقایسه یک مدرک مشکوک و یک مدرک منبع احتمالی، ماتریس متناظر با n -تایی‌های ساخته شده در آن‌ها به صورت زیر ساخته می‌شود:
a. سطرهای ماتریس به n -تایی‌های مدرک مشکوک و ستون‌های آن به n -تایی‌های مدرک منبع احتمالی اختصاص می‌یابد. در ادامه، یک عضو این ماتریس برابر با یک خواهد بود اگر n -تایی‌های متناظر با یکدیگر یکسان باشند.
- ۳- در صورتی که درایه‌های (i, j) -ام و $(i+1, j+1)$ -ام ماتریس برابر با یک باشند، آن‌گاه محل‌های متناظر با این درایه‌ها با یکدیگر ترکیب شده تا بزرگ‌ترین قسمت‌های همانند مدارک شناسایی شوند.
- ۴- برای کم‌تر کردن تعداد موارد همانند تشخیص داده شده، در صورتی که فاصله بین دو مورد از ضربی از طول آن‌ها کمتر باشد با یکدیگر ادغام می‌شوند.
- ۵- موارد همانند تشخیص داده شده‌ای که طول آن‌ها از مقداری آستانه‌ای کمتر باشد، حذف می‌شوند.

¹ Minaei

² Niknam

۲-۱-۵- روش ارائه شده توسط ^۱یرا و ^۲ان جی در روش اول ارائه شده در (Yerra and Ng, 2005)،

- ۱- در ابتدا ایست‌واژه‌ها حذف شده و فرایند ریشه‌یابی لغوی انجام می‌شود.
 - ۲- هر مدرک به جملات و هر جمله به دنباله‌های ۴-کاراکتری تقسیم‌بندی شده و ۳ دنباله ۴-کاراکتری با کمترین وقوع به عنوان نماینده‌های یک جمله انتخاب می‌شوند. فراوانی واحدهای ۴-کاراکتری با استفاده از اطلاعات آماری از یک پایگاه داده از مدارک استخراج می‌شود.
 - ۳- واحدهای متناظر جملات با هم مقایسه شده و در صورت یکسانی این واحدها، دو جمله همانند تشخیص داده می‌شوند.
- لازم به ذکر است که این روش تنها قادر به تشخیص کپی دقیق می‌باشد و در قسمت‌های تغییر یافته مدارک ضعیف عمل می‌کند.

۲-۱-۶- روش ارائه شده توسط ولاسکوئز^۳ و همکاران

در (Velásquez et al., 2016)، گزارش کارهای انجام شده برای طراحی سامانه‌ای برای همانندجویی در متون اسپانیایی ارائه شده است. علاوه بر الگوریتم همانندجوی ارائه شده، نویسندگان توصیفی از سایر قابلیت‌های سیستم ارائه شده مانند بررسی عدم سوء استفاده از سامانه با استفاده از ترندهای فنی شناخته شده، قابلیت همانندجویی روی وب، همانندجویی بین زبانی، قابلیت همانندجویی مبتنی بر ارجاع و ... ارائه کرده‌اند. در ادامه، تنها الگوریتم همانندجویی به کار برده شده در این سامانه را به طور خلاصه بازبینی می‌کنیم:

۱- برای انتخاب مدارک منبع احتمالی:

a. روی هر مدرک موجود در پایگاه داده پیش‌پردازش زیر انجام می‌شود:

i. ایست‌واژه‌ها حذف می‌شوند.

¹ Yerra

² Ng

³ Velásquez

- ii. مدرک به دنباله‌های n -تایی از لغات ($n = 4$) تقسیم‌بندی می‌شود.
- iii. دنباله‌های n -تایی به گروه‌هایی k -تایی تقسیم‌بندی می‌شوند.
- iv. n -تایی‌های موجود در هر گروه مرتب شده و l تا از آن‌ها به عنوان نماینده گروه انتخاب و مابقی در راستای کاهش تعداد مقایسات و افزایش کارایی روش حذف می‌شوند. نحوه مرتب‌سازی و انتخاب نماینده‌ها در مقاله مورد بررسی بیان نشده، اما احتمالاً به گونه‌ای است که n -تایی‌های کم کاربردتر که بیشتر توصیف‌کننده یک مدرک هستند نگهداری شده و مابقی حذف شوند.
- b. پیش‌پردازش بیان شده بر روی مدرک مشکوک انجام می‌شود.
- c. مدارکی که حداقل دارای یک اشتراک با منبع مشکوک هستند به عنوان مدارک منبع احتمالی در نظر گرفته می‌شوند.
- ۲- در فاز "تطبیق متن" یا جستجوی دقیق، پیش‌پردازش دیگری و به صورت زیر روی مدارک منبع احتمالی و مدرک مشکوک انجام می‌شود:
- a. بدون حذف ایست‌واژه‌ها، مدارک به دنباله‌های n -تایی (مقداری کمتر از مقدار قبل، در اینجا $n = 3$) از لغات تقسیم‌بندی می‌شوند.
- b. دنباله‌های n -تایی مرتب‌سازی شده و l تا از آن‌ها انتخاب می‌شوند.
- c. مقایسه بین l -تایی‌های مدرک مشکوک و هر مدرک منبع انتخاب شده انجام گرفته و در صورتی که تعداد n -تایی‌های مشترک از مقداری آستانه‌ای بیشتر باشد، مدرک مربوطه به عنوان منبع اعلام می‌گردد.
- لازم به ذکر است که نویسندگان بنا به دلایل اقتصادی و حق نشر، جزئیات دقیق و مقدار پارامترهای استفاده شده در روش را بیان نکرده‌اند.

۲-۱-۷- روش ارائه شده توسط شربینین^۱ و بوتاکف^۲

روش ارائه شده در (Scherbinin and Butakov, 2009) از الگوریتم‌های تطبیق رشته تقریبی (الگوریتم levenshtein) برای همانندجویی استفاده می‌کند. در این روش:

- ۱- مدارک منبع به واحدهایی با طول مناسب تقسیم می‌شوند.
 - ۲- مقدار چکیده متناظر با هر واحد با استفاده از الگوریتمی درهم‌ساز محاسبه شده و در کنار آن در پایگاه داده نگهداری می‌شود.
 - ۳- برای بررسی همانندی یک مدرک مشکوک با مجموعه مدارک موجود در پایگاه داده، در ابتدا پیش‌پردازش انجام شده بر روی مدارک منبع بر روی مدرک مشکوک نیز انجام می‌شود (مراحل ۱ و ۲).
 - ۴- انتخاب مدارک منبع احتمالی به این صورت است که مجموعه چکیده‌های متناظر با مدرک مشکوک با مجموعه چکیده‌های متناظر با هر مدرک منبع مقایسه شده و در صورت وجود حداقل یک مقدار مشترک، مدرک موردنظر به عنوان مدرک منبع احتمالی در نظر گرفته می‌شود.
 - ۵- از الگوریتم levenshtein برای بررسی میزان همانندی هر واحد از متن مشکوک با هر واحد از هر متن منبع احتمالی و به هم چسباندن قسمت‌های همانند تشخیص داده شده استفاده می‌شود.
- جزئیات دقیق‌تری از روش ارائه شده در این مقاله موجود نمی‌باشد.

۲-۱-۸- روش ارائه شده توسط استاماتاتوس^۳

در (Stamatatos, 2011) روشی جدید برای همانندجویی ارائه شده که تا حدودی برعکس سایر روش‌ها عمل می‌کند. یکی از پیش‌پردازش‌های رایج در اکثر روش‌های همانندجویی حذف ایست‌واژه‌ها می‌باشد اما، روش ارائه شده در (Stamatatos, 2011)، بر عکس این پیش‌پردازش

¹ Scherbinin

² Butakov

³ Stamatatos

عمل کرده و همه متن جز ایست‌واژه‌ها را حذف می‌کند و تنها از الگوی موجود در ایست‌واژه‌ها برای همانندجویی استفاده می‌کند. ایده استفاده شده در این روش این است که ایست‌واژه‌های موجود در یک متن به طور معمول وابسته به الگوی نحوی است. بنابراین، دنباله ایست‌واژه‌ها نکاتی را درباره ساختار نحوی یک مدرک افشا می‌کند و معمولاً این ساختار نحوی در عملیات سرقت علمی ثابت می‌ماند؛ یعنی در هنگام ارتکاب سرقت علمی، معمول‌ترین کار جایگزینی یک لغت یا عبارت با لغات و عبارات مترادف است و بازنویسی و تغییر ساختار نحوی یک مدرک کاری سخت‌تر بوده و کمتر انجام می‌شود.

لازم به ذکر است که ایست‌واژه‌ها قبل از این کار توسط نویسندگان برای تعیین ژانر متن نیز استفاده شده و در آن موضوع نیز نتایج خوبی به دست آمده است.

این روش، با توجه به عدم حساسیت به تغییر ترتیب میان کلمات و جایگزینی یک کلمه با مترادف‌های آن، در همانندجویی متون شدیداً تغییر یافته بهتر عمل می‌کند. جزئیات این روش به شرح زیر است:

- ۱- لیست ۵۰ ایست‌واژه پر کاربرد استخراج شده است.
- ۲- تمام قسمت‌های متون همه مدارک به جز کلمات موجود در مجموعه ۵۰ ایست‌واژه پر کاربرد حذف می‌شود.
- ۳- مدارک جدید به دنباله‌های n -تایی از لغات هم‌پوشان تقسیم‌بندی می‌شوند (نسبت به روش‌های مبتنی بر کاراکتر، در اینجا، انتخاب مقدار مناسب n بسیار مهم‌تر می‌باشد، زیرا مقادیر کوچک n قسمت‌های زیادی از متن را همانند تشخیص داده و مقادیر بزرگ n ، قسمت‌های همانند زیادی را در نظر نخواهد گرفت).
- ۴- برای مدرک مشکوک نیز مراحل فوق دنبال شده و دنباله‌های n -تایی هم‌پوشان از ایست‌واژه‌های متناظر با آن به دست می‌آید.
- ۵- دنباله‌هایی که احتمال وقوع آن‌ها بالا بوده و باعث تشخیص همانندی مثبت اشتباه شده، شناسایی شده و با اعمال شرایطی تشخیص آن‌ها نادیده گرفته می‌شود. برای مثال ۶ ایست‌واژه

- پیر کاربرد شناسایی شده و دنباله‌های n -تایی که تنها شامل این ایست‌واژه‌ها بوده در همانندجویی مورد بررسی قرار نمی‌گیرند، دنباله‌های پیر کاربرد در نظر گرفته نمی‌شوند و ...
- ۶- تمام دنباله‌های n -تایی متناظر با مدرک مشکوک را در دنباله‌های n -تایی متناظر با مدارک مرجع جستجو کرده و تطابق‌ها استخراج می‌شوند.
- ۷- قسمت‌های همانند تشخیص داده شده استخراج شده و فرایند فوق به ازای مقداری جدید و کمتر از مقدار قبل برای پارامتر n تکرار می‌شود. انجام این عمل در راستای شناسایی قسمت‌های شدیداً تغییر یافته می‌باشد.
- ۸- برای بهبود نتایج و حذف قسمت‌های به اشتباه همانند تشخیص داده شده، قسمت‌های همانند تشخیص داده شده با طول بسیار کم حذف شده و قسمت‌های با طول کم نیز در صورتی که میزان همانندی قسمت‌های اصلی متن متناظر با آن‌ها از مقداری کمتر باشد حذف خواهند شد.

۲-۱-۹- روش ارائه شده توسط گیپ^۱ و ماسچه^۲

در روش ارائه شده در (Gipp and Meuschke, 2011)، نویسندگان از ارجاعات و منابع موجود در مدارک برای همانندجویی استفاده کرده‌اند. دلایل استفاده از چنین رویکردی توسط نویسندگان به صورت زیر بیان شده است:

- ۱- مدارکی که دارای منابع مشترک باشند به لحاظ معنایی به هم نزدیک خواهند بود. علاوه بر این، ارجاع به یک منبع توسط دو مدرک در شرایطی که تعداد کل ارجاعات به منبع مورد نظر کم باشد، شک برانگیز است.
- ۲- در صورت وجود الگوی مشترک مابین ارجاعات دو مدرک، احتمال همانندی دو مدرک بیشتر می‌باشد.

^۱ Gipp

^۲ Meuschke

با توجه به این نکات، در (Gipp and Meuschke, 2011) الگوریتمی ارائه شده که در آن، پس از تشکیل دنباله ارجاعات دو مدرک، پارامترهای زیر توسط الگوریتم‌های مناسب محاسبه شده و همانندی موجود بین دو مدرک محاسبه می‌شود:

- ۱- تعداد منابع مشترک با شمارش آن‌ها،
 - ۲- بزرگ‌ترین الگوی ارجاع مشترک بین دنباله‌های ارجاعات با استفاده از الگوریتم محاسبه بزرگ‌ترین زیر رشته مشترک (LCS)،
 - ۳- تمام زیر دنباله‌های مشترک از ارجاعات در دو دنباله،
 - ۴- همانندی‌های محلی بین دو مدرک با استفاده از روش‌های قسمت بندی^۱ (شبیه به آنچه در روش‌های مبتنی بر کاراکتر برای همانندجویی در متن انجام می‌شد).
- لازم به ذکر است که در صورتی که تغییری در منابع مدارک انجام نشده باشد، این روش حتی قادر به تشخیص همانندی مدارک به شدت تغییر یافته می‌باشد. اما، در صورت تغییر منابع و جایگزینی منابع یک مدرک با منابعی دیگر، این روش قدرتی در همانندجویی نخواهد داشت.

۲-۲- روش‌های همانندجویی مبتنی بر بردار

ویژگی‌های نحوی و لغوی را می‌توان با بردار نمایش داد و در ادامه از ضرایب شباهت برداری برای بررسی همانندی آن‌ها استفاده کرد. برخی از روش‌های همانندجویی ارائه شده از این رویه استفاده می‌کنند. در این روش‌ها، مدارک به بردار تبدیل شده، در ادامه همانندی بردار متناظر با مدرک مشکوک با بردار متناظر با هر مدرک منبع با استفاده از معیارهایی مانند جاکارد^۲، کوساین^۳، در برگرفتنی^۴ و ... بررسی شده و همانندی یا ناهمانندی مدارک مشخص می‌شود. این دسته از روش‌های همانندجویی قادر به تشخیص قسمت‌های دقیقاً مشابه، قسمت‌های با تغییرات کم و همچنین قسمت‌های حاصل از تغییر ترتیب حضور کلمات و عبارات می‌باشند. در ادامه برخی از روش‌های موجود در این دسته مورد بازبینی قرار می‌گیرند.

¹ Chunking

² Jaccard

³ cosine

⁴ Containment

۲-۲-۱- روش ارائه شده توسط بارون-سدنو^۱ و همکاران

روش ارائه شده در (Barrón-Cedeño, 2010) نیز از کدگذاری طول کلمه به جای کلمه برای افزایش سرعت محاسبات و کاهش فضای موردنیاز برای ذخیره‌سازی داده‌های اضافه موردنیاز برای همانندجویی استفاده می‌کند. در این روش:

۱- مدارک جدیدی با جایگذاری طول هر کلمه به جای آن کلمه ایجاد می‌شود. با استفاده از این رویکرد، داده‌های جدیدی به وجود می‌آید که الفبای آن اعداد ۱ تا ۹ است (طول کلمات بیشتر از ۹ برابر با ۹ در نظر گرفته شده است) و در ادامه همانندجویی بر روی داده‌های جدید انجام می‌شود.

۲- مدارک جدید به دست آمده به دنباله‌هایی n -تایی تقسیم‌بندی می‌شوند.

۳- از معیار شباهت جاکارد، که به صورت زیر تعریف می‌شود، برای بررسی میزان همانندی متن مشکوک با هر یک از متون منبع استفاده می‌شود:

$$j(d_1, d_2) = \frac{|V_n(d_1) \cap V_n(d_2)|}{|V_n(d_1) \cup V_n(d_1)|} \quad (۲)$$

که d_1 و d_2 مدارک حاصل از کدگذاری مدارک مورد بررسی و $V_n(.)$ مجموعه تمام n -تایی‌های موجود در کدگذاری شده متناظر با یک مدرک می‌باشد.

۲-۲-۲- روش ارائه شده توسط موروگیسان^۲ و همکاران

در روش ارائه شده در (Murugesan, 2010) فرض شده است که مدارک به طور مستقیم در دسترس نمی‌باشند. برای مثال دو کنفرانس در زمینه‌ای مشترک وجود دارند که با توجه به محرمانگی مقالات ارسال شده به آن‌ها نمی‌توانند دسترسی به مقالات دریافتی را به کنفرانس دیگر بدهند اما قصد دارند که مطمئن شوند که مقالات دریافتی، به کنفرانس دیگر ارسال نشده‌اند. به عبارتی دیگر، این کنفرانس‌ها قصد دارند از ارسال چندباره و همزمان مقالات پیشگیری نمایند. در (Murugesan,

¹ Barron-Cedeno

² Murugesan

(2010) روشی امن و کارا برای مقایسه دو مدرک یا قسمت‌هایی از دو مدرک ارائه شده که نسبت به محاسبات چندنفره امن^۱ بسیار کاراتر می‌باشد. جزئیات این روش به شرح زیر می‌باشد:

۱- در ابتدا هر مدرک به صورت برداری نمایش داده می‌شود. بردار مورد نظر به این صورت ساخته شده که فرض می‌شود D برداری از لغات و عبارات ممکن در زبان باشد. در این صورت، مولفه i -ام بردار متناظر با یک مدرک برابر با تعداد تکرار i -امین لغت D در آن مدرک است.

۲- در ادامه هر کاربر (کنفرانس) شکل نرمال بردار متناظر با متن خود را محاسبه می‌کند.

۳- ضرب داخلی دو بردار نرمال شده به صورت توزیع شده و امن (معیار کوساین^۲) توسط دو کاربر و به عنوان میزان همانندی دو متن محاسبه می‌شود:

$$\text{cosine}(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \cdot \|v_2\|} \quad (3)$$

که v_1 بردار متناظر با مدرک اول، v_2 بردار متناظر با مدرک دوم، $\cdot$ نمایانگر ضرب داخلی دو بردار و $\|\cdot\|$ اندازه یا نرم اقلیدسی یک بردار می‌باشد.

۴- با توجه به مقدار به دست آمده و در صورت بیشتر بودن از مقداری آستانه‌ای همانندی دو متن مورد نظر نتیجه‌گیری می‌شود.

۲-۲-۳ روش ارائه شده توسط بارون-سدنو و همکاران

در روش ارائه شده در (Barrón-Cedeño and Rosso, 2009):

۱- ابتدا پیش‌پردازش‌هایی شامل حذف ایست‌واژه‌ها و ریشه‌یابی لغوی انجام می‌شود.

۲- مدارک منع به واحدهای n -تایی از لغات تقسیم می‌شوند.

۳- مدرک مشکوک به جملات تقسیم‌بندی شده و سپس هر جمله از مدرک مشکوک به واحدهای n -تایی از لغات تقسیم می‌شود.

¹ Secure multi party computation

² cosine

۴- همانندی مدرک مشکوک با هر یک از مدارک منبع با استفاده از معیار دربرگرفتنی که به

صورت زیر تعریف می شود، محاسبه می شود:

$$C(d_1, d_2) = \frac{|N(d_1) \cap N(d_2)|}{|N(d_1)|} \quad (۴)$$

که d_1 مدرک مشکوک، d_2 یک مدرک منبع و $N(.)$ تعداد n -تایی های متن مورد نظر می باشد.

نویسندگان روش ارائه شده را برای مقادیر مختلف n آزمایش کرده و در نهایت نتیجه گیری کرده اند که در هنگام تقسیم متن به واحدهایی از کلمات، بهترین گزینه ها انتخاب n برابر با ۲ یا ۳ می باشد.

۲-۲-۴- روش ارائه شده توسط وایت^۱ و جوی^۲

در روش ارائه شده در (White and Joy, 2004)،

- ۱- مدارک منبع به واحدهایی برابر با جمله تقسیم بندی می شوند.
- ۲- تمامی حروف به حالت کوچک تبدیل شده، ایست و اژه ها از مدارک حذف شده و از کلمات تکراری در هر جمله تنها یکی از آن ها نگهداری می شود.
- ۳- مراحل ۱ و ۲ برای مدرک مشکوک نیز انجام می گیرد.
- ۴- هر جمله مدرک مشکوک با هر جمله هر مدرک منبع مورد مقایسه قرار می گیرد. برای انجام مقایسه، معیار زیر که وابسته به تعداد کلمات مشترک بین دو جمله و طول جملات است محاسبه می شود:

$$Sim(S_i, S_j) = \frac{|CW|}{2|S_i|} + \frac{|CW|}{2|S_j|} \quad (۵)$$

که CW مجموعه کلمات مشترک بین دو جمله و $|.|$ اندازه یک مجموعه است. دو جمله

مورد بررسی همانند تشخیص داده شده اگر $Sim(.,.) \geq T_1$ یا $|CW| \geq T_2$.

¹ White

² Joy

- ۵- با توجه به مقادیر محاسبه شده، مدارک منبع را بر حسب میزان همانندی با متن مشکوک رتبه‌بندی کرده و تعدادی مشخص از آن‌ها به عنوان مدارک منبع احتمالی تعیین می‌شوند.
- ۶- می‌توان با رسم نمودار برداری همانندی موجود بین دو متن (مانند آنچه در روش ارائه شده در (Basile et al., 2009) انجام شد)، همانندی‌های یافته شده بین دو متن را به صورت تصویری به کاربر نمایش داد و کاربر می‌تواند بدین طریق از بین مدارک منبع احتمالی، مدارک کمتری را انتخاب و به صورت دستی بررسی کند.

۲-۲-۵- روش‌های ارائه شده توسط کنگ^۱ و همکاران

در روش‌های ارائه شده در (Kong et al., 2013) و (Kong et al., 2015)، از تلفیقی از تکنیک‌های بازیابی اطلاعات و روش‌های مبتنی بر بردار، برای همانندجویی استفاده شده است. جزئیات اقدامات انجام شده در این روش‌ها به صورت زیر می‌باشد:

- ۱- مدارک مشکوک به قسمت‌هایی (شامل ۵ جمله) تقسیم‌بندی می‌شود.
- ۲- با استفاده از یک مدل رتبه‌بندی، کلمات و عبارات موجود در هر قسمت رتبه‌بندی شده (در این مقاله، رتبه‌بندی لغات با استفاده از روش tf-idf انجام شده است) و کلمات کلیدی هر قسمت مشخص می‌شود.
- ۳- با استفاده از یک موتور جستجو، منابع در برگرفته کلمات کلیدی هر بخش استخراج می‌شود.
- ۴- از روش‌های همانندجویی مبتنی بر بردار و معیار کوساین استفاده شده و همانندی مدارک مشکوک با هر یک از مدارک منبع انتخابی در سطح جمله مورد بررسی قرار می‌گیرد.
- ۵- در صورتی که جملات همانند تشخیص داده شده از مقداری به هم نزدیک‌تر باشند (چه در منبع مشکوک و چه در منابع احتمالی)، جملات با یکدیگر ترکیب شده و در صورتی که اندازه قسمت‌های نهایی همانند تشخیص داده شده از مقداری کوچک‌تر باشد، آن قسمت‌ها حذف می‌شوند.

¹ Kong

۲-۲-۶- روش ارائه شده توسط ممتاز^۱ و همکاران

در روش ارائه شده در (Momtaz et al., 2015)، نویسندگان از نمایش گرافی و تشخیص گره‌های همانند گراف برای همانندجویی استفاده کرده‌اند. در این روش:

- ۱- پیش‌پردازشی شامل حذف ایست‌واژه‌ها و نقطه‌گذاری‌ها بر روی مدارک اعمال می‌شود.
 - ۲- مدارک به جملات تقسیم‌بندی می‌شوند.
 - ۳- برای مقایسه مدرک مشکوک با یک مدرک منبع احتمالی، همانندی هر جمله از مدرک مشکوک با هر جمله از مدرک منبع:
- a. معیار همانندی کوساین بیان شده در فرمول (۳) برای دو مدرک محاسبه می‌شود. در صورتی که مقدار محاسبه شده از ۰,۴ کمتر باشد، این دو جمله غیر همانند تشخیص داده شده در غیر این صورت،
- b. متناظر با هر یک از دو جمله یک گراف ساخته می‌شود:
- i. گراف مورد نظر به این شکل ساخته شده که یک گره متناظر با هر لغت از جمله در نظر گرفته شده و هر گره به وسیله یک یال به گره‌های متناظر با هر یک از ۴ لغت قبل و بعد از خود در جمله متصل می‌شود.
- c. همانندی هر گره از جمله اول با هر گره از جمله دوم با استفاده از معیار زیر محاسبه شده و در صورتی که مقدار آن بیش از مقداری آستانه‌ای باشد (در اینجا ۰,۴)، دو گره همانند تشخیص داده می‌شوند:

$$Sim(A, B) = \frac{A \cap B}{\max(len(A), len(B))} \quad (۶)$$

- که A و B به ترتیب همسایه‌های مستقیم گره اول و دوم، $\max(.,.)$ بیانگر مقدار بیشتر و $len(.,)$ بیانگر تعداد همسایه‌های مستقیم یک گره می‌باشد.
- d. در صورتی که تعداد گره‌های همانند متناظر با دو جمله از مقداری آستانه‌ای (یک سوم کلیدواژه‌های مدرک مشکوک) بیشتر باشد، دو جمله همانند تشخیص داده می‌شوند.

¹ Momtaz

۴- جملات همانند تشخیص داده شده که فاصله آن‌ها نسبت به هم در متن از مقداری کمتر باشد ترکیب شده و قسمت‌های همانند نهایی آماده می‌شوند.

۲-۲-۷- روش ارائه شده توسط قروی^۱ و همکاران

در روش ارائه شده در (Gharavi et al., 2016)، نویسندگان از روش‌های یادگیری عمیق برای همانندجویی استفاده کرده‌اند. در این روش، با استفاده از روش‌های شبکه عصبی عمیق^۲، هر لغت با یک بردار نمایش داده می‌شود که این بردار حاوی اطلاعات معنایی و گرامری در مورد آن لغت می‌باشد. این روش برای مسابقات همانندجویی در زبان فارسی طراحی شده و جزئیات چندان دقیقی درباره آن بیان نشده است. اما، به صورت کلی، مراحل اجرای این روش به شرح زیر است:

- ۱- ایست‌واژه‌ها از مدارک حذف می‌شوند.
- ۲- مدارک به جملات تقسیم‌بندی می‌شوند.
- ۳- بردار متناظر با هر جمله به صورت میانگین بردار لغوی متناظر با لغت‌های آن جمله محاسبه می‌شود که بردار لغوی متناظر با هر لغت با استفاده از روش‌های شبکه عصبی عمیق مانند skip-grams یا continuous bag of words و با استفاده از مجموعه بسیار بزرگی از مدارک محاسبه می‌شود.
- ۴- از معیارهای همانندی جاکارد و کوساین (فرمول‌های (۲) و (۳)) استفاده شده و همانندی برداری بردارهای متناظر جملات و همانندی لغوی جملات محاسبه می‌شوند و در صورتی که مقادیر آن‌ها از مقدارهایی آستانه‌ای بیشتر باشد، دو جمله همانند تشخیص داده می‌شوند. مقادیر آستانه‌ای با انجام آزمایش محاسبه می‌شوند.

¹ Gharavi

² Deep neural network

۲-۳- روش‌های همانندجویی مبتنی بر گرامر

در برخی از روش‌های همانندجویی موجود، نویسندگان از ویژگی‌های نحوی برای بررسی همانندی متون و تشخیص سرقت علمی استفاده کرده‌اند. در این روش‌ها، برای دستیابی به نتایج بهتر نسبت به دسته‌های پیش‌تر بررسی شده از روش‌های همانندجویی، با استفاده از ابزارهای گرامری مدارک پردازش شده و همانندجویی بر روی مدارک پردازش شده انجام می‌گیرد. این دسته از روش‌های همانندجویی قادر به تشخیص همانندی‌های از نوع بازنویسی شده می‌باشند که دقت روش حاصل وابسته به ابزار گرامری مورد استفاده و دقت آن خواهد بود. در ادامه این بخش، برخی از روش‌های موجود در این دسته به طور خلاصه مورد بازبینی قرار خواهند گرفت.

۲-۳-۱- روش‌های ارائه شده توسط الهادی^۱ و ال-توبی^۲

اساس روش‌های ارائه شده توسط الهادی و ال-توبی در (Elhadi and Al-Tobi, 2009) و (Elhadi and Al-Tobi, 2008) بر این مهم استوار است که یک لغت یا عبارت را تنها می‌توان با لغت یا عبارتی هم نوع جایگزین کرد. به بیان دقیق‌تر، برای جایگزینی فعل، به طور معمول از فعلی مترادف با آن استفاده می‌شود، برای جایگزینی یک اسم، معمولاً از یک اسم مترادف با آن استفاده می‌شود، برای جایگزینی یک قید معمولاً از یک قید مترادف با آن قید استفاده می‌شود و ... در نتیجه، دنباله اقسام کلمه^۳ متناظر با قسمتی از یک متن و دنباله اقسام کلمه حاصل از بازنویسی آن قسمت از متن یکسان خواهد بود. در روش‌های ارائه شده در این مقالات:

۱- از یک تشخیص‌دهنده اقسام کلمه^۴ استفاده شده و اقسام کلمه استفاده شده در مدارک منبع محاسبه و دنباله محاسبه شده به عنوان داده اضافه برای همانندجویی نگهداری می‌شود. در این روش، هر چه انواع اقسام کلمه در نظر گرفته شده بیشتر باشد، روش دقیق‌تر شده اما از

¹ ElHadi

² Al-Tobi

³ Part of speech

⁴ Part of speech recognizer

کارایی روش کاسته خواهد شد. فرایند ذکر شده موجب می‌شود که به دنباله‌ای از نمادها برای مدارک منبع و مدرک مشترک دسترسی داشته باشیم. در نتیجه می‌توان از الگوریتم‌های تحلیل رشته موجود در زیست‌شناسی برای تشخیص همانندی در این دنباله از نمادها استفاده کرد.

۲- با استفاده از الگوریتم بزرگ‌ترین زیررشته مشترک^۱ (LCS) به دنبال متونی گشته که میزان همانندی آن‌ها از مقداری آستانه‌ای بیشتر باشد.

لازم به ذکر است که این روش با هدف تشخیص صفحات وب با محتویات یکسان ارائه شده است. همچنین ذکر این نکته نیز ضروری به نظر می‌رسد که با توجه به نحوه کار این روش، این روش در برابر بازمرتب‌سازی متون غیر کارا بوده و قادر به تشخیص این نوع همانندی نمی‌باشد.

۲-۳-۲ روش ارائه شده توسط عثمان^۲ و همکاران

در روش ارائه شده در (Osman et al., 2012):

- ۱- ایست‌واژه‌ها حذف شده و فرایند ریشه‌یابی لغوی انجام می‌شود.
- ۲- متون به جملات تقسیم‌بندی می‌شوند.
- ۳- از "برچسب‌زنی نقش معنایی"^۳ برای تبدیل جملات به آرگومان‌ها و تعیین نقش هر مولفه و برحسب محل حضور هر مولفه در جمله استفاده می‌شود.
- ۴- آرگومان‌های به دست آمده به عنوان گره‌های یک گراف در نظر گرفته شده و تمام آرگومان‌های هم‌نوع به یک گره با نام آن نوع متصل می‌شوند.
- ۵- با استفاده از فرهنگ جامع شبکه‌واژگانی^۴، تمام مفاهیم متناظر با هر مولفه از انواع مختلف آرگومان موجود به دست آمده و به عنوان امضا موضوعی متن نگهداری می‌شود.
- ۶- برای مقایسه دو متن، امضا موضوعی دو متن با یکدیگر مقایسه می‌شود.

¹ Longest common subsequence

² Osman

³ Semantic role labeling

⁴ Wordnet

۷- در صورت وجود عنوان مشترک مابین دو امضا، جملات بیانگر آن‌ها مشخص شده و با یکدیگر مقایسه می‌شوند. در انجام مقایسه ما بین دو جمله، تنها آرگومان‌های هم نوع با یکدیگر مقایسه می‌شوند (برعکس روش‌های قبل که تمام کلمات جمله‌ها با یکدیگر مقایسه می‌شد. مثلاً در روش مقایسه n -تایی‌ها، هر n -تایی با هر n -تایی دیگر مقایسه می‌شد. اما در اینجا تنها، فعل با فعل، زمان با زمان و ...). در نتیجه این روش دقیق‌تر بوده و زمان اجرای کمتری نیاز دارد.

در پایان، نویسندگان اهمیت هر یک از انواع آرگومان را در نظر گرفته و نقش آن‌ها در همانندی دو جمله را بررسی کرده‌اند. بر این اساس و با انجام مجموعه‌ای از آزمایشات، یک پارامتر وزن به هر قسم کلمه اختصاص داده و از آن در تعیین همانندی دو جمله استفاده کرده‌اند. به عنوان مثال، آزمایشات انجام شده نشان‌دهنده این هستند که افعال بیشترین نقش را در جملات داشته و در نتیجه در روش نهایی ارائه شده در (Osman et al., 2012) وزن بیشتری به افعال داده شده است.

۲-۳-۳- روش ارائه شده توسط رام^۱ و همکاران

در (Ram et al., 2014) با یافتن الگوهای بازنویسی یک متن توسط کاربر انسانی، روشی برای همانندجویی ارائه شده که الگوهای موردنظر را در نظر گرفته و با اعمال قواعدی برای خنثی کردن اثر اعمال هر یک از این الگوها همانندجویی را انجام می‌دهد. برای به دست آوردن الگوهای مورد استفاده، از یک مجموعه منبع شامل متون همانند تولید شده توسط کاربر انسانی استفاده شده و الگوهای به کار رفته برای بازنویسی متون به دست آمده است. در نهایت با استفاده از این الگوها مجموعه‌ای از قوائد نحوی و معنایی به دست آمده است. با توجه به وابستگی قواعد به اطلاعات نگارشی (نحوی)، این روش نیازمند به انجام پیش‌پردازشی شامل اعمال تشخیص‌دهنده اقسام کلمه بر روی متن، قسمت‌بندی متن و اعمال موتور "میدان تصادفی شرطی"^۲ برای تشخیص "it"‌های زائد بر روی متون می‌باشد. در ادامه و در هنگام بررسی همانندی، در صورت وجود شرایط مشخص و احتمال

¹ Ram

² Conditional random field

ارتکاب نوع مشخصی از بازنویسی، قاعده متناظر اعمال شده و تا در صورت وجود آن نوع بازنویسی، روش قادر به تشخیص آن باشد.

۲-۴- روش‌های همانندجویی معنایی

متون می‌توانند دارای معانی یکسان بوده اما از نقطه نظر نگارشی به گونه‌ای متفاوت نوشته شده باشند. در روش‌های همانندجویی معنایی، سعی بر آن است که محتوای معنایی جملات استخراج شده و همانندجویی بر این اساس انجام شود. ابزارهای مهم مورد استفاده در این گونه روش‌ها، لغت‌نامه‌ها و فرهنگ‌های جامع مانند وردنت می‌باشند. لازم به ذکر است که در روش‌های همانندجویی معنایی، پس از استفاده از ابزارهای معنایی، از دسته روش‌های همانندجویی قبل مانند روش‌های مبتنی بر کاراکتر یا روش‌های برداری برای همانندجویی استفاده می‌شود. استفاده از ابزارهای مورد ذکر در این دسته از روش‌ها، موجب توانایی این روش‌ها در تشخیص همانندی در متون بازنویسی شده با جایگزینی لغات و عبارات با مترادف‌های آن‌ها می‌شود. در ادامه این بخش، روش‌های مهم ارائه شده در این دسته مورد بررسی قرار می‌گیرند.

۲-۴-۱- روش ارائه شده توسط لی^۱ و همکاران

در روش ارائه شده در (Li et al., 2006)، برای بررسی همانندی دو متن، متون به جملات تجزیه شده و همانندی هر جمله از متن اول با هر جمله از متن دوم بررسی می‌شود. برای بررسی وابستگی دو جمله، وابستگی معنایی هر لغت از جمله اول با لغات جمله دوم بررسی شده است. در این روش، علاوه بر وابستگی معنایی لغات دو جمله، ترتیب حضور لغات در جمله نیز در نظر گرفته شده است. دلیل این امر این است که ممکن است جملاتی متفاوت با لغات یکسان وجود داشته باشند در حالی که تنها با استفاده صرف از وابستگی معنایی این جملات یکسان تشخیص داده می‌شوند (برای مثال جملات "تیر به توپ برخورد کرد" و "توپ به تیر برخورد کرد" دارای معانی متفاوت اما لغات

¹ Li

یکسان هستند). جزئیات دقیق‌تر روش ارائه شده در (Li et al., 2006) برای مقایسه دو جمله به صورت زیر است:

- ۱- در ابتدا اجتماع کلمات موجود در دو جمله محاسبه می‌شود.
 - ۲- با استفاده از اطلاعات محتوایی به دست آمده از یک پیکره^۱، اهمیت هر لغت به دست آمده و لغات وزن گذاری می‌شوند.
 - ۳- با استفاده از یک لغت‌نامه (در این مقاله از وردنت استفاده شده است) یک بردار معنایی خام S_j متناظر با هر یک از جملات ساخته می‌شود:
- a. برای محاسبه مقدار مولفه i -ام از بردار متناظر با یک جمله، لغت با بیشترین وابستگی معنایی در جمله دیگر به i -امین لغت از این جمله را مشخص کرده و در صورتی که همانندی معنایی دو لغت بیشتر از مقداری آستانه‌ای باشد، مقدار همانندی محاسبه شده به عنوان مقدار مولفه i -ام از بردار S_j قرار می‌گیرد. در غیر صورت مقدار این مولفه برابر با صفر قرار خواهد گرفت.

b. با استفاده از وزن اختصاص یافته به لغات، مقادیر به دست آمده به صورت زیر به روز می‌شوند:

$$S[i] = S[i].I(w).I(w') \quad (6)$$

که $S[i]$ ، i -امین مولفه بردار S ، w لغت متناظر با این مولفه، w' لغت متناظر با w در جمله دیگر و $I(.)$ وزن یا ضریب تاثیر یک لغت می‌باشد.

- ۴- با استفاده از یک لغت‌نامه (در این مقاله از وردنت استفاده شده است) یک بردار ترتیب r_j متناظر با هر یک از جملات به صورت زیر ساخته می‌شود:

a. برای محاسبه مقدار مولفه i -ام از این بردار در ابتدا، از بین لغات موجود در جمله دیگر، لغت با بیشترین وابستگی معنایی به لغت i -ام از این جمله مشخص می‌شود. در صورتی که همانندی معنایی دو لغت از مقداری آستانه‌ای باشد، اندیس لغت با

¹ Corpus

بیشترین همانندی در جمله دوم به عنوان مقدار مولفه i -ام از بردار r_j قرار می‌گیرد. در غیر این صورت مقدار این مولفه برابر با صفر قرار داده می‌شود.

۵- با توجه به مقادیر محاسبه شده، همانندی معنایی دو جمله به صورت زیر محاسبه می‌شود:

$$S_s = \frac{S_1 \cdot S_2}{\|S_1\| \cdot \|S_2\|} \quad (7)$$

که $\| \cdot \|$ اندازه یک بردار می‌باشد.

۶- همانندی ترتیبی دو جمله به صورت زیر محاسبه می‌شود:

$$S_r = 1 - \frac{\|r_1 - r_2\|}{\|r_1 + r_2\|} \quad (8)$$

۷- با ترکیب دو معیار به دست آمده از بالا، معیار همانندی زیر به عنوان میزان همانندی کلی دو جمله محاسبه می‌شود:

$$S(T_1, T_2) = \delta \cdot S_s + (1 - \delta) \cdot S_r \quad (9)$$

که مقدار بهینه $0.5 \leq \delta \leq 1$ با انجام آزمایش روی مجموعه متون مورد آزمایش به دست می‌آید.

۲-۴-۲- روش ارائه شده توسط مشهدی رجب^۱ و شمس‌فرد^۲

روش ارائه شده در (Mashhadirajab and Shamsfard, 2016)، جهت ارائه به اولین دوره مسابقات pan در ایران و برای زبان فارسی ارائه شده است. با توجه به اینکه در مجموعه مدارک مسابقه در هر مدرک تنها یک نوع سرقت علمی انجام شده است در این روش، در ابتدا نوع سرقت علمی مشخص شده است. برای این کار، از روش‌های دسته‌بندی برای تشخیص نوع سرقت علمی استفاده شده است. نوع سرقت علمی می‌تواند کپی دقیق، تصادفی یا شبیه‌سازی شده باشد. روش ارائه شده در این مقاله به صورت زیر عمل می‌کند:

۱- مدارک به جملات تقسیم‌بندی شده، هر جمله به واحدهای مناسب تجزیه شده، ایست‌واژه‌ها

حذف شده و واژگان مشتق شده به فرم استاندارد بازگردانده می‌شوند.

¹ Mashhadirajab

² Shamsfard

۲- بردار tf-idf متناظر با هر جمله با استفاده از مجموعه مدارک منبع و مشکوک ساخته می‌شود.

با استفاده از معیارهای همانندی برداری کوساین (فرمول (۳)) و دایس^۱ که به صورت زیر محاسبه می‌شود:

$$Dice(S_1, S_2) = \frac{2.\delta(S_1).\delta(S_2)}{\|\delta(S_1)\|^2 + \|\delta(S_2)\|^2} \quad (۱۰)$$

که $\delta(x) = 1$ اگر $x \neq 0$ و ۰ در غیر این صورت و $\|\cdot\|$ بیانگر اندازه یا نرم اقلیدسی می‌باشد.

۳- در صورتی که هر دو معیار میزان همانندی را بالاتر از ۰,۳ اعلام کنند، جملات مورد نظر به عنوان جملات همانند علامت‌گذاری می‌شوند. برای بردارهایی که میزان همانندی آنها بین ۰,۱ تا ۰,۳ به دست آید همانندی معنایی به صورت زیر بررسی می‌گردد:

- a. با استفاده از شبکه عصبی SVM نوع سرقت علمی مشخص می‌شود.
- b. مقدار آستانه‌ای مورد نیاز برای تشخیص همانندی در جملات مورد بررسی با استفاده از نوع سرقت علمی تعیین می‌شود.
- c. با استفاده از فارس‌نت^۲، مجموعه لغات مترادف هر واژه ساخته می‌شود.
- d. هر لغت از جمله اول با هر لغت از جمله دوم مقایسه شده و در صورتی که مجموعه‌های متناظر آنها اشتراک داشته باشد، لغت جمله اول با لغت جمله دوم جایگزین می‌گردد.
- e. همانندی دو بردار دوباره محاسبه شده و میانگین آن با همانندی از قبل محاسبه شده حساب شده و در صورتی که از مقدار آستانه‌ای در نظر گرفته شده بیشتر باشد، این جملات نیز همانند تشخیص داده می‌شوند.

۴- قسمت‌هایی از مدارک که دارای جملات همانند بوده با هم ترکیب شده و همانندی معنایی برای کل قسمت محاسبه می‌شود و در صورتی که از مقداری آستانه‌ای بیشتر باشد، کل

^۱ dice

^۲ FarsNet

قسمت همانند تشخیص داده می‌شود و در صورتی که همانندی آن از مقداری آستانه‌ای کمتر باشد، جملات همانند تشخیص داده شده متناظر حذف می‌شوند. هدف از این مرحله تشخیص کامل قسمت‌های همانند و حذف قسمت‌های به اشتباه همانند تشخیص داده شده در مراحل قبل می‌باشد.

۵- قسمت‌های کوتاه همانند تشخیص داده شده حذف و از قسمت‌های هم‌پوشان تنها یکی از آن‌ها در نظر گرفته می‌شود.

لازم به ذکر است که این روش در مسابقه بیان شده مقام اول را به دلیل نتایج بهتر کسب کرده اما زمان اجرای این روش بسیار زیاد بوده و قابل مقایسه با زمان اجرای سایر روش‌ها نیست.

۲-۵- روش‌های همانندجویی فازی

در روش‌های همانندجویی فازی، تطابق میان قسمت‌های دو متن مانند جملات آن‌ها فازی بوده و به صورت صفر و یک نخواهد بود. مفهوم فازی در همانندجویی را می‌توان به این صورت تصور کرد که یک لغت یا عبارت با مجموعه‌ای از لغات و عبارات همراه بوده و به هر یک از اعضای این مجموعه مقداری مابین صفر و یک تعلق گرفته است. با توسیع این بیان، به قسمت‌های بزرگ‌تر از متن، می‌توان همانندی آن‌ها را نیز به صورت فازی بررسی کرد. با توجه به فازی بودن مقدار همانندی اختصاص یافته به دو قسمت متن، بسته به ابزار مورد استفاده برای بررسی همانندی واحدهای کوچک متن مانند لغات، این دسته از روش‌های همانندجویی قادر به تشخیص همانندی‌های از نوع بازنویسی شده هستند. در ادامه این بخش، روش‌های ارائه شده در این زمینه را بررسی خواهیم کرد.

۲-۵-۱- روش ارائه شده توسط ^۱یرا و ^۲ان‌جی

در روش دوم ارائه شده در (Yerra and Ng, 2005)، استفاده از نظریه مجموعه‌های فازی برای همانندجویی پیشنهاد شده است. در این روش:

^۱ Yerra

^۲ Ng

۱- در ابتدا ایست‌واژه‌ها حذف شده و فرایند ریشه‌یابی لغوی انجام می‌شود.

۲- هر متن به واحدهایی برابر با جملات تقسیم‌بندی می‌شود.

۳- با استفاده از نظریه مجموعه‌های فازی میزان همانندی هر دو کلمه به صورت زیر محاسبه می‌شود:

$$c_{ij} = \frac{n_{ij}}{n_i + n_j - n_{ij}} \quad (11)$$

که n_{ij} تعداد متون شامل دو کلمه w_i و w_j از یک پایگاه داده بزرگ از متون، و n_i تعداد متون شامل کلمه w_i از این پایگاه داده می‌باشد.

۴- برای تعیین میزان مشابهت جمله s_1 به جمله s_2 ، در ابتدا میزان مشابهت هر کلمه از جمله اول با هر کلمه از جمله دوم محاسبه شده و در ادامه میزان مشابهت دو جمله به صورت زیر محاسبه می‌شود:

$$Sim(s_1, s_2) = \frac{\mu_{w_1,2} + \mu_{w_2,2} + \dots + \mu_{w_n,2}}{n} \quad (12)$$

که

$$\mu_{w_i,2} = 1 - \prod_{w_j \in s_2} 1 - (c_{ij}) \quad (13)$$

و n تعداد کلمات موجود در جمله s_1 می‌باشد.

۵- دو جمله همانند تشخیص داده می‌شوند اگر

$$Min(Sim(s_1, s_2), Sim(s_2, s_1)) \geq 0.825 \quad (14)$$

و

$$|Sim(s_1, s_2), Sim(s_2, s_1)| \geq 0.15 \quad (15)$$

که شرط دوم برای کاهش و حذف نتایج مثبت و منفی اشتباه اعمال شده است.

۲-۵-۲- روش ارائه شده توسط الظهرانی^۱ و سلیم^۲

در (Alzahrani and Salim, 2009 a)، نویسندگان بهبودی از روش دوم ارائه شده در (Yerra and Ng, 2005) ارائه داده‌اند. در روش بهبود یافته، با استفاده از یک لغت نامه، کلمات مترادف یک لغت به دست آمده و در محاسبه میزان همانندی دو لغت مورد استفاده قرار می‌گیرند. نحوه استفاده از لغات مترادف به این صورت است که در صورتی که دو واژه با یکدیگر مترادف باشند، همانندی آن‌ها برابر با ۱ (یعنی $C_{ij} = 1$) و در غیر این صورت، همانند روش دوم ارائه شده در (Yerra and Ng, 2005) و با استفاده از فراوانی وقوع همزمان و به تنهایی لغات محاسبه می‌شود. ادامه روال همانندجویی برای تشخیص جملات همانند مانند روش دوم ارائه شده در (Yerra and Ng, 2005) می‌باشد.

۳-۵-۲- روش ارائه شده توسط الظهرانی و سلیم

در (Alzahrani and Salim, 2009 b)، هر دو روش ارائه شده در (Yerra and Ng, 2005) برای زبان عربی استفاده شده و نتیجه‌گیری شده که روش همانندجویی فازی بهتر عمل می‌کند.

۴-۵-۲- روش ارائه شده توسط کوبراشتاین^۳ و ان جی

در روش ارائه شده در (Koberstein and Ng, 2006)، نویسندگان روش ارائه شده در (Yerra and Ng, 2005) را بهبود داده‌اند. اساس کار این روش بر این فرض استوار است که پارامتر C_{ij} به دست آمده تمام اطلاعاتی نیست که می‌توان با تحلیل آماری یک پایگاه داده از متون به دست آورد. علاوه بر معیار C_{ij} ، در این مقاله نویسندگان دو بهبود بر روی این معیار معرفی کرده و از آن‌ها در همانندجویی استفاده کرده‌اند. بهبود اول در نظر گرفته شده بر این باور استوار است که علاوه بر

¹ Alzahrani

² Salim

³ Koberstein

حضور یک لغت در یک متن، تعداد دفعات حضور آن لغت در متن نیز مهم است. در این راستا، نویسندگان معیار همانندی دو لغت (C_{ij}) را به صورت زیر تغییر داده‌اند:

$$C_{ij} = \frac{P_{ij}}{P_{ii} + P_{jj} - P_{ij}} \quad (۱۶)$$

که

$$P_{ij} = \sum_{d \in D} F_{i,d} \times F_{j,d} \quad (۱۷)$$

که D پایگاه داده شامل متون و $F_{i,d}$ فراوانی حضور لغت w_i در متن d می‌باشد.

بهبود دوم تعریف شده در این مقاله، بیان‌گر این است که علاوه بر تعداد حضور، فاصله مابین وقوع دو کلمه نیز در همانندی تاثیرگذار است. بر این اساس، نویسندگان با محاسبه مقادیر P_{ij} با استفاده از رابطه زیر، معیار C_{ij} را بهبود داده‌اند:

$$P_{ij} = \sum_{k_i \in D} \sum_{k_j \in D} \frac{1}{r(k_i, k_j)} \quad (۱۸)$$

که D مجموعه تمام متون، k_i و k_j به ترتیب یک وقوع از لغات w_i و w_j در یک متن از D و $r(k_i, k_j)$ یکی بیشتر از تعداد لغات مابین k_i و k_j می‌باشد.

نویسندگان بیان کرده‌اند که بهبود دوم در نظر گرفته شده جواب‌های دقیق‌تری نسبت به اصل معیار و بهبود اول آن ارائه داده اما استفاده از آن از کارایی روش می‌کاهد. همچنین، استفاده از بهبود اول ارائه شده منجر به ارائه نتایجی دقیق‌تر اما کارایی کمتر نسبت به زمان استفاده از اصل معیار ارائه شده در روش دوم ارائه شده در (Yerra and Ng, 2005) می‌شود.

همچنین، لازم به ذکر است که در زمان استفاده از روش‌های این چنینی که مبتنی بر اطلاعات آماری می‌باشند، باید اطلاعات آماری از مجموعه متون بسیار بزرگ و یکنواخت محاسبه شده باشند و برای مثال پارامترهای به دست آمده از مجموعه متون کوچک یا مجموعه متون مربوط به زمینه‌ای خاص برای استفاده در این روش‌ها قابل قبول نیست.

۲-۵-۵- روش ارائه شده توسط آنگر^۱ و همکاران

در (Unger et al., 2016)، همانند (Murugesan et al., 2010) همانندجویی در متون محرمانه مطرح شده است. در این راستا، نویسندگان از یکی از الگوریتم‌های بازیابی منبع موجود استفاده کرده و به گونه‌ای آن را تغییر داده‌اند که با افشا کمترین اطلاعات ممکن می‌تواند همانندی دو متن را بررسی کند. در مقایسه با (Murugesan et al., 2010)، در این روش، کلمات مشترک بین دو متن و فراوانی آن‌ها برای طرفین مشخص می‌شود.

۲-۵-۶- روش ارائه شده توسط سانچز-پرز^۲

در (Sanchez-Perez et al., 2014)، در راستای دستیابی به نتایج دقیق‌تر، روش رتبه‌بندی tf -idf تغییر یافته و از رتبه‌بندی جدیدی برای استخراج کلمات کلیدی استفاده شده است. روش رتبه‌بندی جدید tf -isf نام‌گذاری شده و به صورت زیر انجام می‌شود:

در روش tf -idf اهمیت یک کلمه از یک متن در آن متن و با توجه به مجموعه‌ای از متون منبع مورد بررسی قرار می‌گرفت اما در گونه تغییر یافته اهمیت یک کلمه در یک جمله و با توجه به یک متن مورد بررسی قرار می‌گیرد. فرض کنیم $tf(t, s)$ بیانگر تعداد تکرار کلمه یا عبارت t در جمله s و

$$idf(t, D) = \log \frac{N}{|\{s \in D \mid tf(t, s) \neq 0\}|}$$

یکی بیشتر از تعداد جملات شامل عبارت یا کلمه t در متن D باشد. در این صورت، پارامتر tf -isf متناظر با یک کلمه یا عبارت t و یک جمله s در یک متن منبع D برابر با $tf(t, s) \times isf(t, D)$ در نظر گرفته می‌شود. پس از بازیابی منابع احتمالی با استفاده از این پارامتر و روش استفاده شده در (Kong et al., 2013) و (Kong et al., 2015)، در ادامه از روش‌های همانندجویی برداری و معیارهای کوساین و دایس برای مقایسه متون منبع احتمالی و متن مشکوک استفاده شده است. به صورت دقیق‌تر، در این روش:

۱- پیش‌پردازش‌های لازم شامل ریشه‌یابی لغوی، تغییر حروف بزرگ به حروف کوچک و ترکیب جملات کوچک با جمله بعدی برای داشتن جملاتی با حداقل ۴ کلمه انجام می‌گیرد.

¹ Unger

² Sanchez-Perez

۲- به ازای هر مدرک منبع احتمالی، مدرک D شامل تمام جملات این منبع و جملات متن مشکوک ساخته می شود و

a. همانندی هر جمله S_1 از مدرک مشکوک و هر جمله S_2 از مدرک منبع به صورت زیر مورد بررسی قرار می گیرد:

i. پارامتر $tf-isf$ را به ازای هر کلمه از هر یک از دو جمله محاسبه می شود.

ii. بردار حاصل از ترکیب دو جمله محاسبه می شود.

iii. جملات S_1 و S_2 با توجه به مقدار $tf-isf$ محاسبه شده برای لغات آن ها و مجموعه لغات موجود در اجتماع آن ها به صورت برداری نشان داده می شوند.

iv. دو جمله همانند تشخیص داده می شوند اگر پس از محاسبه معیارهای همانندی کوساین و دایس از طریق فرمول های (۳) و (۱۰)، داشته باشیم:

$$\cosine(s_1, s_2) \geq t_1 \text{ و } D i d(s_1, s_2) \geq t_2 \text{ که } t_1 \text{ و } t_2 \text{ دو مقدار آستانه ای می باشند.}$$

b. پس از تشخیص تمام زوج جملات احتمالاً همانند در متن منبع احتمالی و متن مشکوک، جملاتی که در متن منبع احتمالی با متن مشکوک به یکدیگر نزدیک بوده با یکدیگر ترکیب شده، قسمت های تکراری و قسمت های با طول کم حذف شده و قسمت های همانند نهایی مشخص می شوند.

۲-۶- روش های همانندجویی ترکیبی

در برخی از روش های ارائه شده برای همانندجویی نویسندگان از ترکیبی از سایر روش ها استفاده کرده اند. از جمله این ترکیب ها می توان به ترکیب روش های مبتنی بر بردار و روش های معنایی و همچنین روش های مبتنی بر گرامر و روش های معنایی اشاره کرد. ترکیب روش های مختلف از سایر دسته روش های همانندجویی موجب توانایی بهتر روش های ترکیبی در انواع همانندجویی شده اما بالطبع در نتیجه این ترکیب و نیازمندی به انجام عملیات بیشتر، از کارایی روش کاسته شده است. در ادامه، برخی از روش های ارائه شده در این دسته مورد بازبینی قرار می گیرند.

۲-۶-۱- روش ارائه شده توسط آدام^۱

در (ADAM, 2014) یک روش همانندجویی ترکیبی ارائه شده است که از مشخص کننده اقسام کلمه و لغت نامه برای همانندجویی استفاده شده است. در این روش:

۱- پیش پردازش های رایج انجام شده و مدارک به واحدهایی برابر با جملات تقسیم بندی می شوند.

۲- همانندی هر جمله از مدرک مشکوک و هر جمله از مدرک منبع به صورت زیر مشخص می شود:

a. با استفاده از مشخص کننده اقسام کلمه، نوع کلمات به کار رفته در جملات مشخص می شود.

b. با استفاده از از وردنت، کلمات مترادف هم نوع با هر کلمه به دست آمده و مجموعه لغات هر جمله به دست می آید.

c. با استفاده از معیار جاکارد (بیان شده در مرور روش ارائه شده توسط بارون-سدنو و همکاران (Barrón-Cedeño et al., 2010)) میزان همانندجویی دو جمله بررسی و در صورت بیشتر بودن از مقداری آستانه ای به عنوان همانند در نظر گرفته می شوند.

۳- پردازش های پسین مورد نیاز برای ترکیب قسمت های همانند تشخیص داده شده نزدیک به یکدیگر و حذف قسمت های با اندازه کم تشخیص داده شده اعمال می شود.

۲-۶-۲- روش ارائه شده توسط الزهرانی و همکاران

در (Alzahrani et al., 2015)، از تلفیقی از روش های همانندجویی معنایی و فازی و همچنین استفاده از ویژگی های گرامری متون، روشی جدید برای همانندجویی ارائه شده است. روش فازی استفاده شده در این روش، روش ارائه شده در (Yerra and Ng, 2005) بوده و نویسندگان با

¹ Adam

تلفیق آن با روش‌های همانندجویی معنایی و همچنین استفاده از تشخیص‌دهنده‌های اقسام کلمه سعی در بهبود نتایج نموده‌اند. جزئیات این روش به شرح زیر است:

۱- ابتدا پیش‌پردازشی شامل استفاده از یک روش استخراج ویژگی، برای مشخص کردن واژگان متن و همچنین برچسب‌زنی اقسام کلمه انجام می‌شود.

۲- متون به واحدهایی مناسب تقسیم‌بندی می‌شوند. در این مقاله، تقسیم‌بندی‌های زیر پیشنهاد شده است:

- دنباله‌های ۳-تایی از لغات،

- دنباله‌های ۵-تایی از لغات،

- دنباله‌های ۸-تایی از لغات با ۳ لغت هم‌پوشان،

- جملات.

۳- اقسام کلمه در نظر گرفته شده با یکدیگر ترکیب شده تا هر لغت در یکی از ۴ نوع اسم، فعل، صفت و قید قرار گیرد. انجام این مرحله احتمالاً با هدف افزایش کارایی روش انجام گرفته است. زیرا هر چه تعداد انواع در نظر گرفته شده بیشتر باشد، روش دقیق‌تر اما ناکاراتر می‌شود.

۴- همانندی هر واحد از مدرک مشکوک با هر واحد از هر مدرک منبع به صورت زیر بررسی می‌شود:

- همانندی معنایی هر واژه از واحد اول و هر واژه از واحد دوم (برای انواع

تقسیم‌بندی‌های مختلف پیشنهاد شده در مرحله ۲) با استفاده از وردنت،

برچسب‌های اقسام کلمه محاسبه شده در مرحله قبل و معیار همانندی ارائه شده در

(Wu and Palmer, 1994) (برحسب کوتاه‌ترین مسیر دو واژه و همچنین

عمقی که آن‌ها در آن قرار گرفته‌اند) محاسبه می‌شود.

- با توجه به مقادیر محاسبه شده در مرحله قبل، همانندی معنایی هر لغت از واحد اول

با واحد دوم و بالعکس محاسبه می‌شود.

- با توجه به مقادیر محاسبه شده در مرحله قبل، همانندی معنایی هر دو واحد محاسبه می‌شود.

- با توجه به مقدار به دست آمده در مرحله قبل و در صورت بزرگ‌تر بودن آن از مقداری آستانه‌ای، این دو واحد همانند تشخیص داده می‌شوند.

۵- پس از محاسبه تمام واحدهای همانند، با انجام پردازشی پسین، واحدهای همانند تشخیص داده شده بررسی شده و در صورت نزدیک بودن به هم، برخی از آن‌ها با یکدیگر ترکیب و در صورت کوتاه‌تر بودن از مقداری، برخی دیگر حذف شده و نتیجه نهایی برای ارائه به کاربر آماده می‌شود.

۲-۶-۳- روش ارائه شده توسط گیپ و ماسچه

در (Meuschke and Gipp, 2014)، روشی جدید برای همانندجویی ارائه شده که ترکیبی از روش ارائه شده در (Gipp and Meuschke, 2011)، روش‌های مبتنی بر کاراکتر و روش‌های مبتنی بر معنا برای همانندجویی استفاده می‌کند. در این مقاله، پیشنهاد شده است که با توجه به کارایی روش ارائه شده در (Gipp and Meuschke, 2011) و روش‌های مبتنی بر کاراکتر، از این دو روش در فاز اول همانندجویی یعنی تشخیص مدارک منبع احتمالی استفاده شود و از روش‌های همانندجویی معنایی که دقت بیشتر اما کارایی کمتری دارند، در کنار الگوریتم‌های تطبیق رشته و روش مبتنی بر ارجاع (روش ارائه شده در (Gipp and Meuschke, 2011)) برای فاز دوم یعنی "تطبیق متن" استفاده شود.

طراحی یک الگوریتم همانندجو برای تشخیص متون بازنویسی شده در زبان فارسی □ ۴۴

**فصل سوم: ابزارهای پردازش زبان‌های طبیعی به کار رفته
برای همانندجویی: بررسی نمونه‌های ارائه شده از این ابزارها
برای زبان فارسی**

در این فصل، ابزارها و مفاهیمی از پردازش زبان‌های طبیعی که در همانندجویی در سایر زبان‌ها به کار رفته را ذکر کرده و مروری بر کارهای انجام گرفته در این زمینه‌ها برای پردازش زبان فارسی انجام می‌دهیم.

بر اساس آنچه در فصل دوم این طرح بیان شد، مفاهیم زیر از پردازش زبان‌های طبیعی برای همانندجویی در سایر زبان‌ها مورد استفاده قرار گرفته‌اند:

۱- ریشه‌یابی لغوی

۲- برچسب‌گذاری اجزای واژگانی کلام

۳- شبکه واژگانی

۴- برچسب‌زنی نقش معنایی

۵- ایست‌واژه‌ها

در ادامه این گزارش، هر یک از این مفاهیم را مورد بررسی قرار داده و الگوریتم‌های ارائه شده برای آن‌ها در زبان فارسی را بیان می‌کنیم.

۳-۱- ریشه‌یابی لغوی

در زبان فارسی یک لغت از ریشه، پیشوندها و پسوندها تشکیل شده است که ریشه بخش اصلی می‌باشد. ریشه‌یابی عبارت است از فرایند کاهش یک لغت به ریشه آن. به عبارت دیگر، یک ریشه‌یاب یک نرم‌افزار است که همه فرم‌های مختلف یک لغت را به یک ریشه یکتا باز می‌گرداند. ریشه‌یابی در بیشتر فعالیت‌های مربوط به پردازش زبان طبیعی مانند بازیابی اطلاعات، جستجوی وب، همانندجویی و ... کاربرد داشته و استفاده از آن موجب بهبود نتایج شده و کارایی سیستم را ارتقا می‌دهد. برای مثال، در بازیابی اطلاعات، می‌توان از ریشه یک لغت به جای خود لغت برای بازیابی استفاده کرد و علاوه بر بهبود نتایج سیستم، فضای کمتری را نیز برای ذخیره‌سازی فایل‌های اندیس‌گذاری مصرف کرد (Karimpour et al., 2009). بررسی‌های انجام شده بیان‌گر دقت

خوب روش‌های ارائه شده برای ریشه‌یابی در زبان فارسی می‌باشد. همانند سایر زبان‌ها، روش‌های ارائه شده برای ریشه‌یابی در زبان فارسی را می‌توان به ۳ دسته تقسیم نمود:

۱- روش‌های ساختاری^۱،

۲- روش‌های مبتنی بر جدول جستجو^۲،

۳- روش‌های آماری^۳.

در ادامه، ویژگی‌های این دسته‌های مختلف را معرفی و با یکدیگر مقایسه کرده و مهم‌ترین الگوریتم‌های موجود در هر دسته را بیان می‌کنیم.

۳-۱-۱- روش‌های ساختاری

الگوریتم‌های موجود در این دسته مبتنی بر ساختار مورفولوژی زبان هستند و برای یافتن ریشه یک لغت، پیشوندها و پسوندهای آن را (در صورت وجود) حذف می‌کنند. روش‌های موجود در این دسته از دقت تقریباً بالایی برخوردار بوده و حافظه کمی برای ذخیره‌سازی نیاز دارند. در (Tashakori et al., 2002)، تشکری^۴ و همکاران اولین الگوریتم ریشه‌یاب برای زبان فارسی را ارائه کرده‌اند. این الگوریتم از قواعد ساخت لغت در زبان فارسی برای شناسایی پیشوندها و پسوندها در یک الگوریتم تکرار شونده استفاده می‌کند تا بتواند ریشه یک لغت را به دست آورد. علاوه بر این، این الگوریتم به لغت‌نامه‌هایی برای شناسایی جمع‌های مکسر و مصادر در زبان فارسی مجهز می‌باشد. در (Taghva et al., 2005) تقوا^۵ و همکاران الگوریتم ساختاری دیگری برای ریشه‌یابی لغات در زبان فارسی ارائه کرده‌اند که علاوه بر قواعد ساختاری، نوع کلمات را نیز در ریشه‌یابی در نظر می‌گیرد. به عبارت دیگر، برای جلوگیری از اشتباه حذف شدن کاراکتر از لغات (و در نتیجه به دست آوردن ریشه اشتباه)، این الگوریتم با تشخیص اولین دنباله پیشوندی یا پسوندی در یک لغت نوع آن را به دست می‌آورد و در ادامه تنها در صورت وجود پیشوند یا پسوند مرتبط با آن نوع لغت اقدام به حذف آن

¹ Structural methods

² Lookup table based methods

³ Statistical methods

⁴ Tashakori

⁵ Taghva

می‌کند. در (Mokhtaripour and Jahanpour, 2006)، مختاری‌پور^۱ و جهان‌پور^۲ علاوه بر لحاظ کردن لغات خارجی، قواعد خاص برای اضافه کردن پیشوند یا پسوند به لغات خاص را در نظر گرفته و به این نکته توجه نمودند که طول ریشه پس از حذف یک رشته از حروف به عنوان پیشوند یا پسوند، نباید کمتر از ۲ حرف باشد. نویسندگان به این ترتیب توانسته‌اند نتایج بهتری برای ریشه‌یابی به دست آورده‌اند. در (Sharifloo and Shamsfard, 2008)، شریف‌لو^۳ و شمس‌فرد^۴ رویکردی متفاوت نسبت به روش‌های قبل برای ریشه‌یابی در زبان فارسی ارائه کرده‌اند. در رویکرد جدید، به جای حذف پیشوندها و پسوندهای ممکن، در ابتدا تمام ریشه‌های ممکن برای یک لغت به دست خواهد آمد. در ادامه با استفاده از قواعد ساخت لغت جدید در زبان فارسی، سعی می‌شود که از هر یک از ریشه‌های موجود در مجموعه ریشه‌های ممکن به دست آمده، لغت ابتدایی و ورودی ساخته شود. انجام این فرایند موجب نمایان شدن این مهم می‌شود که از برخی از ریشه‌های ممکن موجود در لیست نمی‌توان لغت موردنظر را به دست آورد. در نتیجه این لغات از لیست حذف می‌شوند. در ادامه، از استثنائاتی که در قواعد ساخت لغت در زبان فارسی وجود دارد استفاده شده و برخی دیگر از ریشه‌های ناممکن حذف می‌شوند و تنها ریشه‌های ممکن درست باقی می‌مانند. در (Jadidinejad, 2010)، علاوه بر در نظر گرفتن، قواعد ساخت لغات برای حذف پیشوندها و پسوندها، از یک تابع درهم‌ساز برای ذخیره و جستجوی بهینه در میان ریشه‌های ممکن استفاده شده است.

۳-۱-۲- روش‌های مبتنی بر جدول جستجو

الگوریتم‌های موجود در این دسته، همان‌طور که از اسم آن‌ها نیز مشخص است، از یک جدول جستجو شامل لغات و ریشه‌های آن‌ها برای ریشه‌یابی استفاده می‌کنند. این دسته از روش‌ها دارای دقت بالایی بوده اما نیازمند ذخیره‌سازی جدول جستجویی با بزرگی دو برابر لغات موجود در یک لغت‌نامه می‌باشند. در نتیجه کاربرد آن‌ها در برخی موقعیت‌ها ناممکن می‌باشد. از دیگر نقاط ضعف این روش‌ها می‌توان به ناتوانی آن‌ها در برخورد با لغات جدید اشاره کرد، که در این صورت نیازمند به‌روزرسانی خواهند بود. در (Hesamifard and Ghassem-Sani, 2006)، حسامی‌فرد^۵ و

¹ Mokhtaripour

² Jahanpour

³ Sharifloo

⁴ Shamsfard

⁵ Hesamifard

قاسم‌ثانی^۱ یک الگوریتم مبتنی بر جدول جستجو برای زبان فارسی با تمرکز بر روی افعال ارائه کرده‌اند. آزمایش‌های انجام شده بر روی این الگوریتم بیان‌گر نتایج خوب آن می‌باشد. در (Rahimi, 2015)، در راستای دستیابی به نتایج بهتر، روشی ترکیبی از روش‌های ساختاری و روش‌های مبتنی بر جدول جستجو ارائه شده است و از آن برای بازیابی اطلاعات استفاده شده است. لازم به ذکر است که هیچ یک از این جداول جستجو کامل نبوده و همان‌طور که در (Fadaei and Shamsfard, 2010) بیان شده، ساختن جدول جستجوی کامل برای ریشه‌یابی در زبان فارسی کاری سخت است و در حال حاضر چنین جدولی در دسترس نیست.

۳-۱-۳- روش‌های آماری

الگوریتم‌های موجود در این دسته از اطلاعات آماری موجود در یک پیکره بزرگ از مدارک برای به دست آوردن قوانینی برای ریشه‌یابی استفاده می‌کنند. مزیت روش‌های موجود در این دسته عدم وابستگی آن‌ها به زبان می‌باشد. اما این دسته از روش‌ها به طور متوسط نسبت به روش‌های موجود در دو دسته دیگر از دقت پایین‌تری برخوردار هستند. در (Nasiri et al., 2006)، نصیری^۲ و همکاران یک الگوریتم ریشه‌یاب آماری ارائه کرده‌اند که در ابتدا با استفاده از یک پیکره، لیست پیشوندها و پسوندهای زبان فارسی را استخراج می‌کند. در ادامه، هنگام ریشه‌یابی یک لغت، این الگوریتم پیشوند و پسوندهای ممکن را شناسایی کرده و بهترین ریشه را به دست می‌آورد. در (Momenipour and Keyvanpour, 2016)، نویسندگان یک روش مبتنی بر "مدل مخفی مارکوف"^۳ برای ریشه‌یابی در زبان فارسی ارائه کرده‌اند. در این روش، برای ریشه‌یابی یک لغت، در ابتدا به جستجوی آن در یک لغت‌نامه می‌گردد که شامل واژگانی است که ریشه آن‌ها طبق قواعد طبیعی زبان فارسی به دست نمی‌آید (مانند جمع‌های مکسر) و در صورت وجود، ریشه مورد نظر بازگردانده می‌شود. در غیر این صورت، با استفاده از قواعد استنتاج شده با استفاده از مدل مخفی مارکوف، ریشه لغت ورودی محاسبه و بازگردانده می‌شود.

¹ Ghassem-Sani

² Nasiri

³ Hidden Markov model

۲-۳- برچسب‌گذاری اجزای واژگانی کلام

برچسب‌گذاری اجزای واژگانی کلام عمل انتساب برچسب‌های واژگانی مناسب به کلمات و نشانه‌های تشکیل دهنده یک متن است به صورتی که این برچسب‌ها نشان‌دهنده نقش کلمات و نشانه‌ها در جمله باشد. درصد بالایی از کلمات از نقطه نظر برچسب واژگانی دارای ابهام هستند، زیرا کلمات در جایگاه‌های مختلف برچسب‌های واژگانی متفاوتی دارند. به عبارت دیگر، برچسب‌گذاری واژگانی عمل ابهام‌زدایی از برچسب‌ها با توجه به زمینه (متن) موردنظر است. برچسب‌گذاری واژگانی عملی اساسی برای بسیاری از حوزه‌های دیگر پردازش زبان طبیعی از قبیل "ترجمه ماشینی"^۱، خطایابی^۲ و "تبدیل متن به گفتار"^۳ می‌باشد. این روش‌ها از اطلاعات زمینه برای انتخاب برچسب درست یک کلمه در یک جمله استفاده می‌کنند. علاوه بر این، برای برچسب‌گذاری لغاتی که حتی در واژه‌نامه‌ها نیز وجود ندارند راهکارهایی توسط این روش‌ها ارائه شده که از آن جمله می‌توان به اختصاص برچسب لغات کم‌یاب و یا اختصاص تمام برچسب‌های موجود به لغت و در ادامه استفاده از روش‌های قبلی برای تعیین برچسب درست اشاره کرد.

روش‌های ارائه شده برای برچسب‌گذاری اجزای واژگانی کلام را می‌توان به ۲ دسته روش‌های آماری و روش‌های مبتنی بر قاعده^۴ تقسیم‌بندی کرد (Fadaei and Shamsfard, 2010). روش‌های برچسب‌گذاری مبتنی بر قاعده از دو فاز تشکیل می‌شوند. در فاز اول، با استفاده از یک لغت‌نامه، تمام برچسب‌های ممکن که یک لغت می‌تواند داشته باشد استخراج شده و به آن لغت اختصاص داده می‌شود. در فاز دوم، با استفاده از قواعدی محدودکننده (مبتنی بر نحو) تلاش به ابهام‌زدایی می‌شود. روش‌های برچسب‌گذاری آماری روش‌هایی احتمالاتی بوده و با استفاده از اطلاعات آماری، سعی به تخصیص برچسبی به یک لغت می‌کنند که بیش‌ترین احتمال را با توجه به دنباله لغاتی (برچسب‌هایی) که در آن ظاهر می‌شوند، دارد. بر اساس مطالعات صورت گرفته، به نظر می‌رسد که اغلب کارهای انجام شده برای برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی با تکیه بر روش‌های آماری انجام شده است. در این راستا، عاصی^۵ و حاجی‌ابوالحسنی^۱ در ۲۰۰۰ یک

¹ Machine Translation

² Error detection

³ Text to speech conversion

⁴ Rule based methods

⁵ Assi

روش آماری برای برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی ارائه کرده‌اند (Assi, 2000). در این کار از یک روش توزیع‌شده استفاده شده و یک لغت با در نظر گرفتن لغات مجاور برچسب‌گذاری می‌شود. به دلیل در نظر نگرفتن یک سیاست برای ابهام‌زدایی، این روش در مورد لغات کم کاربرد بد عمل می‌کند. علاوه بر این، ویژگی "فاقد ترتیب بودن استفاده از لغات"^۲ در زبان فارسی موجب کاهش دقت این روش شده است. روش‌های ارائه شده در (Raja et al., 2007) و (Azimizadeh et al., 2008)، از دنباله‌های سه کلمه‌ای برای ابهام‌زدایی استفاده می‌کنند. به عبارت دیگر، این برچسب‌گذارها، برچسب یک لغت را با در نظر گرفتن برچسب لغات همسایه آن انتخاب می‌کنند. روش برچسب‌گذار ارائه شده توسط محترمی^۳ در (Mohtarami et al., 2006) نیز از روش‌های آماری می‌باشد که صحت نتایج این روش نیز همانند روش‌های قبل در مورد لغات کم کاربرد پایین می‌باشد. در (Mosavi Miangah and Delavar Khalafi, 2012)، موسوی میانگاه^۴ و دلاورخلفی^۵ یک روش دیگر برای برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی ارائه کرده‌اند. این روش نیازمند پیکره‌ای بزرگ می‌باشد اما، برخلاف سایر روش‌های مطرح شده تا به اینجا که نیازمند به پیکره‌ای برچسب‌گذاری شده بودند، این روش دارای چنین نیازمندی نیست. این روش، در ابتدا به هر لغت مجموعه برچسب‌های ممکن را اختصاص می‌دهد. در ادامه با استفاده از فراوانی دنباله‌های n -تایی از برچسب‌ها سعی به ابهام‌زدایی و اختصاص یگانه برچسب به هر لغت می‌نماید. در (Fadaei and Shamsfard, 2010)، فدایی^۶ و شمس‌فرد یک روش آماری جدید برای برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی ارائه کرده‌اند. در این روش، پس از تجزیه متن به جملات، از واژه‌نامه‌ای شامل لغاتی که تمام برچسب‌های ممکن متناظر با لغات آن را می‌دانیم، استفاده می‌شود. برای برچسب‌گذاری لغات جدید، این روش از قواعد ریختی^۷ زبان فارسی استفاده می‌کند. پس از الصاق تمام برچسب‌های ممکن به مجموعه لغات یک جمله، با در نظر گرفتن فراوانی دوتایی‌های ممکن از برچسب‌های پیاپی، روش ارائه شده سعی به ابهام‌زدایی و

¹ Haji Abdolhosseini

² Free word order characteristic

³ Mohtarami

⁴ Mosavi Miangah

⁵ Delavar Khalafi

⁶ Fadaei

⁷ Morphologically

برچسب‌گذاری یکتای لغات یک جمله می‌نماید. در (Alavi et al., 2014)، علوی^۱ و همکاران یک روش جدید برای برچسب‌گذاری اجزای واژگانی کلام برای زبان فارسی ارائه کرده‌اند. در این روش، در ابتدا با اصلاح فاصله‌ها و نیم‌فاصله‌ها و تشخیص وندها از ریشه‌های هم‌نگاره^۲ موجود در زبان فارسی متن ورودی نرمال‌سازی می‌شود. در ادامه، با توجه به علائم نقطه‌گذاری به کار رفته در متن، متن به جملات تقسیم‌بندی می‌شود. در نهایت، برای هر جمله، با استفاده از مدل مخفی مارکوف، از بین مجموعه برچسب‌های ممکن، محتمل‌ترین برچسب به هر کلمه از این جمله اختصاص داده می‌شود. در (Rezai and Mosavi Miangah, 2016)، نویسندگان اقدامات انجام شده برای طراحی FarsiTag که یک سیستم مشخص‌کننده اجزای واژگانی کلام برای زبان فارسی است را بیان کرده‌اند. در FarsiTag، از قواعد زبانی برای اختصاص مناسب‌ترین برچسب به یک لغت استفاده شده است. با انجام این کار نویسندگان توانسته‌اند بهترین نتایج موجود را برای برچسب‌گذاری اجزای واژگانی کلام در فارسی را ارائه کنند. لازم به ذکر است که بررسی‌های انجام شده بیان‌گر کیفیت بالا و دقت خوب برچسب‌گذارهای اجزای واژگانی کلام ارائه شده برای زبان فارسی در مقایسه با سایر زبان‌ها می‌باشد.

۳-۳- شبکه واژگانی

شبکه واژگانی یک پایگاه داده لغوی الکترونیکی است که در ابتدا برای زبان انگلیسی و در ادامه برای برخی دیگر از زبان‌ها طراحی شده است. شبکه واژگانی، لغات را به مجموعه‌هایی به لحاظ مفهومی هم‌معنا تقسیم‌بندی می‌کند که به هر مجموعه یک "دسته هم‌معنا" می‌گویند. یک دسته هم‌معنا یک مجموعه از لغات با برچسب اجزای واژگانی کلام یکسان هستند که می‌توانند در یک متن مشخص به جای یکدیگر به کار برده شوند. در حقیقت، هر دسته هم‌معنا نمایشگر یک مفهوم مجزا می‌باشد که می‌تواند توسط اعضای خود در متون متفاوت بیان شود. دسته‌های هم‌معنا با استفاده از روابط لغوی (لغت به لغت) و مفهومی-معنایی (دسته هم‌معنا به دسته هم‌معنا) به یکدیگر وابسته‌اند. این روابط ممکن است لغات با یک نوع برچسب اجزای واژگانی کلام (مانند مترادف، متضاد، شمول معنایی و

¹ Alavi

² Homograph

جزء واژگی) و یا دسته‌های متفاوت را به یکدیگر مرتبط کنند. در (Shamsfard et al., 2010)، شمس‌فرد و همکاران اولین شبکه واژگانی کاربردی را برای زبان فارسی طراحی کرده‌اند. روش ارائه شده توسط نویسندگان، که FarsNet نامیده شده است، یک روش نیمه‌خودکار برای ساخت شبکه واژگانی است. نمونه ثانویه این شبکه واژگانی شامل ۱۸۰۰۰ لغت می‌باشد که به شکل ۶۵۰۰ دسته هم‌معنا از اسامی، صفات و افعال سازماندهی شده است. در (Taghizadeh and Faili, 2016)، تقی‌زاده^۱ و فیلی^۲ یک روش تمام خودکار برای ساخت شبکه واژگانی برای زبان‌های با منابع کم ارائه کرده‌اند که با استفاده از یک واژه‌نامه دو زبانه و یک پیکره تک‌زبانه و یک الگوریتم امید ریاضی-بیشینه‌سازی شبکه واژگانی را ایجاد می‌کند. نویسندگان از روش ارائه شده برای تهیه یک شبکه واژگانی برای زبان فارسی استفاده کرده‌اند و به مقایسه نتایج آن با FarsNet پرداخته‌اند. با توجه به کوچک بودن اندازه پیکره مورد استفاده، این شبکه واژگانی لغات و دسته‌های هم‌معنای کمتری را در مقایسه با FarsNet مورد پوشش قرار می‌دهد اما به طور دقیق‌تر دسته‌های هم‌معنا را مشخص می‌کند.

۳-۴- برچسب‌زنی نقش معنایی

درک یک رویداد و عناصر دخیل در آن قسمتی مهم از درک یک زبان طبیعی می‌باشد. به طور کلی، درک یک رویداد بدین معنی است که بتوانیم به سوال "چه کسی چه چیزی را روی چه کس دیگری انجام داد (و شاید کی و چگونه)" پاسخ دهیم (کامل قالی باف و همکاران، ۱۳۸۸). در یک جمله ممکن است پاسخ به این پرسش به راه‌های مختلف بیان شود. برای مثال، اگر بخواهیم با هدف پاسخ دادن به سوالی درباره خرید سهام توسط شرکت X جملاتی را پردازش کنیم، نیازمند درک این رویداد هستیم در حالیکه این رویداد ممکن است به طرق مختلف بیان شده باشد. برای مثال، همانطور که در جملات زیر نشان داده شده است، این رویداد می‌تواند توسط فعل‌های مختلفی مانند خریدن یا فروختن و یا توسط یک اسم مانند خریدار توصیف شده باشد. شرکت X می‌تواند فاعل غیر مستقیم یا مفعول نحوی جمله باشد:

¹ Taghizadeh

² Faili

- شرکت X سهام را خرید.
- آنها سهام را به شرکت X فروختند.
- سهام توسط شرکت X خریداری شد.
- خریدار سهام شرکت X است.
- خرید سهام توسط شرکت X ...

هدف در برجسب زنی نقش معنایی تشخیص اشتراک بین گونه‌های مختلف بیان یک رویداد می‌باشد. در مورد مثال ذکر شده، برجسب زن نقش معنایی باید بتواند تشخیص دهد که یک رویداد خرید اتفاق افتاده است و سهام و شرکت X در این رویداد مشارکت داشته و شرکت X نقش به دست آورنده سهام را داشته است. به بیان دقیق‌تر، برجسب‌زنی نقش معنایی یا تجزیه سطحی معنایی را می‌توان تحلیل معنایی متن در سطح جمله دانست، که در آن فعل جمله مشخص‌کننده رویداد واقع شده و سایر اجزای جمله هر یک نقشی در ارتباط با این رویداد می‌پذیرند. روش‌های موجود برای تعیین نقش‌های معنایی را می‌توان به دو دسته روش‌های مبتنی بر قاعده و روش‌های مبتنی بر پیکره تقسیم‌بندی کرد. در روش‌های مبتنی بر قاعده، تحلیل معنایی متن به کمک لغت‌نامه‌ها، گرامرها و سایر منابع معنایی انجام می‌گیرد. این منابع عمدتاً به صورت دستی تهیه می‌شوند. به کارگیری این رویکرد مستلزم ایجاد گرامرهای دستی وسیع و فراگیر برای انواع جملات با ساختارهای دستوری متفاوت در زبان است که با توجه به تنوع و پیچیدگی جملات و افعال، معمولاً این روش‌ها بخش محدودی از زبان را مورد توجه قرار می‌دهند و برای کاربردهای خاص همچون تعیین نقش‌های معنایی در سیستم محاوره فرودگاه مناسب‌اند. متأسفانه کارهای انجام گرفته در زمینه برجسب‌زنی نقش معنایی در زبان فارسی محدود بوده و روش‌های محدود موجود دارای نواقصی مانند دقت پایین، پوشش تنها قسمتی خاص از زبان و ... می‌باشند. در ادامه، مروری بر روی این کارها انجام می‌شود. در (صدرموسوی و شمس‌فرد، ۱۳۸۶) و (Shamsfard and Sadr Mousavi, 2008)، یک برجسب‌زننده نقش معنایی برای زبان فارسی ارائه شده که در آن، ابتدا اجزای نحوی جمله مشخص می‌شوند، سپس، با توجه به فعل جمله و قواعد مربوط به افعال در زبان فارسی، دسته معنایی متناظر با جمله مشخص می‌شود. در انتها، برجسب نقش معنایی هر جزء با استفاده از خصوصیات نحوی و واژگانی آن جزء و قواعد مربوط به زبان صورت می‌گیرد. در روش ارائه شده در (Jurafsky and

(Martin, 2010)، برای برجسب‌زنی نقش معنایی یک جمله، ابتدا آرگومان‌های فعل به وسیله یک تجزیه‌گر نحوی سطحی شناسایی می‌شود، سپس در مرحله بعد با توجه به رابطه معنایی که این آرگومان‌ها با فعل جمله دارند، نقش معنایی مناسب به آن‌ها اختصاص داده می‌شود. در پیاده‌سازی هر دو فاز از روش دسته‌بندی یادگیری مبتنی بر حافظه، بر روی یک پیکره به صورت دستی برجسب‌گذاری شده استفاده شده است. متأسفانه پیکره مورد استفاده در این روش و تنوع جملاتی که این روش قادر به برجسب‌زنی نقش معنایی در آن‌ها می‌باشد، بسیار محدود است.

۳-۵- ایست‌واژه‌ها

در اغلب روش‌های همانندجویی بررسی شده در گزارش مرحله قبل این طرح، حذف ایست‌واژه‌ها یکی از پیش‌پردازش‌های مورد نیاز بود. ایست‌واژه‌ها لغاتی هستند که علی‌رغم تکرار فراوان در متن، از لحاظ معنایی دارای اهمیت کمی هستند مثل "اگر"، "و"، "ولی"، "که" و در نگاه اولیه کلمات ربط و تعریف، ایست‌واژه به نظر می‌آیند؛ در عین حال بسیاری از افعال، افعال کمکی، اسم‌ها، قیدها و صفات نیز ایست‌واژه شناخته شده‌اند. در اغلب کاربردهای پردازش متن، حذف این کلمات، نتایج پردازش را به شدت بهبود می‌دهد و سبب کاهش بار محاسبات و افزایش سرعت خواهد شد. به همین دلیل این کلمات غالباً در فاز پیش‌پردازش، حذف می‌شوند.

ایست‌واژه‌های زبان فارسی به دو دسته " ایست‌واژه‌های فعلی"^۱ و " ایست‌واژه‌های غیر فعلی"^۲ تقسیم‌بندی می‌شوند (Taghva et al., 2003) که ایست‌واژه‌های فعلی زبان فارسی در جدول ۱ و ایست‌واژه‌های غیر فعلی آن در جدول ۲ نمایش داده شده‌اند. لازم به ذکر است که در مورد ایست‌واژه‌های فعلی تنها سه شکل مصدر، گذشته و وجه امری این افعال در جدول ۱ گزارش شده است اما، هر صرف دیگری از این افعال نیز ایست‌واژه تلقی می‌شود.

جدول ۱: ایست‌واژه‌های فعلی زبان فارسی

مصدر	زمان گذشته	وجه امری
------	------------	----------

^۱ Verbal stopwords

^۲ Nonverbal stopwords

کن	کرد	کردن
باش	بود	بودن
شو	شد	شدن
دار	داشت	داشتن
خواه	خواست	خواستن
گوی	گفت	گفتن
ده	داد	دادن
گیر	گرفت	گرفتن
آی	آمد	آمدن
توان	توانست	توانستن
یاب	یافت	یافتن
آور	آورد	آوردن

جدول ۲: ایست‌واژه‌های غیر فعلی زبان فارسی

دیگران	هم‌چنان	مدت	چیز	سایر
جا	طی	کل	کنونی	بیرون
مثلا	کامل	کاملا	آن‌که	موارد
واقعی	امور	اکنون	بطور	بخشی
تحت	چگونه	عدم	نوعی	حاضر
وضع	مقابل	کنار	خویش	نگاه
درون	زمانی	بنابراین	تو	خیلی
بزرگ	خودش	جز	این‌جا	مختلف

توسط	نوع	هم‌چنین	آن‌جا	قبل
جناح	این‌ها	طور	شاید	ایشان
جهت	طریق	مانند	پیدا	ممکن
کسانی	جای	کسی	غیر	بی
قابل	درباره	جدید	وقتی	اخیر
چرا	بیش	روی	طرف	جریان
زیر	آن‌چه	البته	فقط	چیزی
چون	برابر	هنوز	بخش	زمینه
بین	بدون	استفاده	همان	نشان
بسیاری	بعد	عمل	روز	اعلام
چند	آنان	بلکه	امروز	تمام
بیشتر	آیا	برخی	علیه	دیگری
ویژه	گذشته	انجام	حتی	داده
راه	سوی	ولی	زمان	حال
تنها	بسیار	یعنی	عنوان	همین
هیچ	پیش	وی	یکی	این‌که
وجود	شما	پس	چنین	میان
مورد	چه	اگر	همه	نه
دیگر	آن‌ها	باید	هر	او
ما	من	تا	نیز	اما
یک	خود	بر	یا	هم
را	این	با	آن	برای
و	در	به	که	از

فصل چهارم: الگوریتم‌های همانندجوی پیشنهادی و ارزیابی آنها

در این فصل، از نتایج حاصل از فصل‌های دوم و سوم استفاده شده و دو الگوریتم همانندجوی جدید برای تشخیص متون فارسی بازنویسی شده ارائه می‌شود. الگوریتم اول ارائه شده در دسته الگوریتم‌های همانندجویی معنایی قرار گرفته و دومین الگوریتم در دسته الگوریتم‌های همانندجویی فازی قرار می‌گیرد. ارزیابی‌های انجام شده نشان‌گر نتایج بهتر الگوریتم اول ارائه شده می‌باشد.

لازم به ذکر است که روش‌های ارائه شده معادل برخی روش‌های همانندجویی موجود برای سایر زبان‌ها هستند. همان‌طور که در اهداف این طرح بیان شده است، الگوریتم‌های موجود برای سایر زبان‌ها قابل استفاده در زبان فارسی نیستند. در نتیجه، الگوریتم‌های پیشنهادی با اعمال تغییراتی در گونه‌های خارجی خود، بومی‌سازی و مناسب برای استفاده در زبان فارسی گشته‌اند.

در ادامه این فصل، ابتدا الگوریتم‌های همانندجوی ارائه شده را بررسی کرده و پس از بیان معیارهای بررسی کیفیت الگوریتم‌های همانندجو، کیفیت دو الگوریتم ارائه شده را بررسی می‌کنیم.

۴-۱- الگوریتم‌های همانندجوی ارائه شده

در این بخش، دو الگوریتم همانندجوی جدید برای تشخیص متون فارسی بازنویسی شده ارائه می‌شود. همان‌طور که بیان شد الگوریتم‌های ارائه شده در دسته الگوریتم‌های همانندجویی معنایی و فازی قرار می‌گیرند. در ادامه، در ابتدا الگوریتم همانندجویی معنایی ارائه شده و سپس الگوریتم همانندجویی فازی ارائه شده برای تشخیص متون بازنویسی شده فارسی را بیان خواهیم کرد. لازم به ذکر است که رویکردهای در نظر گرفته شده در این طرح با توجه به ابزارهای موجود برای پردازش زبان فارسی (نتایج فصل سوم این طرح) انتخاب شده‌اند.

۴-۱-۱- الگوریتم همانندجوی معنایی پیشنهادی

الگوریتم همانندجوی معنایی پیشنهادی همانند سایر الگوریتم‌های موجود در این دسته از ابزار لغت‌نامه جهت تعیین همانندی متون استفاده می‌کند. همان‌طور که پیش‌تر و در فصل اول این طرح پژوهشی بیان شد، روش‌های همانندجویی موجود برای سایر زبان‌ها قابل استفاده در همانندجویی در زبان فارسی نبوده و اغلب روش‌های همانندجویی ارائه شده برای زبان فارسی نیز تنها همانندی از نوع

آشکار را در نظر گرفته و قادر به تشخیص همانندجویی حاصل از بازنویسی متون نیستند. تنها روش‌های همانندجوی موجود که قادر به تشخیص متون فارسی بازنویسی شده هستند نیز در اواسط اجرای این طرح و جهت شرکت در مسابقاتی با این هدف ارائه شده‌اند (Momtaz et al., 2016)، (Minaei and Niknam, 2016)، (Mashahdirajab and Shamsfard, 2016) و (Gharavi et al., 2016). روش همانندجویی معنایی ارائه شده همانند روش مشهدی‌رجب و شمس‌فرد (Mashahdirajab and Shamsfard, 2016) از لغت‌نامه‌ای فارسی به فارسی جهت همانندجویی استفاده می‌کند. در مقایسه با این روش، روش پیشنهادی نیازمند تشخیص نوع همانندی به کار رفته نبوده، از معیار tf-idf برای وزن‌دهی به واژگان استفاده نکرده، از ابتدا همانندی معنایی هر دو جمله از متن مشکوک و متن منبع احتمالی را بررسی کرده و از معیاری جدید برای بررسی دقیق‌تر همانندی دو جمله استفاده می‌کند.

در الگوریتم پیشنهادی برای مقایسه یک متن مشکوک با یک متن منبع احتمالی به صورت زیر عمل می‌شود:

۱- پیش‌پردازش‌های زیر بر روی متن مشکوک و متن منبع احتمالی انجام می‌شود:

a. هر دو متن نرمال‌سازی می‌شوند. در نرمال‌سازی تنها حروف نگهداری شده و مابقی کاراکترها حذف می‌شوند.

b. ایست‌واژه‌ها از هر دو متن حذف می‌شوند. با توجه به فراوانی استفاده از ایست‌واژه‌ها، این پیش‌پردازش از تاثیر منفی ایست‌واژه‌ها در همانندجویی جلوگیری می‌کند.

c. ریشه‌یابی لغوی بر روی هر دو متن اعمال می‌شود.

۲- هر دو متن به جملات تقسیم‌بندی می‌شوند. در پیاده‌سازی روش پیشنهادی، جمله به عنوان واحد مناسب برای تجزیه متن انتخاب و استفاده شده است. دلیل این امر این است که به نظر می‌رسد تقسیم به دنباله‌چندتایی از کاراکترها تنها در جستجوی دقیق کاربرد داشته و در مقایسه با تقسیم‌بندی به جملات، در تقسیم متون به دنباله‌های چندتایی از لغات، همانندی

متون بازنویسی شده قابل بررسی نیست و در صورت درج یا حذف لغات از جملات همانندی ممکن است غیر قابل تشخیص شود.

۳- هر جمله S_1 از متن مشکوک با هر جمله S_2 از متن منبع احتمالی به صورت زیر مقایسه می شود:

a. جمله S_1 به مجموعه لغات تقسیم بندی می شود. فرض کنیم $\{S_1^1, S_1^2, \dots, S_1^n\}$

مجموعه لغات جمله S_1 و $\{S_2^1, S_2^2, \dots, S_2^m\}$ مجموعه لغات جمله S_2 باشند.

b. همانندی هر لغت S_1^i از جمله S_1 هر واژه S_2^j از جمله S_2 که با $Sim_{S_1^i \leftrightarrow S_2^j}$

نمایش داده می شود به صورت زیر محاسبه می شود:

i. در صورتی که دو واژه یکسان باشند، $Sim_{S_1^i \leftrightarrow S_2^j}$ برابر با یک در نظر گرفته می شود.

ii. در غیر این صورت، مقدار $Sim_{S_1^i \leftrightarrow S_2^j}$ با استفاده از لغت نامه مورد استفاده

در الگوریتم محاسبه می شود. در آزمایشات صورت گرفته از لغت نامه ای

معمولی، تنها شامل لغات مترادف و متضاد استفاده شده و در نتیجه، در

صورت وجود یکی از این کلمات مابین مترادف ها و متضادهای دیگری

مقدار $Sim_{S_1^i \leftrightarrow S_2^j}$ برابر با β قرار خواهد گرفت (مقدار β توسط

آزمایش ها محاسبه می شود). لازم به ذکر است که رویه مورد استفاده در

هنگام استفاده از لغت نامه موجب بازگرداندن تنها دو مقدار صفر و β به

عنوان میزان همانندی دو لغت می شود اما در صورت استفاده از شبکه

واژگانی (Shamsfard et al., 2010)، این پارامتر می تواند از تمام

مقادیر ممکن مقدار اختیار کند. به عبارت دیگر، شبکه واژگانی به صورت

دقیق تر وابستگی معنایی دو واژه را بررسی می کند.

c. همانندی هر لغت S_1^i از جمله S_1 با جمله S_2 که با $Sim_{S_1^i \leftrightarrow S_2}$ نمایش داده

می شود برابر با بیشینه $Sim_{S_1^i \leftrightarrow S_2^j}$ ($j = 1, \dots, m$) قرار داده می شود.

d. به همین ترتیب همانندی هر واژه S_2^j از جمله S_2 با جمله S_1 محاسبه شده و با $Sim_{S_2^j \leftrightarrow S_1}$ نمایش داده می شود.

e. میزان همانندی جمله S_1 نسبت به جمله S_2 با استفاده از رابطه زیر محاسبه می شود:

$$Sim_{S_1 \rightarrow S_2} = \frac{\sum_{i=1}^n Sim_{S_1^i \rightarrow S_2}}{n}$$

f. میزان همانندی جمله S_2 نسبت به جمله S_1 با استفاده از فرمول زیر محاسبه می شود:

$$Sim_{S_2 \rightarrow S_1} = \frac{\sum_{j=1}^m Sim_{S_2^j \rightarrow S_1}}{m}$$

g. دوجمله همانند تشخیص داده می شوند اگر:

$$i. Sim_{S_2 \rightarrow S_1} > \alpha_1$$

$$ii. (Sim_{S_2 \rightarrow S_1})^{\frac{1}{\log_2^{n/2}}} > \alpha'_1$$

$$iii. Sim_{S_1 \rightarrow S_2} > \alpha_2 \text{ و}$$

$$iv. (Sim_{S_2 \rightarrow S_1})^{\frac{1}{\log_2^{n/2}}} > \alpha'_2$$

که α_1 ، α'_1 ، α_2 و α'_2 مقادیری آستانه ای بوده که با انجام آزمایشات محاسبه می شوند.

۴- پس از مشخص شدن جملات مشابه دو متن، فرایند زیر بر روی متن مشکوک و متن منبع احتمالی انجام شده تا قسمت های همانند تشخیص داده شده مجزا به یکدیگر متصل شده و قسمت های با طول کم حذف شوند:

a. در صورتی که دو جمله نزدیک به هم در متن مشکوک (در متن منبع احتمالی) همانند تشخیص داده شود و فاصله آنها در متن کمتر از مقداری آستانه ای باشد، قسمت همانند تشخیص داده شده برابر با از ابتدای جمله اول تا انتهای جمله دوم در

نظر گرفته می‌شود. در پیاده‌سازی‌های صورت گرفته، از نصف طول دو جمله همانند تشخیص داده شده به عنوان این مقدار آستانه‌ای استفاده شده است.

b. در صورتی که قسمتی از متن همانند تشخیص داده شده و پس از انجام مرحله قبل طول آن هم‌چنان کمتر از مقداری آستانه‌ای باشد، این قسمت از مجموعه متون همانند تشخیص داده شده حذف می‌گردد. در پیاده‌سازی‌های صورت گرفته از ۴ کلمه به عنوان این مقدار آستانه‌ای استفاده شده است.

ملاحظه: با توجه به بررسی‌های صورت گرفته در فصل دوم این طرح به نظر می‌رسید اختصاص وزن به واژگان و اثردهی وزن واژگان در همانندجویی منجر به تاثیر بیشتر لغات کم‌کاربرد و در نتیجه کیفیت بهتر الگوریتم همانندجویی شود. آزمایشات انجام شده بیان‌گر عکس این موضوع بوده و در نتیجه، در هیچ یک از الگوریتم‌های پیشنهادی از وزن‌دهی به لغات استفاده نشده است.

۴-۱-۲- الگوریتم همانندجوی فازی پیشنهادی

در این قسمت یک الگوریتم فازی برای همانندجویی در متون فارسی بازنویسی شده ارائه می‌کنیم. لازم به ذکر است که الگوریتم ارائه شده، تنها الگوریتم همانندجوی فازی ارائه شده برای زبان فارسی است. همانند الگوریتم ارائه شده در بخش قبل، این الگوریتم نیز ابتدا پیش‌پردازش‌های ریشه‌یابی لغوی، نرمال‌سازی متن و حذف ایست‌واژه‌ها را بر روی هر دو متن ورودی انجام می‌دهد. در ادامه، پس از تقسیم متون ورودی به واحدهای مناسب، همانندی هر دو واحد را با استفاده از فراوانی‌های لغات و فراوانی‌های هم‌رخدادی آن‌ها محاسبه می‌کند. جزئیات دقیق الگوریتم فازی پیشنهادی به شرح زیر است.

در الگوریتم پیشنهادی برای مقایسه یک متن مشکوک با یک متن منبع احتمالی به صورت زیر عمل می‌شود:

۱- پیش‌پردازش‌های زیر بر روی متن مشکوک و متن منبع احتمالی انجام می‌شود:

a. هر دو متن نرمال‌سازی می‌شوند. در نرمال‌سازی تنها حروف نگهداری شده و مابقی کاراکترها حذف می‌شوند.

b. ایست‌واژه‌ها از هر دو متن حذف می‌شوند.

c. ریشه‌یابی لغوی بر روی هر دو متن ورودی اعمال می‌شود.

۲- هر دو متن به جملات تقسیم‌بندی شده و هر جمله S_1 از متن مشکوک با هر جمله S_2 از متن منبع احتمالی به صورت زیر مقایسه می‌شود:

a. جمله S_1 به مجموعه لغات تقسیم‌بندی می‌شود. فرض کنیم $\{S_1^1, S_1^2, \dots, S_1^n\}$

مجموعه لغات جمله S_1 و $\{S_2^1, S_2^2, \dots, S_2^m\}$ مجموعه لغات جمله S_2 باشند.

b. همانندی هر لغت S_1^i از جمله S_1 هر واژه S_2^j از جمله S_2 که با $Sim_{S_1^i \leftrightarrow S_2^j}$

نمایش داده می‌شود به صورت زیر محاسبه می‌شود:

i. اگر دو لغت یکسان باشند، $Sim_{S_1^i \leftrightarrow S_2^j}$ برابر با β قرار می‌گیرد، در غیر

این صورت، $Sim_{S_1^i \leftrightarrow S_2^j}$ به صورت زیر محاسبه می‌شود:

$$Sim_{S_1^i \leftrightarrow S_2^j} = \frac{n_{ij}}{n_i + n_j - n_{ij}}$$

که n_i فراوانی لغت S_1^i ، n_j فراوانی لغت S_2^j و n_{ij} فراوانی هم‌رخدادی این دو لغت می‌باشد که این فراوانی‌ها با توجه به پیکره‌ای از متون فارسی محاسبه شده‌اند.

c. همانندی هر لغت S_1^i از جمله S_1 با جمله S_2 که با $Sim_{S_1^i \leftrightarrow S_2}$ نمایش داده

می‌شود برابر با بیشینه $Sim_{S_1^i \leftrightarrow S_2^j}$ ($j = 1, \dots, m$) قرار داده می‌شود.

d. به همین ترتیب همانندی هر واژه S_2^j از جمله S_2 با جمله S_1 محاسبه شده و با

$Sim_{S_2^j \leftrightarrow S_1}$ نمایش داده می‌شود.

e. میزان همانندی جمله S_1 به جمله S_2 و میزان همانندی جمله S_2 به S_1 به ترتیب

با استفاده از فرمول‌های زیر محاسبه می‌شود:

$$Sim_{S_1 \rightarrow S_2} = \frac{\sum_{i=1}^n Sim_{S_1^i \rightarrow S_2}}{n}$$

$$Sim_{S_2 \rightarrow S_1} = \frac{\sum_{j=1}^m Sim_{S_2^j \rightarrow S_1}}{m}$$

f. دو جمله همانند تشخیص داده می‌شوند اگر:

i. $Sim_{S_2 \rightarrow S_1} > \alpha_1$ و

ii. $Sim_{S_1 \rightarrow S_2} > \alpha_2$

که α_1 و α_2 مقادیری آستانه‌ای بوده و با انجام آزمایشات محاسبه می‌شوند.

۳- پس از مشخص شدن جملات مشابه دو متن، فرایند زیر بر روی متن مشکوک و متن منبع احتمالی انجام شده تا قسمت‌های همانند تشخیص داده شده مجزا به یکدیگر متصل شده و قسمت‌های با طول کم حذف شوند:

c. در صورتی که دو جمله نزدیک به هم در متن مشکوک (در متن منبع احتمالی)

همانند تشخیص داده شود و فاصله آن‌ها در متن کمتر از مقداری آستانه‌ای باشد، قسمت همانند تشخیص داده شده برابر با از ابتدای جمله اول تا انتهای جمله دوم در نظر گرفته می‌شود. در پیاده‌سازی‌های صورت گرفته، از نصف طول دو جمله همانند تشخیص داده شده به عنوان این مقدار آستانه‌ای استفاده شده است.

d. در صورتی که قسمتی از متن همانند تشخیص داده شده و پس از انجام مرحله قبل

طول آن هم‌چنان کمتر از مقداری آستانه‌ای باشد، این قسمت از مجموعه متون همانند تشخیص داده شده حذف می‌گردد. در پیاده‌سازی‌های صورت گرفته از ۴ کلمه به عنوان این مقدار آستانه‌ای استفاده شده است.

۴-۲- ارزیابی الگوریتم‌های همانندجوی پیشنهادی

در این بخش، در ابتدا معیارهای ارزیابی یک الگوریتم همانندجو را بررسی کرده و در ادامه، پس از بیان جزئیات پیاده‌سازی و آزمایشات صورت گرفته بر روی الگوریتم‌های پیشنهادی، نتایج این آزمایشات را ارائه کرده و الگوریتم‌های ارائه شده را با هم مقایسه می‌کنیم.

۴-۲-۱- معیارهای ارزیابی یک الگوریتم همانندجو

در این بخش، معیارهای بررسی کیفیت یک الگوریتم همانندجو ارائه می‌شود تا در بخش‌های بعد، با استفاده از آن‌ها، کیفیت الگوریتم پیشنهادی بررسی گردد. برای ارزیابی کیفیت یک الگوریتم همانندجو از ۳ پارامتر دقت^۱، فراخوانی^۲، دانه‌دانه‌بودن^۳ (Potthast et al., 2010) استفاده می‌شود. لازم به ذکر است که پارامترهای دقت و فراخوان بیان شده در مورد همانندجویی اندکی با آنچه در سایر حوزه‌های پردازش زبان طبیعی مورد استفاده قرار می‌گیرد متفاوت است. در ادامه این پارامترها و نحوه محاسبه آن‌ها بیان خواهد شد.

فرض کنیم یک الگوریتم همانندجو را بر روی یک متن مشکوک و یک متن منبع احتمالی اجرا کنیم. مجموعه R را به عنوان نمایش‌گر مجموعه موارد همانند تشخیص داده شده توسط الگوریتم در نظر بگیریم. همچنین، فرض کنیم S نمایش‌گر مجموعه موارد همانند بین این دو متن باشد. در این صورت، پارامترهای دقت و فراخوان به صورت زیر تعریف می‌شوند:

$$prec(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{|U_{s \in S} s \cap r|}{|r|}$$

و

$$rec(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{|U_{r \in R} s \cap r|}{|s|}$$

که $s \cap r$ کاراکترهای هم‌پوشان بین s و r است زمانی که آن‌ها در هر دو قسمت از متن مشکوک و متن منبع احتمالی حداقل یک کاراکتر اشتراک داشته باشند. در غیر این صورت $s \cap r$ تهی است. این معیارها، بدون توجه به طول مورد همانندی، وزن یکسانی به همه موارد همانند می‌دهند. علاوه بر این، لازم به ذکر است که این معیارها درصد همانندی اختصاص یافته توسط الگوریتم همانندجو به موارد همانند تشخیص داده شده را در نظر نمی‌گیرند (Stamatatos, 2011).

¹ Precision

² Recall

³ Granularity

در همانندجویی، پارامترهای دقت و فراخوان تصویری کلی از تاثیرگذاری روش‌ها ارائه نمی‌کنند. در شرایطی که یک الگوریتم همانندجو قسمت‌های هم‌پوشان از متن منبع احتمالی را به عنوان منبع برای یک مورد همانندی ارائه کند یا در مواردی که قسمت‌های متن منبع احتمالی به قسمت‌های کوچک‌تر تقسیم شده و بازگردانده شوند، پارامترهای مورد نظر تغییر خواهند کرد (افزایش خواهند یافت). در نتیجه، به معیار دیگری نیاز است که این موارد را در نظر بگیرد. برای تعریف معیار جدید، فرض کنیم $S_R \subseteq S$ موارد تشخیص داده شده در R باشند و $R_S \subseteq R$ موارد تشخیص داده شده مربوط به S باشند. حال معیار دانه‌دانه‌بودن به صورت زیر تعریف می‌شود:

$$Gran(S, R) = \frac{1}{|S_R|} \sum_{S \in S_R} |R_S|$$

کم‌ترین و بهینه‌ترین مقدار برای این پارامتر ۱ می‌باشد. هر چقدر مقدار این پارامتر بیشتر باشد قسمت‌های متن (احتمالاً هم‌پوشان) بیشتری برای یک مورد مشابه همانندی تشخیص داده شده است. پارامترهای دقت، فراخوان و دانه‌دانه‌بودن را می‌توان با یکدیگر ترکیب و به صورت یک معیار کلی $plag\ det$ که به صورت زیر تعریف می‌شود نمایش داد:

$$plag\ det(S, R) = \frac{F_1}{\log_2(1 + Gran(S, R))}$$

که F_1 میانگین هارمونی^۱ مقادیر پارامترهای دقت و فراخوان است. لازم به ذکر است که $plag\ det$ پارامتر نهایی است که از آن در رتبه‌بندی الگوریتم‌های همانندجویی ارائه شده در مسابقات همانندجویی استفاده شده است.

۴-۲-۲- نتایج تجربی ارزیابی الگوریتم‌های همانندجویی پیشنهادی

در راستای بررسی نحوه کارکرد الگوریتم‌های پیشنهادی، آن‌ها را پیاده‌سازی نموده و آزمایشاتی بر

^۱ میانگین هارمونیک (harmonic mean) چند پارامتر برابر است با تعداد آنها تقسیم بر مجموع وارون آنها.

روی آن‌ها انجام داده شده است که در ادامه جزئیات فعالیت‌های انجام شده در این قسمت و نتایج آن بیان خواهد شد.

(۱) نحوه پیاده سازی

در این قسمت، ملاحظات صورت گرفته در پیاده‌سازی الگوریتم‌های پیشنهادی بیان می‌شود. همان‌طور که بیان شد یکی از پیش‌پردازش‌های موردنیاز به انجام در روش‌های پیشنهادی، انجام ریشه‌یابی لغوی است. در آزمایشات صورت گرفته، برای ریشه‌یابی لغوی از نسخه ۱,۶ "مجموعه نرم‌افزار پردازش زبان طبیعی دانشگاه فردوسی مشهد"^۱ استفاده شده است. متأسفانه به دلیل عدم امکان دسترسی به فایل کتابخانه‌ای این مجموعه، امکان استفاده از این ابزار به عنوان تابعی از برنامه نوشته شده وجود نداشته و ریشه‌یابی‌های لغوی مورد نیاز در آزمایشات به صورت دستی و با استفاده از فایل اجرایی^۲ این مجموعه انجام شده است. از دیگر پیش‌پردازش‌های مورد نیاز در الگوریتم‌های ارائه شده حذف ایست‌واژه‌ها می‌باشد که در الگوریتم‌های پیشنهادی از لیست ایست‌واژه‌های فراهم شده در (Taghva et al., 2003) استفاده شده است.

ابزار دیگر لازم برای پیاده‌سازی الگوریتم معنایی پیشنهادی، لغت‌نامه‌ای جهت بررسی نزدیکی معنایی دو واژه می‌باشد. نتایج تحقیقات فصل دوم این طرح پژوهشی نشان می‌دهد که در زبان انگلیسی، اغلب پژوهشگران از شبکه واژگانی ارائه شده برای این زبان به عنوان لغت‌نامه استفاده کرده‌اند. شبکه واژگانی موجود برای زبان فارسی به کاملیت نمونه مشابه برای زبان انگلیسی نبوده اما به نظر می‌رسد که بهترین انتخاب به عنوان لغت‌نامه می‌باشد. در صورت استفاده از شبکه واژگانی، با توجه به ساختار این ابزار، علاوه بر در نظر گرفتن مترادف و متضادهای واژگان، نزدیکی معنایی و مفهومی آن‌ها را نیز می‌توان در نظر گرفت. متأسفانه فایل منبع این ابزار مفید تنها به زبان برنامه‌نویسی جاوا موجود بوده و از آن‌جا که آزمایشات صورت گرفته توسط نویسنده با استفاده از زبان برنامه‌نویسی C# انجام شده است، امکان استفاده از این ابزار در آزمایشات وجود نداشته است. با توجه به این موضوع، برای ارزیابی الگوریتم همانندجوی معنایی ارائه شده، در پیاده‌سازی‌های صورت گرفته از "فرهنگ جامع

^۱ http://wtlab.um.ac.ir/index.php?option=com_content&view=article&id=220&Itemid=200&lang=fa

^۲ .exe

واژگان مترادف و متضاد زبان فارسی " تألیف فرج الله خداپرستی (Khodaparasti, 1997) استفاده شده است که تنها در برگیرنده مترادف ها و متضادهای واژگان می باشد.

الگوریتم همانندجوی فازی پیشنهادی نیز نیازمند محاسبه فراوانی واژگان و فراوانی هم‌رخدادی واژگان با توجه به پیکره‌ای عظیم از متون فارسی است. برای محاسبه این مقادیر و با توجه به قدرت محاسباتی در دسترس، از قسمتی از پیکره همشهری (AleAhmad et al., 2009) (حدوداً ۲۰٪ تصادفا انتخاب شده از این پیکره) به عنوان پیکره مرجع استفاده شده و جدول هم‌رخدادی لغات در هر دنباله ۳۱ کلمه‌ای پیاپی از متون این مجموعه محاسبه شده است.

لازم به ذکر است که با توجه به ابزارهای مورد استفاده، مقادیر بهینه برای پارامترهای α_1 ، α_2 ، α'_1 ، α'_2 و β در الگوریتم همانندجوی معنایی ارائه شده به ترتیب برابر با ۰,۵۵، ۰,۵۵، ۰,۵۵، ۰,۵۵ و ۰,۹ و مقادیر پارامترهای α_1 ، α_2 و β در الگوریتم فازی پیشنهادی به ترتیب برابر با ۰,۰۴۵، ۰,۰۴۵ و ۰,۱ قرار گرفته است.

(۲) ارزیابی

برای ارزیابی الگوریتم‌های ارائه شده از مجموعه داده آزمایش منتشر شده با این هدف و برای کنفرانس تشخیص سرقت علمی در زبان فارسی^۱ استفاده شده است. این مجموعه داده شامل مجموعه‌ای از متون منبع و مجموعه‌ای از متون مشکوک می‌باشد که در آن متن منبع احتمالی هر متن مشکوک نیز مشخص شده است. علاوه بر این، قسمت‌های همانند موجود در هر زوج نیز مشخص شده می‌باشد که با استفاده از آن بررسی صحت نتایج الگوریتم‌های پیشنهادی ممکن می‌باشد.

موارد سرقت علمی موجود در این مجموعه به ۴ دسته تقسیم شده است:

- ۱- No plagiarism
- ۲- No obfuscation
- ۳- Random obfuscation
- ۴- Simulated obfuscation

¹ <http://www.ictrc.ac.ir/plagdet/>

برای ارزیابی الگوریتم‌های پیشنهادی آن‌ها را بر روی قسمتی (که به صورت تصادفی انتخاب شده است) از زوج متون مشکوک و منبع احتمالی هر دسته اعمال کرده و نتایج آن‌ها را محاسبه می‌کنیم. لازم به ذکر است که دلیل اعمال الگوریتم‌های پیشنهادی بر روی قسمتی از داده‌های آزمایش عدم دسترسی به فایل کتابخانه‌ای ابزار ریشه‌یابی لغوی و سختی انجام دستی این فرایند است.

در دسته Simulated obfuscation که شامل متون همانند شبیه‌سازی شده توسط کاربر انسانی است، مقادیر پارامترهای موردنیاز برای بررسی کیفیت الگوریتم معنایی طراحی شده به شرح زیر محاسبه شده است: پارامتر دقت برابر با ۰,۹۷ و پارامتر فراخوان برابر با ۰,۵۶. با توجه به این مهم و مقدار یک به عنوان پارامتر دانه‌دانه‌بودن با توجه به پیوستگی قسمت‌های همانند تشخیص داده شده، پارامتر نهایی همانندجویی یعنی plagdet برابر با ۰,۷۱ می‌شود.

این آزمایش با داده‌های یکسان بر روی الگوریتم فازی پیشنهادی نیز اجرا شده و مقادیر پارامترهای موردنظر به شرح زیر محاسبه شده است. پارامتر دقت برابر با ۰,۹۰ و پارامتر فراخوان برابر با ۰,۵۰. با توجه به این مهم و مقدار یک به عنوان پارامتر دانه‌دانه‌بودن، پارامتر نهایی همانندجویی یعنی plagdet برابر با ۰,۶۴ محاسبه شده است.

جدول ۳ نمونه‌هایی از اجرای الگوریتم‌های ارائه شده را ارائه می‌دهد. هر دو سطر از این جدول نمایش‌گر یک زوج متن مشکوک و منبع می‌باشد. هر سطر با شماره فرد این جدول نمایش‌گر یک متن منبع و سطر بعدی نمایش‌گر متون بازنویسی شده متناظر با آن است. قسمت‌های همانند تشخیص داده شده توسط الگوریتم معنایی ارائه شده در سطرهای زوج با فونت برجسته و قسمت‌های همانند تشخیص داده شده توسط الگوریتم فازی پیشنهادی با خط کشیده شده در زیر آن مشخص شده‌اند. لازم به ذکر است که در این جدول، در راستای جلوگیری از هدر رفتن صفحات، تنها قسمت‌های مشابه و آن‌چه توسط الگوریتم همانند تشخیص داده شده، ارائه شده است و متن منبع و متن مشکوک کامل گزارش نشده است.

جدول ۳: برخی نتایج حاصل از اجرای الگوریتم‌های ارائه شده در دسته Simulated obfuscation.

۱	وی که تحت نفوذ موبدان و نجبا، برای مقابله با گرایش‌های استقلال‌جویانه ^۱ ارامنه، الزام آنها را به قبول آیین زرتشتی، لازم می‌شمرد با مقاومت شدید مسیحی‌های آنجا مواجه شد. پیشرفت آیین عیسی در ارمنستان از مدتی پیش باعث اضطراب دولت ایران شده بود و زمامداران ایران دریافته بودند که تا اختلافات مذهبی در بین است، تصاحب ارمنستان امری ناپایدار و بی ثبات خواهد بود و مهنررسی شخصاً طرفدار اجرای فشار و تضییق بود. نتیجه مشاوراتی که بین شاهنشاه و مهنررسی و صاحبان مراتب عالی و رؤسای دیانت زرتشتی واقع شد فرمانی بود که مهنررسی بنام شاه به نجبای ارمنستان ابلاغ کرد. (۳) متن فرمان مهنررسی لازار فریبی که ظاهراً نیم قرن پس از وقایع مذکور می‌زیسته است، آن فرمان را بشرح ذیل نقل کرده است «ما اصول دیانت خود را که متکی بر حقیقت و مبتنی بر اساس و قواعد محکم است نوشته و برای شما فرستادیم.
۲	او که زیر فشار موبدان و نجیب زادگان، برای مبارزه با تمایلات استقلال‌جویانه ^۲ ارمنی‌ها، دستور آنها را به قبول دین زرتشتی، اجباری حساب میکرد با برخورد سخت مسیحیان آن منطقه روبرو شد. فراگیری مسیحیت در آن منطقه از زمانی قبل موجب آشفتگی حکومت ایران شده بود و شاهان ایران فهمیده بودند که تا درگیریهای آیینی وسط است، اشغال ارمنستان غیر قابل اطمینان خواهد بود و مهنررسی شخصاً طرفدار اجرای زور و تضییق بود. نتیجه گفتگوهایی که بین شاه و مهنررسی و مسئولان و موبدان بلندپایه انجام شد دستور حکومتی بود که مهنررسی بنام حاکم به نجیب زادگان ارمنی رساند. (۳) متن دستور مهنررسی لازار فریبی که احتمالاً ۵۰ سال پس از حوادث اشاره شده زندگی میکرده است، آن حکم را به صورت زیر بیان کرده است «ما پایه‌های دین خویش را که بر پایه حقیقت و قواعد محکم استوار است مرقوم کرده و برایتان ارسال کردیم.
۳	در حالی که نیروی انتظامی خروج خودرو از چرخه معاملات رسمی را راهکار مقابله با چنین مفسده‌ای میدانند. راهکاری که پس از یکصد سال سابقه ثبت رسمی در ایران و سابقه‌ای طولانی در سایر کشورهای نظام حقوق نوشته معقول و منطقی به نظر نمی‌رسد.
۴	<u>نیروی انتظامی راهکار مقابله با اینگونه جرم‌ها را پایان معاملات خودرو دانسته است</u> . این راه حل سابقه بسیار زیادی در کشورهای جهان سوم به خصوص ایران داشته است.
۵	علاوه بر این، تاریخ‌گرایان نوین آمریکا به درک بدبینانه‌ی قدرت‌گفتمانی در بازنمایی ادبی نظم اجتماعی عصر الیزابت و جاکوب شناخته شده‌اند، اما ماتریالیست‌های فرهنگی بریتانیایی نوعی تاریخ‌گرایی را پرورانده‌اند که بدبینی کم‌تری دارد ولی از لحاظ سیاسی رادیکال‌تر است.
۶	<u>همچنین، تاریخ‌گرایان جدید آمریکا به درک بدبینانه از قدرت‌گفتمانی در بازنمایی ادبی نظم اجتماعی عصر الیزابت و جاکوب شناخته شده‌اند، اما ماتریالیست‌های فرهنگی انگلیسی</u>

	نوعی تاریخ گرایی را پرورش داده‌اند که کمتر دچار بدبینی است ولی به لحاظ سیاسی تندروتر است.
۷	محدوده جغرافیائی: واقع در شمال خاوری شهر شیراز از شمال به اراضی منابع طبیعی و از جنوب به بلوار مدرس، خاور به میدان گل سرخ و بلوار خلیج فارس و از باختر در امتداد پل غدیر به بلوار سرداران می‌باشد.
۸	موقعیت این مکان در شمال شیراز می‌باشد از شرق به میدان گل سرخ و از غرب به پل غدیر می‌رسد.
۹	این قوانین انجام عقد یک وکالت معمولی بین فروشنده و دلال خودرو بمنظور فروش خودرو را بر هزینه نموده‌است. لذا مشاورین و دلالان خودرو بجای اینکه برای فروش خودرو از فروشنده وکالت بگیرند اقدام به اخذ سفید امضای سند بیع قطعی از فروشنده می‌نمایند. لذا تغییر مقررات مالیاتی و ثبتی و کاهش هزینه‌های اسناد وکالت فروش خودرو می‌تواند تنظیم اسناد سفید امضا در دفترخانه‌ها را محدود نماید. ناجا شرکتهای زیرمجموعه بنیاد تعاون ناجا را دارای صلاحیت تنظیم اسناد خودرو میداند. ولی کانون سردفتران و معاون قوه قضائیه در سازمان ثبت اسناد و املاک فقط دفاتر اسناد رسمی را به سبب صلاحیت ذاتی این نهاد در ثبت و تنظیم روابط اجتماعی و اقتصادی و صدور اسناد لازم‌الاجرا مرجع صالح میدانند.
۱۰	انجام عقد یک وکالت معمولی بین فروشنده و دلال خودرو به منظور فروش خودرو، توسط قوانینی پرهزینه شده است. به‌همین‌منظور، دلالان و مشاوران خودرو به جای داشتن وکالت فروشنده برای فروش خودرو، اقدام به اخذ سفید امضای سند بیع قطعی از فروشنده می‌نمایند. لذا تغییر مقررات مالیاتی و ثبتی و همچنین کاهش هزینه‌های اسناد وکالت فروش خودرو، می‌تواند باعث محدودیت تنظیم اسناد سفید امضا در دفترخانه‌ها شود. ناجا، شرکت‌های زیر مجموعه بنیاد تعاونش را دارای صلاحیت تنظیم اسناد خودرو می‌داند. در صورتی که کانون سردفتران و معاون قوه قضائیه در سازمان ثبت اسناد و املاک فقط دفاتر اسناد رسمی را به علت صلاحیت ذاتی این موسسه در تنظیم و ثبت روابط اجتماعی و اقتصادی و صدور اسنادی که باید اجرا شوند، مرجع صالح می‌دانند.

در دسته Random obfuscation که شامل متون همانند شبیه‌سازی شده توسط کامپیوتر است، مقادیر پارامترهای موردنیاز برای بررسی کیفیت الگوریتم معنایی طراحی شده به شرح زیر محاسبه شده است: پارامتر دقت برابر با ۰٫۹۵ و پارامتر فراخوان برابر با ۰٫۷۴. با توجه به این مهم و مقدار یک به عنوان پارامتر دانه‌دانه‌بودن با توجه به پیوستگی قسمت‌های همانند تشخیص داده شده، پارامتر نهایی همانندجویی یعنی plagdet برابر با ۰٫۸۳ می‌شود.

این آزمایش با داده‌های یکسان بر روی الگوریتم فازی پیشنهادی نیز اجرا شده و و مقادیر پارامترهای موردنظر به شرح زیر محاسبه شده است. پارامتر دقت برابر با ۰,۹۳ و پارامتر فراخوان برابر با ۰,۸۱ با توجه به این مهم و مقدار یک به عنوان پارامتر دانه‌دانه‌بودن با توجه به پیوستگی قسمت‌های همانند تشخیص داده شده، پارامتر نهایی همانندجویی یعنی plagdet برابر با ۰,۸۶ می‌شود.

جدول ۴ نمونه‌هایی از اجرای الگوریتم‌های ارائه شده بر روی دسته Random obfuscation را نمایش می‌دهد. همانند جدول شماره ۳، در این جدول نیز هر دو سطر از این جدول نمایش‌گر یک زوج متن مشکوک و منبع می‌باشد. هر سطر با شماره فرد این جدول نمایش‌گر یک متن منبع و سطر بعدی نمایش‌گر متون بازنویسی شده متناظر با آن است. قسمت‌های همانند تشخیص داده شده توسط الگوریتم معنایی ارائه شده در سطرهای زوج با فونت برجسته و قسمت‌های همانند تشخیص داده شده توسط الگوریتم فازی با خط کشیده شده در زیر آن مشخص شده‌اند.

جدول ۴: برخی نتایج حاصل از اجرای الگوریتم ارائه شده بر روی دسته سرقت علمی Random obfuscation.

۱	عده‌ای از افسران مبارز به فرمان ستاد ارتش به کرمان تبعید شدند. طرح دستگیری روزبه نیز ریخته شد اما روزبه که در مرخصی یک‌ماهه به سر می‌برد، به جای بازگشت به خدمت، مخفی شد و از بازداشت نجات یافت.
۲	تبعید افسران مبارز به فرمان ستاد ماهی‌گیر ارتش سر کرمان رقم . مرخصی دستگیری روزبه نیز ریخته شد اما روزبه و به طرح یک‌ماهه به سر می‌برد ، به جای بازگشت وظیفه ، مخفی شد که از بازداشت رهایی یافت .
۳	در مرحله دوم هر سال نیز، حدود ۴۰ نفر برگزیده می‌شوند و برای آن‌ها دوره‌ای در طول تابستان برگزار می‌شود. در طی دوره تابستانی و در انتهای آن تعدادی آزمون برگزار می‌شود. با توجه به نمره امتحانات دوره به دانش‌پژوهان مدال‌های طلا، نقره یا برنز اهدا می‌شود. دارندگان مدال طلا در دوره دیگری که معمولاً در نیمه دوم همان سال برگزار می‌شود شرکت می‌کنند. پس از پایان دوره طلا آزمون‌های انتخاب تیم برگزار می‌شود که از میان آن‌ها اعضای تیم ملی المپiad جمهوری اسلامی ایران برگزیده می‌شوند. فهرست المپیادها شیوه مدال دهی طی مرحله سوم، در کنار کلاس‌ها آزمون‌هایی نیز برگزار می‌شود. حدود یک سوم دانش آموزان که نمره بالاتری گرفته باشند مدال طلا، کمی بیشتر از افرادی که مدال طلا گرفته‌اند، نقره و بقیه برنز می‌گیرند. در

	صورتی که نمره فردی خیلی پایین باشد، از مدال محروم می‌شود و صرفاً لوح افتخار می‌گیرد.
۴	والی دهانه مرحله دوم هر سال نیز، حدود ۴۰ نفر و لوح پودنه آن‌ها آشنا ساختن دوره‌ای در طول برنز می‌شود. <u>در طی دوره تابستانی و در فردیت برگزار دم احصائیه آزمون برگزار می‌شود فهمیده . با</u> <u>لطف به نمره امتحانات دوره به دانش پژوهان مرحله برنز ، نقره یا طلا پیشکش کردن می‌شود .</u> <u>دارندگان مدال طلا در دوره دیگری که معمولاً در نیم دوم ذات اقدس همان خداوند</u> <u>سال برگزار می‌شود شرکت می‌کنند پس از پایان مرتبه زر جستجو تیم برگزار می‌شود که از</u> <u>کانون آن‌ها اعضای تیم ملی المپیاد جمهوری اسلامی جمهوری ایران انتخابی می</u> <u>شوند . فهرست المپیادها شیوه مدال دهی طی مدال‌های ، در کنار کلاس‌ها آزمون‌هایی</u> <u>انتهای می‌شود انتخاب . حدود یک سوم دانش که شماره پایین گرفته باشند نشان طلا ، صرفاً</u> <u>بیشتر از افرادی که مدال طلا گرفته‌اند ، نقره و امل بقیه آبادانی تابستان می‌گیرند . درب رخ</u> که فردی فراوان بالاتری باشد ، از نشان محروم می‌شود و متأسفانه کمی برای افتخار می‌گیرد .
۵	گروه کر فیلامونیک ایران (با نام پیشین گروه کر نوری) یک گروه کر ایرانی به رهبری علیرضا شفق نژاد است. تاریخچه در ابتدا این گروه با نام هم‌آواز، در سال ۱۳۷۹ توسط علیرضا شفق نژاد، با اجتماع گروهی از دوستان و شاگردانش که عمدتاً از دانشجویان و دانش‌آموختگان دانشکده موسیقی بودند، شکل گرفت. (۱) پس از اجرای چندین برنامه و تنها مدتی بعد از شکل‌گیری و آغاز همکاری گروه با محمد نوری و چندین اجرا در برخی کنسرت‌های وی، و با اضافه شدن تعدادی دیگر از شاگردان نوری و شفق نژاد که روی هم حدود ۱۰۰ نفر می‌شدند، به پیشنهاد محمد نوری در سال ۱۳۸۰ به گروه کر نوری تغییر نام داد که این نام تا سال ۱۳۸۸ باقی ماند. (۲) این گروه در سال ۱۳۸۸ و با کسب اجازه از محمد نوری به علت گستردگی فعالیت‌های هنری و کنسرت‌هایی که بدون حضور او انجام می‌گرفت و نیز دوری از شائبه بهره‌جویی از نام وی، به گروه کر فیلامونیک ایران تغییر نام داد.
۶	<u>گروه‌بندی نام گروه کر فیلامونیک (با نوری کر فیلامونیک گروه محمد اپتیکی کر یک نام</u> <u>ایرانی رهبری علیرضا شفق نژاد است . گیری جلو رفتن درگاه منتها درجه کمال این</u> <u>گروه‌بندی با نام هم‌آواز درب سال ۱۳۷۹ توسط شفق نژاد ، با اجتماع گروهی از دوستان و</u> <u>شاگردانش که عمدتاً از دانشجویان و دانش‌آموختگان دانشکده موسیقی بودند ، شکل گرفت .</u> (۱ . پس کارمند از اجرای تاریخچه و تنها مدتی بعد از گونه برنامه اوایل همکاری گروه با سرایه طلافروشی اپتیکی و زیبایی چندین اجرا در برخی کنسرت‌های وی ، و با شدن دیگر از شاگردان گروه و شفق نژاد و

روی هم حدود ۱۰۰ آمار می شدند ، به پیشنهاد محمد اپتیکی در سال ۱۳۸۰ آخر گروه کر نوری تغییر شهرت داد که این نام همانند ۱۳۸۸ باقی ماند) (۲) این در سال ۱۳۸۸ و با کسب اجازه از محمد نوری به علت گستردگی فعالیتهای هنری که کنسرتهاى که بدون حضور او به می گرفت و هم دورى از شائبهى بهره جوئى از نام به گروه رگ برنده پیشین ایران تغییر کرم داد سراسر .	
(۳) دباغ اکنون در دانشگاه تورنتو در مقطع پست دکترا مشغول به تحقیق است. (۱). پیش از این، او استادیار موسسه پژوهشی حکمت و فلسفه بود. در آذرماه ۱۳۹۱، عبدالحسین خسروپناه، رئیس این مؤسسه حکم اخراج او را به دلیل «غیبت غیرموجه» صادر کرد. دباغ با انتشار مطلبی تحت عنوان «هر آنکه کشته نشد از قبیله ما نیست»، دلایل اخراجش را شرح داد. (۴) زندگی او دوره متوسطه را در مدرسه نیکان تهران گذراند و با دیپلم علوم تجربی در سال ۱۳۷۱ فارغ التحصیل شد. همان سال در رشته داروسازی دانشگاه علوم پزشکی تهران پذیرفته شد و از تدریس کسانی چون علی خلیج، پروین پاسالار و مسعود محمودیان بهره برد. او پایان نامه خود را به راهنمایی علی خلیج نوشت و در سال ۷۷ از همان دانشگاه با مدرک دکترای داروسازی فارغ التحصیل شد. از سال ۷۷ تا سال ۷۹ مشغول خدمت سربازی بود و از سال ۷۸ تا ۸۰ با نام «سروش وفا» در بعضی نشریات داخلی با موضوعیت اندیشه و ادبیات و با تاکید بر اندیشه های سهراب سپهری قلم می زد.	۷
(۳) دباغ جلو در تورنتو در متوسطه دکترا قسمت الحاقی ۱۳۷۱ به تحقیق است. (۱) الان از این ، او باز کردن راندمان الگو خدمت موسسه پژوهشی فارغ التحصیل و مکتب فکری بود . غیبت آذرماه ۱۳۹۱) عبدالحسین خسروپناه تصمیم ، سالار این مؤسسه حکم باگیجی او را به » اندرون سال ، صادر کرد . دباغ با انتشار مطلبی زیر عنوان اب « هر ما کشته نشد از قبیله آنکه نیست » ، دلایل اخراجش را شرح داد . (۴) « زندگی او دوره مقطع را در مدرسه نیکان کارمند پایتخت ایران گذراند جاده و با دیپلم علوم تجربی در غیرموجه مشغول خرد شد . همان سال داخل رشته دیر داروسازی دانشگاه علوم داروسازی تهران پذیرفته شد و از تدریس کسانی چون سناتور ابوتراب ، ثریا پاسالار جبهه هوا و محمودیان بهره بانکی راهکار برد . او پایان نامه خود را به ارشاد علی خلیج نوشت و در سال ۷۷ از همان قلم با مدرک دکترای سال فارغ التحصیل شد . از ۷۷ تا پزشکی ۷۹ سرگرم استادیار سربازی بود و از سال ۷۸ همتا ۸۰ با نام « سروش باوفایی بعضی داخلی با موضوعیت اندیشه و ادبیات و با تاکید بر اندیشه های سپهری دانشگاه .	۸
رشته ها	۹

واحد شوشتر هم اکنون دارای ۱۲۲ رشته تحصیلی در نظام پاره وقت و تمام وقت درمقاطع کاردانی، کارشناسی، کارشناسی ارشد و دکتری حرفه‌ای می‌باشد. یکی از پر متقاضی‌ترین رشته‌های دایر در منطقه شش (دانشگاه آزاد اسلامی استان خوزستان) با عنوان دکتری حرفه‌ای دامپزشکی در این واحد تاسیس گردیده است (۳). کلینیک دامپزشکی با ۱۸۰۰ متر مربع مساحت و مجهز به بخشهای آبریان (اتاق آکواریوم و مزارع پرورش ماهی)، بخش آناتومی، بخش طیور، بخش داخلی دام کوچک، بخش داخلی دام بزرگ، بخش رادیولوژی، آزمایشگاه تشخیص، موزه جانور شناسی و بخش ژنتیک و اصلاح نژاد دام آماده ارائه خدمات مشاوره و درمانی در سطح استان می‌باشد.	
<u>رشته‌ها شهر شوشتر هم اکنون مرفه ۱۲۲ رشته تحصیلی اندرون ماشین نظامی پاره وقت و گیتی تمام وقت در کاردانی، کارشناسی و دکتری حرفه‌ای می‌باشد. مرغوب کلفت از پر، متقاضی‌ترین فیبر دایر دهانه ناحیه ۱۸۰۰ (آزاد مسلمان شدن خوزستان) با عنوان دکتری حرفه‌ای دامپزشکی در این واحد اجتماعی ردوبدل سپیدار تاسیس شرعی گردیده است، ۳) بخش کلینیک دریاچه دامپزشکی با شش متر مربع سبط مساحت به دام طیور (اتاق آکواریوم و مزارع پرورش ماهی) بخش علم تشریح جناح سهم آبریان بخش تله کوچک، مریخ نژاد داخلی بزرگ، بخش رادیولوژی (آزمایشگاه تشخیص شناسی و استخوان جناح سهم پزشکی و اصلاح کردن ارائه مهیا. سرویس مشاوره تشریفات و ژنتیک در سطح ایالت می‌باشد دانشگاه (۱)</u>	۱۰

در دسته No obfuscation که شامل متون دقیقاً کپی شده است، مقادیر پارامترهای موردنیاز برای بررسی کیفیت الگوریتم طراحی شده به شرح زیر محاسبه شده است: پارامتر دقت برابر با ۰,۹۸ و پارامتر فراخوان برابر با ۰,۹۵ با توجه به این مهم و مقدار یک به عنوان پارامتر دانه‌دانه‌بودن با توجه به پیوستگی قسمت‌های همانند تشخیص داده شده، پارامتر نهایی همانندجویی یعنی plagdet برابر با ۰,۹۶ می‌شود.

این آزمایش با داده‌های یکسان بر روی الگوریتم فازی پیشنهادی نیز اجرا شده و مقادیر پارامترهای موردنظر به شرح زیر محاسبه شده است. پارامتر دقت برابر با ۰,۸۰ و پارامتر فراخوان برابر با ۰,۹۵ با توجه به این مهم و مقدار یک به عنوان پارامتر دانه‌دانه‌بودن، پارامتر نهایی همانندجویی یعنی plagdet برابر با ۰,۸۶ می‌شود. همان‌طور که مشخص است در استفاده از الگوریتم فازی پیشنهادی و

بر روی این دسته از داده‌های آزمایش سرقت علمی پارامتر دقت مقداری نسبتاً پایین اختیار کرده است. دلیل این امر قسمت‌های اشتباه‌ها همانند تشخیص داده شده از متن مشکوک است. برای بهبود پارامتر دقت می‌توان پارامترهای به کار رفته در الگوریتم (α_1 و α_2) را افزایش داد تا دقت نتایج بهتر شود. اما در آزمایشات صورت گرفته، در راستای دستیابی به بهترین نتایج کلی این کار صورت نگرفته است.

جدول ۵ نمونه‌هایی از اجرای الگوریتم‌های ارائه شده بر روی دسته No obfuscation را نمایش می‌دهد. نحوه توصیف موارد همانند در این جدول، همانند جدول‌های ۳ و ۴ می‌باشد.

جدول ۵: برخی نتایج حاصل از اجرای الگوریتم ارائه شده بر روی دسته سرقت علمی No obfuscation.

۱	به طوری که در کتاب «کمیترون و خاور» (صفحه ۱۰۵) آمده‌است: «هنوز قبل از کنگره» چپ‌ها بین گروه رهبری کمونیست‌های ایران تفوق داشتند. اگر از جزئیات صورتجلسه و دیگر اسناد کنگره و مباحثاتی که پیرامون گرایش‌های «چپ» و راست در این کنگره شده‌است، بگذریم، یک نکته جای تردید باقی نمی‌گذارد و آن وجود بعضی اشتباهات «چپ» است، که در آن دوره (دوره تکوین و طفولیت این احزاب) برای خیلی از احزاب کمونیست روی می‌داد. برای کمک به این احزاب بود که لنین اثر مشهور خود کمونیسم چپ: بیماری کودکی را نوشت. اسناد کنگره اول حزب کمونیست ایران حاکی از آنست که ۵۱ نفر نماینده و ۹ نفر مهمان به کنگره اعزام شده بودند، که ۴۸ اعتبارنامه) نمایندگان با رای قطعی (به تصویب می‌رسد.
۲	<u>به طوری که در کتاب «کمیترون و خاور» (صفحه ۱۰۵) آمده‌است: «هنوز قبل از کنگره»</u> <u>چپ‌ها بین گروه رهبری کمونیست‌های ایران تفوق داشتند. اگر از جزئیات صورتجلسه و</u> <u>دیگر اسناد کنگره و مباحثاتی که پیرامون گرایش‌های «چپ» و راست در این کنگره</u> <u>شده‌است، بگذریم، یک نکته جای تردید باقی نمی‌گذارد و آن وجود بعضی اشتباهات</u> <u>«چپ» است، که در آن دوره (دوره تکوین و طفولیت این احزاب) برای خیلی از احزاب</u> <u>کمونیست روی می‌داد. برای کمک به این احزاب بود که لنین اثر مشهور خود کمونیسم</u> <u>چپ: بیماری کودکی را نوشت.</u> <u>اسناد کنگره اول حزب کمونیست ایران حاکی از آنست که ۵۱ نفر نماینده و ۹ نفر مهمان به</u> <u>کنگره اعزام شده بودند، که ۴۸ اعتبارنامه) نمایندگان با رای قطعی (به تصویب می‌رسد.</u>
۳	اگر از جزئیات صورتجلسه و دیگر اسناد کنگره و مباحثاتی که پیرامون گرایش‌های «چپ» و راست در این کنگره شده‌است، بگذریم، یک نکته جای تردید باقی نمی‌گذارد و آن وجود بعضی اشتباهات «چپ» است، که

	در آن دوره (دوره تکوین و طفولیت این احزاب) برای خیلی از احزاب کمونیست روی می‌داد اگر از جزئیات صورتجلسه و دیگر اسناد کنگره و مباحثاتی که پیرامون گرایش‌های «چپ» و راست در این کنگره شده است، بگذریم، یک نکته جای تردید باقی نمی‌گذارد و آن وجود بعضی اشتباهات «چپ» است، که در آن دوره (دوره تکوین و طفولیت این احزاب) برای خیلی از احزاب کمونیست روی می‌داد
۴	اگر از جزئیات صورتجلسه و دیگر اسناد کنگره و مباحثاتی که پیرامون گرایش‌های «چپ» و راست در این کنگره شده است، بگذریم، یک نکته جای تردید باقی نمی‌گذارد و آن وجود بعضی اشتباهات «چپ» است، که در آن دوره (دوره تکوین و طفولیت این احزاب) برای خیلی از احزاب کمونیست روی می‌داد
۵	ب - چنانکه بابت هر دوره مالی که برای حساب‌های کمپانی نفت انگلیس و ایران محدود (از تاریخ دوره مالی منتهی به ۳۱ دسامبر ۱۹۴۸ به بعد) تنظیم شده جمع مبلغی که کمپانی باید طبق بند الف ماده ۴ قرارداد فعلی و جزء (I - ب) ماده ۱۰ قرارداد اصلی به دولت پردازد احياناً از چهار میلیون لیره استرلینگ کمتر بشود کمپانی مابه‌التفاوت جمع مبلغ مزبور و چهار میلیون لیره استرلینگ را به دولت خواهد پرداخت. مع هذا مقرر است که چنانچه در طی یکی از دوره‌های مالی مذکور کمپانی بر اثر حوادثی که خارج از اختیار او باشد نتواند نفت از ایران صادر نماید، از میزان مبلغی که کمپانی باید بابت چنین دوره طبق مقررات سابق‌الذکر جزء (ب فرعی) قرارداد فعلی به دولت پردازد باید مبلغی کسر گردد که نسبت آن با میزان سابق‌الذکر همان نسبتی را دارا باشد که بین مدت این تعطیل و دوره مالی مذکور وجود دارد.
۶	<u>ب - چنانکه بابت هر دوره مالی که برای حساب‌های کمپانی نفت انگلیس و ایران محدود (از تاریخ دوره مالی منتهی به ۳۱ دسامبر ۱۹۴۸ به بعد) تنظیم شده جمع مبلغی که کمپانی باید طبق بند الف ماده ۴ قرارداد فعلی و جزء (I - ب) ماده ۱۰ قرارداد اصلی به دولت پردازد احياناً از چهار میلیون لیره استرلینگ کمتر بشود کمپانی مابه‌التفاوت جمع مبلغ مزبور و چهار میلیون لیره استرلینگ را به دولت خواهد پرداخت. مع هذا مقرر است که چنانچه در طی یکی از دوره‌های مالی مذکور کمپانی بر اثر حوادثی که خارج از اختیار او باشد نتواند نفت از ایران صادر نماید، از میزان مبلغی که کمپانی باید بابت چنین دوره طبق مقررات سابق‌الذکر جزء (ب فرعی) قرارداد فعلی به دولت پردازد باید مبلغی کسر گردد که نسبت آن با میزان سابق‌الذکر همان نسبتی را دارا باشد که بین مدت این تعطیل و دوره مالی مذکور وجود دارد</u>
۷	پس از چهل روز، حاج عبدالوهاب مرشد نائینی دوباره او را به منزل خویش برد و از آن زمان به مقام حقیقت دست یافت. (روحانی وصال، همان). پس از چندی، میرزا ابوالقاسم به فرمان مرادش به شیراز بازگشت و در آن جا ازدواج نمود. چندی بعد، به شوق

دیدار پیر خود، دوباره عازم نائین شد و با کسب اجازه از محضر وی، به زیارت مشهد مقدس رفت و از آنجا بار دیگر به زیارت مکه مکرمه رفت و از راه حلب، شام و عراق به شیراز بازگشت. در این حال مقام وی در شیراز به جایی رسید که «مردمان هوشمند و بزرگان فارس و پاره‌ای از رجال ایران به دور شمع وجودش پروانه‌سان جمع شدند». (همان، ص ۲۶). میرزا محمد شفیع وصال (۱۲۶۲ق. - ۱۱۹۷ق.) یکی از مریدان وی بود که تا پایان عمر شیخ، در خدمت وی به سر می‌برد. وی هم‌چنین قصیده‌ای در رثای حاج عبدالوهاب مرشد نائینی سروده‌است. (دیوان کامل وصال شیرازی، ج ۱، ص نه پیشگفتار/، ماهیار نوایی، ص ۱۵۰). میرزا ابوالقاسم پس از بازگشت به شیراز دوباره به نائین سفر کرد و به مدت بیست روز در خدمت پیر خویش به سر برده، به شیراز بازگشت و پس از یک سال مجدداً به دیدار حاج عبدالوهاب مرشد نائینی رهسپار شد. به امر وی به کاشان سفر کرد و به هنگام سکونت در کاشان از درگذشت پیر خویش، حاج عبدالوهاب مرشد نائینی مطلع شد. پس از آن، چهار سال پریشان‌وار در کاشان و اصفهان به سر برد، سپس به شیراز بازگشت، گوشه انزوا گزید و مهر سکوت بر لب زد.	۸ <u>پس از چهل روز، حاج عبدالوهاب مرشد نائینی دوباره او را به منزل خویش برد و از آن زمان به مقام حقیقت دست یافت. (روحانی وصال، همان).</u> <u>پس از چندی، میرزا ابوالقاسم به فرمان مرادش به شیراز بازگشت و در آن جا ازدواج نمود. چندی بعد، به شوق دیدار پیر خود، دوباره عازم نائین شد و با کسب اجازه از محضر وی، به زیارت مشهد مقدس رفت و از آنجا بار دیگر به زیارت مکه مکرمه رفت و از راه حلب، شام و عراق به شیراز بازگشت. در این حال مقام وی در شیراز به جایی رسید که «مردمان هوشمند و بزرگان فارس و پاره‌ای از رجال ایران به دور شمع وجودش پروانه‌سان جمع شدند». (همان، ص ۲۶).</u> <u>میرزا محمد شفیع وصال (۱۲۶۲ق. - ۱۱۹۷ق.) یکی از مریدان وی بود که تا پایان عمر شیخ، در خدمت وی به سر می‌برد. وی هم‌چنین قصیده‌ای در رثای حاج عبدالوهاب مرشد نائینی سروده‌است. (دیوان کامل وصال شیرازی، ج ۱، ص نه پیشگفتار/، ماهیار نوایی، ص ۱۵۰).</u> <u>میرزا ابوالقاسم پس از بازگشت به شیراز دوباره به نائین سفر کرد و به مدت بیست روز در خدمت پیر خویش به سر برده، به شیراز بازگشت و پس از یک سال مجدداً به دیدار حاج عبدالوهاب مرشد نائینی رهسپار شد. به امر وی به کاشان سفر کرد و به هنگام سکونت در کاشان از درگذشت پیر خویش، حاج عبدالوهاب مرشد نائینی مطلع شد. پس از آن، چهار سال پریشان‌وار در کاشان و اصفهان به سر برد، سپس به شیراز بازگشت، گوشه انزوا گزید و مهر سکوت بر لب زد.</u>
---	--

۹	(۲) نقطه ضعف اولین فیلم محمدرضا اعلامی به عنوان کارگردان بود. (۳) او پس از این فیلم فیلم‌های موفق دیگری ساخت که در گیشه هم موفق بودند. اما اکثریت منتقدان سینمای ایران، نقطه ضعف را بهترین فیلم او می‌دانند.
۱۰	<u>(۲) نقطه ضعف اولین فیلم محمدرضا اعلامی به عنوان کارگردان بود. (۳) او پس از این فیلم فیلم‌های موفق دیگری ساخت که در گیشه هم موفق بودند.</u> اما اکثریت منتقدان سینمای ایران، نقطه ضعف را بهترین فیلم او میدانند.

الگوریتم‌های ارائه شده بر روی متون موجود در دسته No plagiarism که شامل متون فاقد سرقت علمی است نیز اجرا شده است. نتایج به دست آمده نشان‌گر این است که مطابق انتظار الگوریتم‌های پیشنهادی در این قسمت هیچ تشخیصی انجام نداده‌اند.

جدول ۶ به طور خلاصه عملکرد ۲ الگوریتم ارائه شده در دسته‌های مختلف متون آزمایشی را نشان می‌دهد. نتایج گزارش شده نشان‌دهنده کیفیت بهتر الگوریتم همانندجویی معنایی پیشنهادی در مقایسه با الگوریتم همانندجویی فازی پیشنهادی می‌باشد.

ملاحظه: لازم به ذکر است که با توجه به حجم پیکره استفاده شده برای محاسبه ماتریس هم‌رخدادی و فراوانی حضور واژگان (۲۰٪ از پیکره همشهری)، نتایج به دست آمده از الگوریتم فازی پیشنهادی از کیفیت بالایی برخوردار نبوده و قطعا در صورت استفاده از پیکره مرجع کامل‌تری برای این هدف، نتایج این الگوریتم بهتر خواهد شد.

جدول ۶ مقایسه الگوریتم‌های پیشنهادی.

Simulated obfuscation				Random obfuscation				No obfuscation				الگوریتم
plagdet	دانه‌دانه‌بودن	فازی	معنایی	plagdet	دانه‌دانه‌بودن	فازی	معنایی	plagdet	دانه‌دانه‌بودن	فازی	معنایی	
۰٫۷۱	۱	۰٫۵۶	۰٫۹۷	۰٫۸۳	۱	۰٫۷۴	۰٫۹۵	۰٫۹۶	۱	۰٫۹۵	۰٫۹۸	معنایی
۰٫۶۴	۱	۰٫۵۰	۰٫۹۰	۰٫۸۶	۱	۰٫۸۱	۰٫۹۳	۰٫۸۶	۱	۰٫۹۵	۰٫۸۰	فازی

(۳) تحلیل کارایی

در این قسمت، کارایی روش‌های ارائه شده را با یکدیگر و همچنین با روش‌های همانندجویی دقیق مقایسه می‌کنیم.

در ادامه، برای سادگی انجام مقایسات، فرض می‌کنیم که در تمام روش‌ها از واحد جمله برای تقسیم‌بندی متون استفاده می‌شود. همچنین، فرض می‌کنیم که n تعداد واژگان موجود در یک واحد از متن (چه از متن مشکوک و چه از متن منبع احتمالی) باشد. همچنین فرض می‌کنیم که t تعداد متوسط واژگان مترادف و متضاد یک واژه در یک لغت‌نامه باشد.

در روش‌های همانندجویی دقیق، برای مقایسه یک واحد از متن مشکوک با یک واحد از متن منبع احتمالی، چکیده هر دو واحد با استفاده از تابعی درهم‌ساز محاسبه شده و این مقادیر با یکدیگر مقایسه می‌گردند. محاسبه چکیده واحدهای متون منبع احتمالی می‌تواند به صورت غیر برخط انجام شود. علاوه بر این، چکیده واحدهای متن مشکوک نیز یک بار انجام شده و در تمام مقایسات مورد استفاده قرار می‌گیرد و در نتیجه محاسبه چکیده به تعداد مقایسات مربوط نیست و می‌توان از آن چشم‌پوشی کرد. با توجه به این موارد، می‌توان بیان کرد که در الگوریتم‌های همانندجویی دقیق برای مقایسه یک واحد از متن مشکوک با یک واحد از متن منبع احتمالی تنها انجام یک مقایسه مورد نیاز است.

روش همانندجوی معنایی ارائه شده، برای مقایسه یک واحد از متن مشکوک و یک واحد از متن منبع احتمالی، $2n^2t$ مقایسه انجام می‌دهد. با توجه به این مهم، این روش از روش‌های همانندجویی دقیق ناکارتر می‌باشد.

روش همانندجویی فازی ارائه شده محاسبه میزان شباهت میان هر دو واژه را به صورت غیر برخط انجام می‌دهد. با توجه به این موضوع، برای مقایسه یک واحد از متن مشکوک با یک واحد از متن منبع احتمالی، این الگوریتم n^2 مقایسه انجام می‌دهد. در نتیجه، روش فازی پیشنهادی از روش معنایی پیشنهادی کارتر بوده اما در مقایسه با روش‌های همانندجویی دقیق از کارایی کمتری برخوردار است.

لازم به ذکر است که با توجه به اینکه طول جملات و همچنین تعداد واژگان مترادف و متضاد متناظر با هر واژه‌ای در لغت‌نامه در مقایسه با تعداد مقایسات مورد نیاز به انجام در عمل همانندجویی ناچیز و

عددی ثابت می‌باشد، می‌توان نتیجه گرفت که محاسبات موردنیاز به انجام در هر دو روش پیشنهادی برابر با مضربی کوچکی از محاسبات موردنیاز به انجام در روش‌های همانندجویی دقیق خواهد شد. به عبارت دیگر، می‌توان نتیجه گرفت که الگوریتم‌های پیشنهادی و الگوریتم‌های تشخیص کپی دقیق از پیچیدگی محاسباتی یکسانی برخوردار هستند.

نتیجه گیری

در این طرح پژوهشی، در ابتدا الگوریتم‌های موجود برای همانندجویی در سایر زبان‌ها مورد بررسی قرار گرفت. در ادامه و پس از استخراج لیست ابزارهای پردازش زبان طبیعی مورد استفاده در سایر زبان‌ها برای همانندجویی، نمونه‌های ارائه شده از این ابزارها برای پردازش زبان فارسی مورد بررسی قرار گرفت. سپس، دو الگوریتم جدید برای همانندجویی در متون فارسی بازنویسی شده ارائه شد. الگوریتم اول ارائه شده یک الگوریتم همانندجویی معنایی بوده که در آن همانندی دو واحد از متن با توجه به یک لغت‌نامه بررسی و تعیین می‌شود. الگوریتم دوم در دسته الگوریتم‌های همانندجویی فازی قرار گرفته و همانندی واحدهای متن 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 درحالی که کارایی الگوریتم فازی پیشنهادی بهتر است، الگوریتم معنایی پیشنهادی کیفیت بهتری را ارائه می‌نماید. لازم به ذکر است که در الگوریتم‌های پیشنهادی، همانندی هر جمله از متن مشکوک با هر جمله از متن منبع احتمالی به صورت مستقل بررسی شده است. اما، با توجه به این نکته که اغلب قسمت‌های همانند در متن مشکوک و متن منبع کنار هم‌دیگر می‌باشند، به نظر می‌رسد بتوان با در نظر گرفتن همانندی هم‌زمان جملات همسایه کیفیت الگوریتم‌های پیشنهادی را بهبود داد که این مهم از عهده این طرح خارج بوده و به عنوان یک کار آتی در نظر گرفته می‌شود.

واژه‌نامه فارسی به انگلیسی

Stopword	ایست‌واژه
Nonverbal stopwords	ایست‌واژه‌های غیر فعلی
Verbal stopwords	ایست‌واژه‌های فعلی
Semantic role labeling	برچسب‌زنی نقش معنایی
Longest common subsequence	بزرگ‌ترین زیررشته مشترک
Corpus	پیکره
Hash function	تابع درهم‌ساز
Text to speech conversion	تبدیل متن به گفتار
Machine Translation	ترجمه ماشینی
Part of speech recognizer	تشخیص‌دهنده اقسام کلمه
String matching	تطبیق رشته
Approximate string matching	تطبیق رشته تقریبی
Jaccard	جاکارد
Error detection	خطایابی
Granularity	دانه‌دانه‌بودن
Dice	دایس
Containment	دربرگرفتن
Precision	دقت
Part of speech	دنباله اقسام کلمه
Statistical methods	روش‌های آماری
Structural methods	روش‌های ساختاری
Lookup table based methods	روش‌های مبتنی بر جدول جستجو

Rule based methods	روش های مبتنی بر قاعده
Morphologically	ریختی
Stemming	ریشه یابی لغوی
Deep neural network	شبکه عصبی عمیق
Wordnet	شبکه واژگانی
Haming distance	فاصله همینگ
Recall	فراخوانی
Chunking	قسمت بندی
Cosine	کوساین
Secure multi party computation	محاسبات چند نفره امن
Hidden Marcov model	مدل مخفی مارکوف
harmonic mean	میانگین هارمونیک
Homograph	هم نگاره

واژه‌نامه انگلیسی به فارسی

Approximate string matching	تطبیق رشته تقریبی
Chunking	قسمت‌بندی
Containment	درب‌گرفتنی
Corpus	پیکره
Cosine	کوساین
Deep neural network	شبکه عصبی عمیق
Dice	دایس
Error detection	خطایابی
Granularity	دانه‌دانه بودن
Haming distance	فاصله همینگ
harmonic mean	میانگین هارمونیک
Hash function	تابع درهم‌ساز
Hidden Marcov model	مدل مخفی مارکوف
Homograph	هم‌نگاره
Jaccard	جاکارد
Longest common subsequence	بزرگ‌ترین زیررشته مشترک
Lookup table based methods	روش‌های مبتنی بر جدول جستجو
Machine Translation	ترجمه ماشینی
Morphologically	ریختی
Nonverbal stopwords	ایست‌واژه‌های غیر فعلی
Part of speech	دنباله اقسام کلمه
Part of speech recognizer	تشخیص‌دهنده اقسام کلمه

Precision	دقت
Recall	فراخوانی
Rule based methods	روش های مبتنی بر قاعده
Secure multi party computation	محاسبات چندنفره امن
Semantic role labeling	برچسب زنی نقش معنایی
Statistical methods	روش های آماری
Stemming	ریشه یابی لغوی
Stopword	ایست واژه
String matching	تطبیق رشته
Structural methods	روش های ساختاری
Text to speech conversion	تبدیل متن به گفتار
Verbal stopwords	ایست واژه های فعلی
Wordnet	شبکه واژگانی

مراجع

- صدرموسوی، م. و م. شمس فرد (۱۳۸۶). برجسب‌زنی نقش‌های معنایی با استفاده از تجزیه سطحی جملات فارسی. سیزدهمین کنفرانس سالانه انجمن کامپیوتر ایران، جزیره کیش، ایران.
- کامل قالی‌باف، آ.، س. راحتی قوچانی، و ا. استاجی (۱۳۸۸). برجسب‌زنی خودکار نقش معنایی جملات فارسی با رویکرد مبتنی بر پیکره. پانزدهمین کنفرانس بین‌المللی سالانه انجمن کامپیوتر ایران، تهران، ایران.
- Adam, A. R. (2014). Plagiarism detection algorithm using natural language processing based on grammar analyzing. *Journal of Theoretical & Applied Information Technology*, 63(1).
- Alavi, S. A., B. Minaei, and M. Okhovvat (2014). PSTM: Part-Of-Speech Tagger for Persian Based on Markov Model. *Bulletin of the Georgian national academy of sciences*, 8(2); 406-412.
- AleAhmad, A. et al. (2009). Hamshahri: A Standard Persian Text Collection. *Knowledge-Based Systems*, 22(5); 382–387.
- Alzahrani, S. M., and N. Salim (2009 a). On the use of fuzzy information retrieval for gauging similarity of Arabic documents. *Applications of Digital Information and Web Technologies*, 2009. ICADIWT'09. Second International Conference on the. IEEE; 539-544.
- Alzahrani, S. M., and N. Salim (2009 b). Statement-Based Fuzzy-Set Information Retrieval versus Fingerprints Matching for Plagiarism Detection in Arabic Documents. 5th Postgraduate Annual Research Seminar (PARS'09).
- Alzahrani, S. M., N. Salim, and A. Abraham (2012). Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42(2); 133-149.
- Alzahrani, S. M., N. Salim, and V. Palade (2015). Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model. *Journal of King Saud University-Computer and Information Sciences*, 27(3); 248-268.
- Assi, S. M. (2000). Grammatical tagging of a Persian corpus. *International Journal of Corpus Linguistics*, 5(1); 69-81.

- Azimizadeh, A., M. Arab, and S. Rahati Quchani (2008). Persian part of speech tagger based on Hidden Markov Model. 9th International Conference on the Statistical Analysis of Textual Data, Lyon, France; 121-128.
- Barrón-Cedeño, A., P. Gupta, and P. Rosso (2013). Methods for cross-language plagiarism detection. Knowledge-Based Systems 50; 211-217.
- Barrón-Cedeño, A., and P. Rosso (2009). On automatic plagiarism detection based on n-grams comparison. European Conference on Information Retrieval; 696-700.
- Barrón-Cedeño, A. et al. (2010). Word length n-Grams for text re-use detection. International Conference on Intelligent Text Processing and Computational Linguistics; 687-699.
- Basile, C., et al. (2009). A plagiarism detection procedure in three steps: Selection, matches and “squares”. Proc. SEPLN, Donostia, Spain; 19-23.
- Eissen, Z., S. Meyer, and B. Stein (2006). Intrinsic plagiarism detection. Advances in Information Retrieval; 565-569.
- Elhadi, M., and A. Al-Tobi (2008). Use of text syntactical structures in detection of document duplicates. Digital Information Management, 2008. ICDIM 2008. Third International Conference on. IEEE; 520-525.
- Elhadi, M., and A. Al-Tobi (2009). Duplicate detection in documents and webpages using improved longest common subsequence and documents syntactical structures. Computer Sciences and Convergence Information Technology, 2009. ICCIT'09. Fourth International Conference on. IEEE.
- Fadaei, H., and M. Shamsfard (2010). Persian POS tagging using probabilistic morphological analysis. International Journal of Computer Applications in Technology, 38(4); 264-273.
- Gharavi, E. et al. (2016). A Deep Learning Approach to Persian Plagiarism Detection. Working notes of FIRE 2016-Forum for Information Retrieval Evaluation; 154-159.
- Gipp, B., and N. Meuschke (2011). Citation pattern matching algorithms for citation-based plagiarism detection: greedy citation tiling, citation chunking and longest common citation sequence. Proceedings of the 11th ACM symposium on Document engineering; 249-258.

- Grozea, C., C. Gehl, and M. Popescu (2009). ENCOPLLOT: Pairwise sequence matching in linear time applied to plagiarism detection. 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse.
- Hannabuss, S. (2001). Contested texts: Issues of plagiarism. *Library Management*, 22; 311-318.
- Hesamifard, R., and G. Ghassem-Sani (2006). A stemming algorithm for the Persian words. In *Proceedings of the 11th Annual International CSI Computer Conference (CSICC2006)*, Tehran, Iran; 515-519.
- Jadidinejad, A., F. Mahmoudi, and J. Dehdari (2010). Evaluation of PerStem: a simple and efficient stemming algorithm for Persian. In *Workshop of the Cross-Language Evaluation Forum for European Languages*; 98-101.
- Jurafsky, D., and J. H. Martin (2010). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Chapter 22, Boston, MA: MIT Press.
- Karimpour, R. et al. (2009). Improving Persian information retrieval systems using stemming and part of speech tagging. *Evaluating Systems for Multilingual and Multimodal Information Access*; 89-96.
- Kasprzak, J., M. Brandejs, and M. Kripac (2009). Finding plagiarism by evaluating document similarities. *Proc. SEPLN*, 9(4); 24-28.
- Khodaparasti, F. (1997). *A Comprehensive Dictionary of Persian Synonyms and Antonyms*. Fars Encyclopedia.
- Koberstein, J., and Y.-K. Ng (2006). Using word clusters to detect similar web documents. *International Conference on Knowledge Science, Engineering and Management*; 215-228.
- Kong, L. et al. (2013). Approaches for Source Retrieval and Text Alignment of Plagiarism Detection. *Notebook for PAN at CLEF 2013*.
- Kong, L. et al. (2015). HawkEyes Plagiarism Detection System. *International Conference of Young Computer Scientists, Engineers and Educators*; 459-463.
- Li, Y. et al. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE transactions on knowledge and data engineering*, 18(8); 1138-1150.

- Mashhadirajab, F., and M. Shamsfard (2016). A Text Alignment Algorithm Based on Prediction of Obfuscation Types Using SVM Neural Network. Working notes of FIRE 2016-Forum for Information Retrieval Evaluation; 167-171.
- Meuschke, N., and B. Gipp (2014). Reducing computational effort for plagiarism detection by using citation characteristics to limit retrieval space. Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries; 197-200.
- Minaei, B., and M. Niknam (2016). An n-gram based Method for Nearly Copy Detection in Plagiarism Systems. Working notes of FIRE 2016-Forum for Information Retrieval Evaluation.
- Mohtarami, M. et al. (2006). Using heuristic rules to improve Persian part of speech tagging accuracy. 6th International Conference on Informatics and Systems, INFOS 2006.
- Mokhtaripour, A., and S. Jahanpour (2006). Introduction to a new Farsi stemmer. In Proceedings of the 15th ACM international conference on Information and knowledge management; 826-827.
- Momenipour, F., and M. R. Keyvanpour (2016). PHMM: Stemming on Persian Texts using Statistical Stemmer Based on Hidden Markov Model. International Journal of Information Science & Management, 14(2); 107-117.
- Momtaz, M., et al. (2016). Graph-based Approach to Text Alignment for Plagiarism Detection in Persian Documents. Working notes of FIRE 2016-Forum for Information Retrieval Evaluation; 176-179.
- Mosavi Miangah, T., and A. Delavar Khalafi (2012). Unsupervised part of speech tagging for Persian. International Journal of Artificial Intelligence & Applications, 3(2); 33.
- Murugesan, M. et al. (2010). Efficient privacy-preserving similar document detection. The VLDB Journal—The International Journal on Very Large Data Bases, 19(4); 457-475.
- Nasiri, M. M., K. Sheikh Esmaeili, and H. Abolhassani (2006). A statistical stemmer for Persian language. 11th International CSI Computer Conference (CSICC'2006), School of Computer Science, IPM.
- Osman, A. H. et al. (2012). An improved plagiarism detection scheme based on semantic role labeling. Applied Soft Computing, 12(5); 1493-1502.

- Potthast, M., et al. (2009). Overview of the 1st international competition on plagiarism detection. In Proceedings of the 3rd Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse; 1-9.
- Potthast, M. et al. (2010). An evaluation framework for plagiarism detection. In Proceedings of the 23rd international conference on computational linguistics: Posters. Association for Computational Linguistics; 997-1005.
- Potthast, M., et al. (2013). Overview of the 5th International Competition on Plagiarism Detection. In: Forner, P., Navigli, R., Tufis, D. (eds.) Working Notes Papers of the CLEF 2013 Evaluation Labs.
- Rahimi, A. (2015). A new hybrid stemming algorithm for Persian. arXiv preprint arXiv: 1507.03077.
- Raja, F., S. Tasharofi, and F. Oroumchian (2007). Statistical POS tagging experiments on Persian text. Second Workshop on Computational Approaches to Arabic Script-based Languages; 128-133.
- Ram, R., V. Sundar, E. Stamatatos, and S. L. Devi (2014). Identification of Plagiarism Using Syntactic and Semantic Filters. International Conference on Intelligent Text Processing and Computational Linguistics; 495-506.
- Rezai, M. J., and T. Mosavi Miangah (2016). FarsiTag: A Part-of-Speech Tagging System for Persian. Digital Scholarship in the Humanities; fqw017.
- Sanchez-Perez, M. A., G. Sidorov, and A. F. Gelbukh (2014). A Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014. CLEF (Working Notes); 1004-1011.
- Scherbinin, V., and S. Butakov (2009). Using Microsoft SQL server platform for plagiarism detection. In Proc. SEPLN; 36-37.
- Shamsfard, M., and M. Sadr Mousavi (2008). Thematic role extraction using shallow parsing. International Journal of Computational, Intelligence, 4(2); 126-132.
- Shamsfard, M. et al. (2010). Semi automatic development of FarsNet; the Persian wordnet. In Proceedings of 5th global WordNet conference, Mumbai, India.
- Sharifloo, A.A., and M. Shamsfard (2008). A Bottom Up approach to Persian Stemming. ” in Proceedings of the Third International Joint Conference on Natural Language Processing; 583-588.
- Stamatatos, E. (2011). Plagiarism detection using stopword n- grams. Journal of the American Society for Information Science and Technology, 62; 2512-2527.

- Taghizadeh, N., and H. Faili (2016). Automatic Wordnet Development for Low-Resource Languages using Cross-Lingual WSD. *Journal of Artificial Intelligence Research*, 56; 61-87.
- Taghva, K., R. Beckley, and M. Sadeh (2003). A list of Farsi stopwords. Technical Report 2003-01, Information Science Research Institute, University of Nevada, Las Vegas, July 2003.
- Taghva, K., R., Beckley, and M. Sadeh (2005). A stemming algorithm for the Farsi language. *Information Technology: Coding and Computing*, 2005. ITCC 2005. International Conference on. Vol. 1. IEEE; 158-162.
- Tashakori, M., M. Meybodi, and F. Oroumchian (2002). Bon: The Persian stemmer. *EurAsia-ICT 2002: Information and ommunication Technology*; 487-494.
- Unger, N., S. Thandra, and I. Goldberg (2016). Elxa: Scalable Privacy-Preserving Plagiarism Detection. *Proceedings of the 2016 ACM on Workshop on Privacy in the Electronic Society*; 153-164.
- Velásquez, J. D., et al. (2016). DOCODE 3.0 (DOcument COpy DEtector): A system for plagiarism detection by applying an information fusion process from multiple documental data sources. *Information Fusion*, 27; 64-75.
- White, D. R., and M. S. Joy (2004). Sentence-based natural language plagiarism detection. *Journal on Educational Resources in Computing*, 4(4); 2-2.
- Wu, Z., and M. Palmer (1994). Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*; 133–139.
- Yerra, R., and Y.-K. Ng (2005). A sentence-based copy detection approach for web documents. *International Conference on Fuzzy Systems and Knowledge Discovery*; 481-482.

Abstract

Text alignment is the main part of plagiarism detection process in which on input of a suspicious document and a small set of candidate source documents, the suspicious document is compared to each of the candidate sources with the aim of the detection of highly similar passages. The text alignment is a highly language dependent task in a way that the existing algorithms for one language can not be employed in another one. Unfortunately, the existing approaches for the text alignment task in Persian language are inefficient when implemented in real world or are only able to detect exactly copied passages. To overcome these issues, in this research project, we propose two new algorithms for detection of paraphrased passages in Persian language. While the first algorithm employs semantic similarity of pairs of words obtained from a dictionary to detect highly similar passages, the second one uses the coocurrence frequencies of pairs of words obtained from a large corpus. Compared to each other, while the proposed fuzzy based algorithm provides more efficiency, the semantic based one offers better results.

Keywords: Plagiarism, similarity detection, text alignment, paraphrased passages.



**Iranian Research Institute
for Information Science and Technology (Irandoct)**

**Designing a similarity search algorithm for detection
of paraphrased texts in Persian language**

Nasrollah Pakniat

Summer 2017