ترجمه فا
TarjomeFa.Com

ارائه شده توسط:

سایت ترجمه فا

مرجع جدیدترین مقالات ترجمه شده

از نشریات معتبر

DOCODE 3.0(شناساگر کپی اسناد): سیستمی برای تشخیص سرقت ادبی با استفاده از فرایند تلفیق اطلاعات از منابع داده های اسنادی مختلف

چکیده

سرقت ادبی اشاره به فرایند ارایه کلمات، افکار و ایده های افراد دیگر به صورت کلمات، افکار و ایده های خود بدون رفرنس دادن به منابع آن ها دارد. رشد ازمایشی منابع اسناد دیجیتالی مختلف موجود در اینترنت موجب تسهیل توسعه این عمل شده و در نهایت موجب شده است تا تشخیص دقیق آن به یک فرایند مهم برای سازمان های اموزشی تبدیل شود.در این مقاله، DOCODE 3.0 که یک سیستم اینترنتی برای موسسات اموزشی جهت تحلیل مقادیر زیادی از اسناد دیجیتال در رابطه با درجه اصلیت است بررسی می شود. چون سرقت ادبی یک مسئله پیچیده است، سیستم ما از الگوریتم هایی برای فرایند تلفیق اطلاعات از منابع چند دادهای به همه این سطوح استفاده میکند. این الگوریتم ها به طور موفق در جامعه علمی درحل مسائلی نظیر شناسایی متن های سرقت شده و بازیابی کاندید های منبع از اینترنت استفاده شده اند.ما این الکوریتم ها را به معماری JEE جند لایه ای و قوی تلفیق کرده و به مشتریان مختلف با نیاز های مختلف امکان می دهیم تا خدمات ما را مصرف کنند. برای کاربران، DOCODE تولید گزارشاتی می کند که معلمان و پرفسور ها امکان دست یابی به اطلاعات در خصوص اصلیت اسناد را می دهد. تجربه ما مربوط به کشور شیلی با زبان اسپانیایی است که راه حل هایی را برای اسناد اموزشی شیمی را در هر یک از محیط های یادگیری مجازی مطلوب ارایه می کیند. با اینحال،DOCODE به اسانی قادر به افزایش پوشش زبان است.

کلمات کلیدی: تشخیص سرقت ادبی، ترکیب اطلاعات الگوهای متنی، منابع داده چند اسنادی

1- مقدمه

سناریوی امروز یک تغییر معنی دار را در شیوه دست یابی به اطلاعات نشان داده و بر استفاده از وب به عنوان یک منبع دانش تاکید می کند(48-49). با این حال، دسترسی به وب به عنوان یکی از منابع اصلی برای کاهش ادراک شده در صداقت تحصیلی به خصوص در رابطه با سرقت ادبی استناد شده است(44).

سرقت ادبی متشکل از استفاده از کار ها و نام های دیگران به اسم خود است. هم چنین سرقت ادبی فرایند کپی کردن نوشته های دیکران بدون استناد است. وقتی که به محیط اموزشی اعمال شود، نتایج نشان داده است که اصطلاح سرقت ادبی اغلب اشاره به وقوع سرقت ادبی مورد استفاده توسط دانشجویانی دارد که در موسسات اموزشی قرار دارند(20) و این بیانگر موارد مربوط به سرقت ادبی متنی است. در این زمینه، چو.ن حجم زیادی از اطلاعات به راحتی قابلدسترس وجود دارد، پدیده سرقت ادبی به اسانی گزارش شده است. مطالعات بین المللی بر بزرگی این رفتار تاکید کرده است که در آن درصد زیادی از دانشجویان از وب برای سرقت ادبی استفاده می کنند. نظر سنجی 2010 توسط وزارت مهندسی صنعتی دانشگاه شیلی نشان داد که 55درصد دانش اموزان راهنمایی و 42 درصد دانشجویان اطلاعات را بدون استناد سرقت کرده اند(31).

با توجه به حجم زیادی از اسناد و منابع اطلاعاتی که امروز وجود دارند، بررسی اصلیت و تشخیص سرقت ادبی به یک مسئله بسیار پیچیده تبدیل شده است. اگرچه موتور های جست وجوگر را می توان برای تشخیص سرقت ادبی استفاده کرد، با این حال فرایند شناسایی سخت و خسته کننده است(20). در سناریوی امروزه، بررسی دستی به صورت یک فرایند زمان بر و غیر ممکن می باشد. معلمان اغلب فاقد زمان کافی و لازم برای ارزیابی جامع هستند. هم چنین برخی از دانش اموزان صرف نظر از میزان ممنوعیت و منع، سرقت ادبی را انجام می دهند(22). در شیلی، نبود سیستم تشخیص مناسب سرقت ادبی بهزبان اسپانیایی موجب بدتر شدن وضعیت شده است.

سرقت ادبی یک مسئله مهم برای اهداف اموزشی در هر سطح است زیرا میتواند بر فرایند یادگیری دانش جویان اثر دارد(27). معلمان ودانشگاهیان از سرقت ادبی تنفر دارند زیرا با اهداف اموزشی تناقض دارد. در نتیجه تمایل زیادی از طرف معلمان برای حمله به این مسئله با توسعه شیوه های مختلف برای شناسایی اصلیتی کار(44) وجود دارد.بررسی بزرگی مسئله(16) نشان می دهد که بدیهی است که دانشگاهیان نیازمندابزاری برای بهبود تشخیص سرقت ادبی می باشد. این ابزار ها که اغلب موسوم به موتور های تشخیص سرقت ادبی می باشد و به این ترتیب معلمان بهتر قادر به تحلیل تعداد زیادی از اسناد میباشند.

مرور منابع مربوط به سرقت ادبی در موسسات اموزشی نشان می دهد که بسیاری از نویسندگان پیشنهاد کرده اند که این مجموعهای از رفتار های نامناسب است. برای حل این پیچیدگی، برخی از محققان از سطوح مختلف و انواع سرقت ادبی استفاده کرده اند.

از دیدگاه ما، هنگام استناد به اهداف اموزشی، موتور های تشخیص سرقت ادبی مجموعه ای از ابزار را برای کسب اطلاعات در خصوص اسناد ارزیابی شده ارزیابی می شوند. از این روی ، مطالعه ما سیستمی را ارایه می کندکه قادر به تشخیص سرقت ادبی متنی برای موسسات اموزشی با استفاده از یک دیدگاه چند سطحی می باشد. سیستم ما که موسوم به شناساگر کپی سنداست، با معلمان و اساتید همکاری کرده و یک رابط کاملی را برای ابزار ها جهت کشف،درک و مدیریت سطوح سرقت ادبی ارایه می کند.DOCODE یک سیستمی بر اساس معماری مقیاس پذیر و پیاده سازی مجموعه ای از الگوریتم ها برای تشخیص سرقت ادبی میباشد. اگرچه تجربه ما محدود به شرایط شیلی و زبان اسپانیایی می باشیم و بیشتر الگوریتم ها به اسانی قادربه افزایش پوشش زبانی هستند.

ادامه این مقاله به صورت زیر سازمان دهی شده است. بخش 2 به بررسی منابع مربوط به موضوع سرقت ادبی پرداخته و برخی از الگوریتم ها و چارچوب های پیشرفته سرقت ادبی را در اختیار می گذارد. در بخش 3 ما به توضیح این می پردازیم که چگونه Docode کار می کند و چه خدماتی را ارایه می کند. الگوریتم های اصلی مربوطبه این سیستم ارایه شده اند. بخش 4 به بررسی سازمان دهی docode، توضیح معماری می پردازد.بخش 5 به معرفی رابط های کاربری می پردازد. بخش 6 شامل نتیجه گیری و کار های اینده است.

2- کار های مربوطه

در این بخش مرور کوتاهی در خصوص سرقت ادبی از جمله مهم ترین تعاریف بیان شده توسط جامعه علمی،رویکرد های پیشرفته در تشخیص سرقت ادبی ونیز و مرور مهم ترین سیستم های شناساگرکپی ارایه می شود.

2-1 به سوی طبقه بندی سرقت ادبی

اگرچه محققان تعاریف مختلفی را از سرقت ادبی در سال های مختلف ارایه کرده اندف با این حال بدون تعمیم دادن سرقت ادبی به ایده ها، متن و غیره نیز می توان آن را بیان کرد(18). با این حال نگاهی دقیق به همه این

تعاریف امکان بررسی این را می دهد که برخی از محققان پیشنهاد کرده اند که سرقت ادبی مجموعه ای از رفتار های نامناسب می باشد. در واقع، یکی از اولین تلاش ها برای تعریف انواع سرقت ادبی در اوایلدهه 90 ارایه شده است. همین مقاله بیان می دارد که سرقت ادبی می تواند شش فرم را داشته باشد که در ادامه بحث شده است. با این حال در هر دو صورت اگر رابطه سرقت ادبی بین دو متن وجود داشته باشد این نشان می دهد که متون درجاتی از فرایند درون منتنی را نشان می دهند که به طور مستقل نوشته شده است.

از سوی دیگر(1) به بررسی مسئله تقلب دانش اموزان و تعاریف دیگر پرداخته و انواع رفتار های تقلب مربوط به سرقت ادبی را به کپی کردن سوالات امتحانی، همکاری و فریب طبقه بندی کرده است. در 33 محققان خاطر نشان کرده اند که دانش اموزان از فنون مختلف برای پنهان کردن سرقت ادبی استفاده می کنند. از نظرما طبقه بندی سرقت ادبی در درک چالش های موجود اهمیت دارد. به این ترتیب ما از این چالش برای ارزیابی شیوه مقابله DOCODE3استفاده می کنیم.اساسا ما موارد سرقت ادبی زیر را در نظر می گیریم

1- کپی کلمه به کلمه: کپی از منابع الکترونیکی از جمله سرقت نام نویسنده
2- پارافریز:افزودن، جایگزین کردن ویا حذف کاراکتر ها و کلمات. افزودن اشتباهات گرامری واملایی. جایگزین کردن کلمات با مترادف ها.
3- شیوه های فنی برای بهره گیری از ضعف سیستم: استفاده از حروف سفید و بی رنگ در سند
4- بررسی واستفاده غیر دقیق از رفرنس ها. ارایه منابع تقلبی وکاذب واستفاده از لینک های منقضی شده

باور ما این است که این مقوله ها برای تحلیل مسئله تشخیص سرقت ادبی از دیدگاه اموزشی مناسب است زیرا هر یک از آنها قادر به تشخیص و تعریف مسئله تشخیص میباشند. ما ادعا نمی کنیم که این طبقه بندی تنها صحیح است. از این روی این طبقه بندی می تواند مورد تحلیل واقع شود(33)

2-2 تشخیص خودکار سرقت ادبی

طیف وسیعی از تحقیقات در زمینه تشخیص سرقت ادبی به صورت خودکار صورت گرفته اند. با این حال در بیشتر کار های اخیر همانند مطالعه 41، سرقت ادبی به صورت استفاده مجدد از کار های فرد دیگر برای خود تعریف می شود. در این زمینه، 43 بیان میدارد که منابع مربوطه اغلب تشخیص سرقت ادبی را همانند شناسایی بخش های بسیار مشابه در متن می داند.از این روی 39 یک تعریف رسمی از سرقت ادبی را به صورت

$s = \langle s_{plg}, d_{plg}, s_{src}, d_{src} \rangle$ ارایه کرده است که متشکل از s_{plg} در سند بودهو نسخه سرقت شده از متن s_{src} در d_{src} است. در d_{plg} ، شناساگر سرقت قادر به شناسایی $r = \langle r_{plg}, d_{plg}, r_{src}, d'_{src} \rangle$ است که متشکل از متن سرقت شده r_{plg} در d_{plg} و منبع آن r_{src} در d'_{src} است .

همین محققان خاطر نشان کرده اند که دیدگاه های موجود در خصوص سرقت ادبی یک تصویر کلی را نشان نمی دهد و تشخیص سرقت ادبی را به دو بخش تقسیم می کند یعنی خارجی و داخلی. از یک سو، سرقت ادبی نشان می دهد که متن منبع برای یک سند سرقت ادبی را می توان در تشخیص سند استفاده کرد. از سوی دیگر، در سرقت ادبی درونی، شناساگر ادبی قادربه شناسایی متون سرقت شده بر اساس اطلاعات است. از نظر ما، بخش بندی درونی و بیرونی جالب است زیرا هر رویکرد شامل تعدادی از مسائل مورد استفاده برای تعریف ویژگی های ارایه شده توسط سیستم تشخیص سرقت ادبی است. به منظور توسعه اطلاعات جدید در خصوص موشوع، رقابت های سالانه تحت PAN از زمان 2009 سازمان دهی شده است. برای این رقابت ها،سازمان دهنده نیز اولین مجموعه را ارایه کرده اند که شامل سرقت ادبی متون و شاخیص های تشخیص سرقت ادبی است(3).بعد ها محققان یک چارچوب ارزیابی را برای تشخیص سرقت ادبی در(42) ارایه کرده اند که شامل نسخه ای از متون بوده وتعریف رسمی از معیار ها برای ارزیابی عملکرد شناساگر سرقت ادبی ارایه شده است. در 2012، محققان تصمیم به ایجاد مجموعه جدید گرفته اند که شامل اسناد کتبی و شبیه سازی کل فرایند سرقت ادبی است(41). علاوه بر PAN PC، امکان یافتن Clough09 وجود دارد که متشکل از 57 پاسخ کوتاه به یکی از 5 سوال علمی است.

در نتیجه رقابت های PAN و به دلیل علاقه جامعه علمی به مسئله سرقت ادبی، چندین روش برای تشخیص انواع سرقت ادبی مختلف ارایه شده است. بسته به نوع سرقتادبی مورد استفاده،مجموعه مهمی از رویکرد ها بر اساس ویژگی های مختلف متن را می توان استفاده کرد(12). منابع موجود در خصوص هر موضوع گسترده هستند و برخی محققان به بررسی رویکرد ها در تشخیص خودکار سرقت ادبی پرداخته اند. در این جا ما به بررسی مهم ترین مطالعات 14-2-17-28 می پردازیم.

3-2 سیستم های تشخیص سرقت ادبی

علاقه به تشخیص سرقت ادبی خودکار تنهاشامل دانشگاهاین نیست. امروزه چندین ابزار تجاری در بازار موجود است.هر ابزار دارای رویگرد خاص خود است. سیستم های تشخیص سرقت ادبی را می توان به هرمتیک و وب تقسیم کرد. سیستم هایتشخیص وب سعی میکنند تاانطباق هایی را در اسناد مشکوک درمنابع انلاین پیدا کنند در حالی که سیستم های هرمتیک به دنبال تمونه های سرقت ادبی از مجموعه ای از اسناد هستند(33). رخی از سیستم های موجود نظیر Turniton نیز خدماتی را ارایه می کند با این حال سایرین را می توان دانلود کرد یک توصیف کوتاه در زیر ارایه شده است

- تارنیتون: یک شرکت تجاری که خدماتی را برای تشخیص سرقت ادبی ارایه میکند. کار های دانشجویان را بر اساس دیتابیس های به روز کنترل می کند. امروزه دارای بیش از یک میلیون مقاله است و بیش از 12 میلیون صفحه اینترنتی وجود دارد
- Eve 2» یکابزار تجاری که وب را برای منابع مشکوگ جست وجو می کند. URL را استفاده می کند و مقاله را به استاد گزارش می کند
- PlagiarismDetect.com : یک ابزار تجاری برای جست و جودر وب بوده و مشابه با Eve 2 است
- سرویس سرقت ادبی گلت: سه قطعه از نرم افزار راارایه می کند. اولین مورد برنامه اموزشی برای کمک به اموزش دانشجو.یان در موردسرقت ادبی است.دومین مورد یک برنامه فیلترینگ برای اجتناب است در حالیکه سومین مورد یک برنامه غربال گری برای تشحیص سرقت ادبی است
- Ephorus7: دیگر ابزار تجاری برای تشخیص سرقت ادبی است که شامل پشتیبانی هایی برای کنترل استنادات صحیح است که امکان مکان یابی منابع را در زمان نوشتن مقاله می دهد.
- WCopyfind8 :یک برنامه ویندوز متن باز که اسناد را مقایسه کرده و تشابهات را در عبارات گزارش می کنند. این خود تحت راهنمای GNU قرار دارد

تا کنون چندین مقاله به بررسی ابزارها و مقایسه آن ها تحت معیار های مختلف پرداخته اند از جمله 23-28-10-20-22. بیشتر ابزار های موفق جزییاتی را در خصوص الگوریتم ها ارایه نمی کند و به صورت جعبه سیاه کار میکند.از دیدگاه کاربران، این کمبود اطلاعات بیانگر یک مانع مهم در درک شیوه نشان دادن نتایج است. دراین رابطه،Dococe متفاوت است زیرا الگوریتم ها و برنامه ها با جامعه علمی نشان داده می شود.

3- سیستم پیشنهادی

در این دو بخش ما به توصیف کاملی از سییستم خود می پردازیم. در اولین مورد، جزییاتی در خصوص الگوریتم های پشتیبانی کننده عملکرد ها ارایه می کنیم. سپسفما توضیح می دهیم که چگونه DOCOODE سازمان دهی می شود. در نهایت ما به بررسی رابط های کاربری در این خصوص می پردازیم. این سازمان دهی به این واقعیت پاسخ می دهد که سیستم ما سه نقش مهم را دارد

- توسعه و پیاده سازی چندین الگوریتم برای بررسی ومقابله با همه ابعاد مسئله سرقت ادبی . برخی از این الکوریتم ها بهتر از رویکرد های پیشرفته هستند
- ارایه یک معماری مقیاس پذیر، کارامد و قوی برای پشتیبانی از عملکرد ها و ارایه خدمات با کیفیت بالا.
- طراحی و پیاده سازی رابط های کاربری با مجموعه ای از منشور ها و طرح ها و ابزار های پشتیبانی از فرایند تصمیمگیری در رابطه با سرقت ادبی. سومین هدف مربوط به اثر متقابل و تعامل انسان و کامپیوتر است

3-1 شناساگرئ کپی در اسناد:FASTDOCODE

اولین مورد ارایه شده توسط DOCOODE براساس الکوریتم شناساگر کپی می باشد که موسوم به FASTDOCODE است. همان طورکه در 34 گفته شد، FASTDOCODE بر اساس دو فاز اصلی به کار برده شده پس از مرحله پیش پردازش در هر مورد است. عموما الگوریتم ما فضای جست و جو را با استفاده از جست وجوی بخش ها بر اساس طرح پیشنهادی 4 بررسی می کند و به این ترتیب از الگوریتم جست وجو در اسناد بهره می برد

در زیر اگر V بردار کلمات تعریف کننده واژگان باشد.کلمه با w تعریف شده و واحد اساسی داده های گسسته است که با $\{1,\ldots,|V|\}$ نماینده می شود. سند d به صورت ترتیب کلمات p می باشد که با $d=(w_1,\ldots,w_p)$ تعریف می شود و wo بیانگر pمین کلماتدر سند است که با جمع اوریnسند نشان داده شده با $D=(d_1,\ldots,d_n).$ تعریف می شود

با توجه به مجموعهd حاوی مجموعه ای از اسناد، اولین مرحله مربوط به روش کاهش فضای جست و جو است که هدف آن شناسایی جفت اسناد در مجموعه ای است که دارای متن های مشترک می باشد. این مشابه با راهبرد نمونه گیری برای هر بخش است و موجب کاهش تعداد مقایسه های انجام شده و بهبود زمان اجرای

الگوریتم می شود(34). در نهایت، راهبرد روش تحلیل را بر اساس کلمه 4-گرم در نظر میگیرد و به این ترتیب اسناد به صورت مشکوک در نظر گرفته می شود.

دومین مرحله ، جست وجوی جامع برای یافتن پیام های سرقت شده در اسنادی است که مشکوک است. بر عکس، برای این مرحله، کلمات توقف حذف نمی شوند و کلمه 3 گرم استفاده می شود. اساسا، تقاطع بین 3 گرم بخش مختلف محاسبه می شود. پس از پایان مقایسه دو به دوی گروه ها، شاخص تشابه نهایی بازگشته می شود. مقادیر پارامتر های L وr ونیز جزییات الگوریتم در این جا اشکار نمی شوند.

FASTDOCODE در طیPAN2010وpan2011 در تشخیص سرقت ادبی تست شده و رتبه سوم وپنجم را در میان رقبا کسب کرده است.پارامتر های الگوریتم ها از جمله تعداد N گرم برای استفاده برای دو فاز تعدیل شده اند. با این حال با استفاده از تحلیل گسترده بر رو ی عملکرد الکوریتم بررسی شده است. PAN PC 2010 الگوریتمی است که قادر به دست یابی به یک دقت است و با استفاده از PAN PC 2011 برای فراخوانی و 91.17 در نظرگرفته شده است.

2-3 تغییر شناساگر سبک نوشتاری

هدف دومین عملکرد استخراج شواهد سرقت ادبی از یک سند مشکوک است. این رویکرد تلاش می کند تا متون مشکوک را در خصوص استفاده از راهبرد تشخیص سرقت ادبی درونی پیدا کند

الگوریتم ما بر اساس شناسایی سبک نوشتاری نویسنده است. اگر برخی لغات استفاده شدهدر سند مختص نویسنده باشند، این کلمات در پاراگراف هایی متمرکز می شوند که قبلا ذکر شده اند. به طور مشابه، 36 یک مدلی را برای کمی سازی سبک نوشتاری ارایه کرده است که هدف آن یافتن انحرافات در سبک است. یک رویکرد به شرح زیر است:اول، سند به کاراکتر های الفبایی پیش پردازش شده و به این ترتیب در طبقه های مختلف قرار می کیرد. بدون حذف کلمات توقف، یونیگرامها استخراج شده و تعدادی از بخش های سند ها با عبارت M تشکیل میشوند. رد پا یا سبک عمومی سند با میانگین همه اختلافات در کلمات نمایش داده می شوند. از این روی اگر کلمات خاصی در هر بخش استفاده شود، مقایسه این بخش با کل سند منجر به مقدار پایین در تفاوت بین سبک ها می شودزیرا فراوانی این کلمات در کل سند یکسان است. اگر تغییر معنی دار باشد، سبک کم تر از مقدار متوسط منهای استانه از پیش تعریف شده است. در نهایت همه بخش ها بر طبق

فاصله با توجه به مقدار سبک سند طبقه بندیمی شود.اگر سبک بخشکم تر از مقدار سبک منهای استانه باشد این بخش به صورت مشکوک طبقه بندی می شود.

در طی رقابت های PAN بر روی تشخیص سرقت ادبی، الگوریتم با استفاده از PAN-PC-2010 و 2011 تست شد. برای پارامتر های بیان شده در 35، 400کلمه و پارامتر استانه0.075 استفاده شد. این موارد بسته به طول متن تعدیل شد.در اولین مورد، عملکرد 38ک97 درصد برای دقت و 31.91 درصد برای فراخوانی بود. برای مجموعه 2011، دقت و فراخوانی 33.98 درصد و 31.23 درصد برای امتیاز کل 32.54 درصد بدست امد. در مقایسه با رویکرد های پیشرفته، الکوریتم به نتایج مطلوبی دست یافت و اولین مورد در PAN 2011مشاهده شد. لازم به ذکر است که این روش از ویژگی هایوابسته به زبان نظیر کلمات استفاده می کند.جزییات بیشتر در خصوص این نتایج در مقاله ها ارایه شده است(40-25-36).

3-3 شناساگر متن پنهان

همان طور که در بخش 2 گفته شد برخی از روش های سرقت ادبی شامل استفاده از فنونی است که از نقطه ضعف مختلف سیستم های تشخیص سرقت ادبی استفاده میکند. در این جا ما از کلمات سفید رنگ به صورت جایگزینی برای فضا های خالی استفاده می کنیم. بر اساس تجربه در شیلی، این رفتار یکی از رایج ترین راهبرد های مورد استفاده توسط دانشجویانی هستند که تلاش می کنند تا سرقت ادبی را پنهان می کنند.

رویکرد ما در این زمینه ، الگوریتمی است که کنترل می کند که آیا طول متوسط هرجمله در دامنه نرمال قابل قبول است یا خیر. پارامتر های استانه بر اساس قواعد زبانی واسنجی شده اند. در صورتی که الگوی غیر طبیعی شناسایی شود، الگوریتم یک مقداری را بدست می دهد که برای ارایه هشدار به کاربر استفاده می شود.

3-4 بازیاب سند وب مشابه

ویژگی اساسی دیگر ارایه شده توسط DOCODE یک بازیاب سند وب مشابه می باشد. با توجه به سند مشکوک Dو مجموعه D از اسنادی که نویسنده آن ها سرقت ادبی کرده است، نخستین گام بازیابی تعدادی از اسناد کاندید $D_x \subseteq D$ است که منبع سرقت ادبی محسوب می شود. این حود در نظر می گیرد که D بسیار بزرگ است(45).

بر اساس سند مشهود، الگوریتم به بررسی مسئله دست یابی به اسناد مشابه از وب با موتور های جست و جوگرمی پردازد. از این روی ما این مسئله را از دیدگاه بازیابی اطلاعات در نظر می گیریم. اگرچه منابع معمولا پیشنهاد می کندکه کوئری های جست و جو را می توان به سه مقوله طبقه بندی کرد یعنی اطلاعاتی، هدایتی و تراکنشی. در(8)، محققان اینمسئله را مسئله بازیابی تشابه سند وب دانسته اند. تفاوت کلیدی بین مقوله های کوئری و WDSRP این است که ورودی سند را به جای پرس و جوی مبتنی بر متن در نظر می گیرد.

همان طور که در 24 گفته شده است، با توجه به سند D، که Vرا می توان از آن استخراج کرد، مدل زبان MD از d یک تابعی است که شاخص احتمال را از v اندازه گیری می کند. مدل های زبانی به عنوان توابع رتبه بندی در بازیابی اطلاعات استفاده شده و احتمال ایجاد یک کوئریq را با توجه به مدل زبان براورد می کند.در این جا فرض اصلی این است که اطلاعات جدیدی به موتور های جست و جوگر ارسال می شود و این موضوع را در نظر می گیرد که موتور های جست و جوگر امکان طول ماکزیمم کوئری های ورودی را می دهد. این فرایند با استخراج واژگان از d شروع می شود(36). سپس، امکان ایجاد کوئری به منظور کوئری ها با محدود سازی استخراج ها بدون جایگزینی اصطلاحات استخراج شده وجود دارد. طول پرس و جو ها یا سوالات قابل تغییر است. علاوه بر این الگوریتم، یک سیستم پرس و جویدیگر برای استخراج نمونه ای از N- گرم ارایه میشود. ازاین روی این الکوریتم از رویکرد انگشت نگاری استفاده میکند.

در نهایت ما یک الگوریتم را برای براورد تشابه بین سند و هر سند بازیابی شده از وب پیشنهاد می کنیم. رویکرد ما ترکیبی از دو ویژگی است. اولین ویژگی بر اساس تابع توزیع Zipf در محتویاسناد بازیای شده،مدل سازی اهمیت پاسخ موتور جست وجو گر کوئری به عنوان توزیع zipf است. از این روی، اهمیت نتایج ارایه شده در موتور های جست و جوگر ارتباط معکوسی با رتبه بندی آن ها دارد. در این رابطه، ما اهمیت پاسخ سوالات را با ترکیب رتبه بندی و پایداری نتایج موتور های جست و جوگر بررسی می کنیم. دومین ویژگی ناشی از ترکیب عنوان و خلاصه می باشد. سپس، ما از دو ویژگی برای پیش بینی تشابه استفاده می کنیم با فرض این که آن ها ارتباط نزدیکی با تشابه بین d و سند وب داشته باشند. مدل ما با استفاده از روش هایی نظیر شبکه های عصبی مصنوعی و یا سایر فنون رگرسیون مربوطه برازش می یابد. راهبرد ما شامل اندازه گیری کارایی مدل برای رفع نیاز های اطلاعاتی مربوط به wdsrp است. ما ابتدا یک مجموعه دستی از 160 پاراگراف را تولید کرده ایم که

از سایت های مختلف اینترنتی انتخاب شده است. سپس پاراگراف ها به صورت ورودی به سیستم ارسال شده و 15 پاسخ از هر پاراگراف مرور و طبقه بندی شد.جدول 1 عملکرد را نشان می دهد.

همان طور که دیدیم نتایج نشان داد که طرح پیشنهادی ما قادر به حل مسئله بازیابی تشابه اسناد است. هم چنین آن ها نشان داده اند که مدل جست وجو به طورمعنی داری موجب بهبود ظرفیت بازیابی نسبت به نتایج موتورجست و جوگر می شود. الگورتیم ما مرتبط با سه موتور جست و جو گر است.

5-3 شناساگر اقتباسی برای زبان اسپانیایی

یک ویژگی دیگر ارایه شده در سیستم، شناساگر منابع کتاب شناسی است. در این رابطه، اگرچه ما الگوریتمی را برای زبان اسپانیایی طراحی کرده ایم، و رویکرد های مشابه را برای زبان های دیگر در نظر گرفته ایم.بر طبق 32، نقل قول استعمال محتوی غیر اصلی نویسنده است، در این صورت اگرچه یک الکوریتم برای زبان اسپانیایی ارایه شده است رویکرد های مشابه نیز مطلوب هستند.

سازمان دهی مبتنی بر عبارت نحوی: اشاره به بخشهایی از جمله با توابع نحوی متناظر دارد. در این مقوله امکان یافتن الگوهای تیپوگرافی وجود دارد. این الگو ها شامل نام نویسنده، فعل و خود نقل قول است.

جدول 1: دقت مربوط به اسناد مربوطه بازیابی شده در نتایج برتر

k	1	2	3	4	5
Precision (%)	86.9	70.9	60.6	53.0	46.9

امکان تمایز دو مقوله وجود دارد 1- نقل قول های مستقیم که در آن عبارت و کلمات نویسنده مرجوع به طور متنی استفاده می شود. 2- نقل قول های غیر مستقیم که عباراتی هستند که اشاره شده اند

طبقه بندی مبتنی بر گفتمان: در این صورت،نقل قول، نسبت به سطح اهمیتی که نویسنده می دهدتجزیه تحلیل می شود. این به صورت استفاده یاعدم استفاده از نقل قبول می باشد. این مقوله ها تعریف شده اند:1- نقل قول های اصلی 2-نقل قول های غیر اصلی

در(46)، دستورالعمل سبک های نوشتاری مختلف برای نویسنده، سردبیران و ناشران ارایه میشود. از اینروی فرمت های مختلفی برای نوشتن نقل قبول ها وحود دارد کهبا در نظر گرفتن مقوله های فوق گروه بندی شده است. از این روی مقوله های خاص برای شناسایی هر نقل قول به شکلی دقیق توسعه یافته است. این راهبرد ها

شامل استفاده از عبارات منظماست. 13عبارت مختلف برای پوشش دادن همه دستور العمل ها در نظر گرفته شده است. در این میان از الگوهای خارج از دستور العمل استفاده می شود.

1- استخراج متن سند

2- علامت گذاری متن با مختصاتی که در آن عبارات به طور منظم تشخیص داده می شوند.

3- جست وجوی الگوها و مقایسه آن ها با مقادیر از پیش تعیین شده

4- در صورتی که یک الگوی شناسایی شده وجود دارد، بدیهی است که نقل قول وجود دارد

به منظور آزمون اثر بخشی رویکرد، ما به طور دستی مجموعه ای از نقل قول ها را با استفاده از تز های کارشناسی دانشگاه زراعت و مهندسی شیمی تعیین می کنیم.مااقدام به مرور این تزها و استخراج250 و530 نقل قول کردیم. هدف اولین ازمایش،ارزیابی الگوریتم شناسایی نقل قول تحت شرایط کنترل شده می باشد. د برای انجام این کار، 484 کتاب از حوزه عمومی و 5800 عبارت انتخاب شد. ایده ما شبیه سازی های مختلف را در نظر می گیریم. دومین ازمایش ارزیابی الگوریتم در شرایط واقعی است. برای انجام این کار، 20 تز زراعت بررسی شد. اسناد دارای 15169 جمله می باشد و از این روی 536 مورد به صورت نقل قول در نظر گرفته شده است. لازم به ذکر است که مجموعه ما یک مجموعه داده غیرمتعادل می باشد و ظاهرا از اهمیت زیادی برخوردار است. یک بررسی دقیق نتایج نشان داد که با توجه به این که شناساگر نقل قول بر اساس عبارات منظمی است که شامل استفاده از الگوهای مختلف می باشد،بسیاری از موارد نادر تنها بخش هایی از متن هستند که به آسانی قابل تشخیص می باشند. اگرچه الگوریتم قادر به فعال سازی بسیاری از موارد متنی نیست، با این حال دقت بایستی با استفاده از الگوریتم های اضافی برای فیلتر کردن نتایج بهبود یابد. در این رابطه،نتایج ما موثر و امیدوار کننده است.

6-3 انالیزور موضوعی چند اسنادی

ویژگی های دیگر ارایه شده توسط DOCODE یک انالیزور موضوعی از مجموعه زیادی از اسناد می باشد که امکان ترسیم نمودار موضوعی اسناد را می دهد. به عبارت دیگر، تحلیل موضوعی به تعیین این که آیا اسناد تحلیل شده دراای محتوای مشابه است یا خیر کمک می کند. این روش به تحلیل روابط بین یک مجموعه ای

از اسناد واصطلاحات برای ایجاد مفاهیم جدید می پردازد. برای انجام این کار، این الکوریتم فرض می کند که از نظر معنی مشابه با قطعات متن است.

رویکرد ما از مطالعه(11) الهام گرفته است که درآن محققان روش ها و فنون تشخیص سرقت ادبی را بر اساس LSAپیشنهاد می کنند. به بیان ساده تر، ما ماتریس Tf-idf را برای اسناد محاسبه می کنیم و این موضوع را در نظر می گیریم که همه پارامترها قابل تحلیل هستند. svd شامل مدل جدیدی از فضای ویژگی است که در آن رابطه معنایی بین این دو در نظر گرفته شده است. وقتی که از svd در همه ماتریس ها استفاده شد،مجموعه ای از ماتریس های اسناد بعدی را میتوان داشت. چون ستون های M مدل های فضای مفهومی از هر سند می باشند، تشابه بین اسناد و ستون ها را میتوان با استفاده از شاخص های مختلف محاسبه کرد.

7-3 بازیاب سند دیتابیس داخلی

سیستم ما قادر به ذخیره سازی همه اسناد پردازش شده در گذشته است وهمه اسناد وب توسط بازیاب وب در دیتابیس درونی دانلود شده اند. با توجه به سند منبع، الگوریتم بازیاب دیتابیس بر اساس اپاچی بوده و در بر گیرنده مجموعه ای از اسناد مشابه است. لوسن یک مدل رتبه بندی بر اساس مدل فضای بردار کلاسیک و مدل بولین از بازیابی اطلاعاتی است که هدف آن انتخاب اسناد مشابه است.

4-معماری

DOCODE خدمات مختلف را به طیف وسیعی از موسسات و افراد ارایه می کند. بیشتر کارکرد های آن شامل سرویس های وب هستند. ر اساس برنامه SAS، برنامه ها توسط کاربر با استفاده از یک مشتری قابل دسترس هستند. در این مورد DOCODE از طریق رابط های برنامه نویسی ارایه می شود که بر اساس دو پروتوکل هستند. در همین رابطه، کلاینت ها قادر به مصرف خدمات ارایه شده توسط سیستم هستند

جزییات بیشتر در مورد لایه ها در زیر نشان داده شده اند

- لایه کلاینت: که در آن کاربر به برنامه دسترسی دارد و خدمات را مصرف می کند. همانطور که گفته شد سیستم ما تنها تیازمند یک کلاینت سرویس و یا بروزر وب است. برای کلاینت وب سرویس هر درخواست به سرویس ارایه می شود که با یک ادرس اینترنتی همراهاست.

- لایه وب: متناظر باپیادهسازی رابط وب متفاوت است که به کاربرامکان مصرف خدمات ارایه شده توسط سیستم و تعادل بانتایج را می دهد. جزییات رابط موجود در بخش 5 ارایه شده است
- لایه سرویس: این لایه مسئول ارایه خدمات وب برای شبکه خارجیاست. اولین گام تعریف اشیایی است که بخشی از یک ارتباط هستند. در این مورد، همانطور که در پارامتر JEE قرار دارد، فایل های XSD XML) برای تعریف ساختارهایی که در پیام های SOAP ارسال می شوند استفاده می شوند. سپس ما فایل) WSDL) را که شامل تمام ساختارهای تعریف شده در فایل XSD و همچنین روشهایی است که برای اجرای آن در دسترس است، تولید می کنیم. گام بعدی اجرای این سرویس HSJ. ما تصمیم گرفتیم از API جاوا برای خدمات وب XML (JAX-WS) استفاده کنیم، زیرا این API در پلاتفورم JEE ما متن باز است
- لایه منطق کسب وکار: این لایه اقدامات و فرآیندهای را ذخیره می کند که حاوی قوانین کسب و کار به عملیات صحیح DOCODE است. قوانین منطق کسب و کار نیازهای تجاری را برآورده می کند و تجزیه و تحلیل آنها از طریق الگوریتم های مختلف شرح داده شده در بخش 3، در نهایت نتایج نشان داده شده است. تجارت مدل فرایند طراحی شده است تا یک عکس بزرگ از کامل داشته باشد فرایند کسب و کار [6]. به این ترتیب، امکان مدل سازی کل سیستم با استفاده از BPMN (مدیریت فرآیند کسب و کار نشانه گذاری) وجود دارد

لایه مشتری
لایه وب
لایه خدمات
لایه منطق کسب و کار
لایه پایداری
لایه داده

شکل 1: نمودار لایه سیستم

- لایه داده ها: این لایه یک مولفه مهم بوده و بخش اصلی راه حل است. از این روی درنظر گرفتن مسائل مربوط به افزودگی داده ها، استفاده مجدد از داده ها، کنترل دسترسی مهم است. دراین صورت، لایه داده ها مستقل از توسعه پروژه است. هر دو ساختار امکان مدیریت و پردازش مجموعه بزرگی از اسناد را می دهد.

ما قبلا از کلاینت هایی برای برخی از محیط های یادگیری مجازی مهم استفاده کرده ایم که امکان یکپارچه سازی مودل، ساکای و نیز U-Cursor را می هد. چون بسیاری از کاربران مختلف خدمات ارایه شده را مصرف می کنند، امنیت یک مسئله مهم برای سیستم ما است. ما بایستی الگوریتم های خود را از دسترسی خارجی محافظت کنیم که امکان دست یابی به داده های مختلف را می دهد. از سوی دیگر کپی اسناد ذخیره شده در سرور نیز می تواند مورد استفاده قرار گیرد

هسته DOCODE یک موتور اصلی است که همه الگوریتم های پشتیبانی کننده را پیاده سازی می کند. هر یک از این الگوریتم ها به عنوان بخشی از یک کتابخانه محصور شده اند. سپس،

ما یک الگوریتم encapsulated را با استفاده از مشخصات EJB3 یکپارچه کردیم. هر زمان درخواستی به DOCODE CORE می رسد، رابط کاربری مسئولیت فراخوانی الگوریتم های مربوط به آن را دارد درخواست، با توجه به آنچه در MDB موضوع موضوع تعریف شده است. مانند شکل 2 نشان می دهد، رابط در حالت همگام در دسترس است که به طور مستقیم از EJB، و همچنین در یک حالت ناهمزمان که در آن یک پیام در صف الگوریتم درونی داخلی برای بعد باقی می ماند . پارادایم انعطاف پذیر مبتنی بر EJB به سیستم اجازه می دهد الگوریتم های مختلف را به صورت موازی اجرا کنید، زمان پاسخ را کاهش دهید. آی تی همچنین باعث می شود که الگوریتم های موجود را اصلاح و اضافه کنید.

برای پیاده سازی، ما بهترین شیوه های استاندارد JEE را استفاده می کنیم. سرقت ادبی اشاره به فرایند ارایه کلمات، افکار و ایده های افراد دیگر به صورت کلمات، افکار و ایده های خود بدون رفرنس دادن به منابع آن ها دارد. رشد ازمایشی منابع اسناد دیجیتالی مختلف موجود در اینترنت موجب تسهیل توسعه این عمل شده و در نهایت موجب شده است تا تشخیص دقیق آن به یک فرایند مهم برای سازمان های اموزشی تبدیل شود.در این مقاله، DOCODE 3.0 که یک سیستم اینترنتی برای موسسات اموزشی جهت تحلیل مقادیر زیادی از اسناد دیجیتال در رابطه با درجه اصلیت است بررسی می شود. چون سرقت ادبی یک مسئله پیچیده است، سیستم ما از الگوریتم هایی برای فرایند تلفیق اطلاعات از منابع چند دادهای به همه این سطوح استفاده میکند. این الگوریتم ها به طور موفق در جامعه علمی درحل مسائلی نظیر شناسایی متن های سرقت شده و بازیابی کاندید های منبع از اینترنت استفاده شده اند.ما این الکوریتم ها را به معماری JEE جند لایه ای و قوی تلفیق کرده و به

مشتریان مختلف با نیاز های مختلف امکان می دهیم تا خدمات ما را مصرف کنند. برای کاربران، DOCODE تولید گزارشاتی می کند که معلمان و پرفسور ها امکان دست یابی به اطلاعات در خصوص اصلیت اسناد را می دهد. تجربه ما مربوط به کشور شیلی با زبان اسپانیایی است که راه حل هایی را برای اسناد اموزشی شیمی را در هر یک از محیط های یادگیری مجازی مطلوب ارایه می کیند. با اینحال،DOCODE به اسانی قادر به افزایش پوشش زبان است.

رابط کاربری

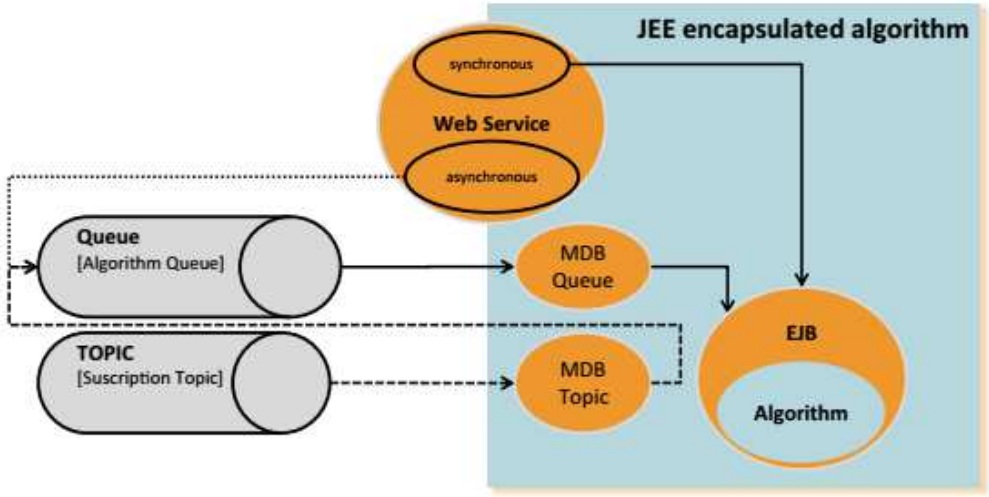
در این بخش، ما نشان می دهیم که چگونه رابط های کاربری DOCODE طراحی می شوند. به طور کلی، خدمات ارایه شده توسط همه مشتریان یا کلاینت ها یکسان است. با این حال به دلیل ویژگی های هر کلاینت، رابط انسانی این ویژگیها تفاوت اندکی رانشان داده اند.همان طور که دربخش 3 گفته شد، کلاینت ها برای ساکای،مودل و U-SCRE در نظر گرفته می شوند. موسسات با استفاده ازاین پلتفرم ها به همه سرویس ها دسترسی دارند. با این حال، چون ما قادر به محدود سازی تنها این ابزار ها نمی باشیم، DOCODE ASP را محدود کرده ایم. چون همه این ویژگی ها را بر اساس مودل VLE پیاده می کنیم DOCODE ASP تنها پردازنده ای است که از هر نرم افزار برای فعالیت های خود پشتیبانی می کند.

قبل از معرفی ویژگیهای رابط کاربری، برخی مفاهیم مربوط به سرقت ادبی تعریف وجود دارد و به تعریف مفاهیم کمک می کند. یک سازمان اموزشی را در نظر بگیرید که از نرم افزار VLE استفاده می کند. امکان شتاسایی تشابهات خاص در شیوه سازمان دهی و جوددارد. بر این اساس فایل حاوی پاسخ ها به صورت سند تحویل داده شده در نظر گرفته می شود. در نهایت ما مجموعه را به صورت مجموعه ای از اسناد در نظر می گیریم که برای هر تکلیف خانگی در نظر گرفته می شود.DOCODE بر اساس دوره های اموزشی، تکالیف خانگی و یا مجموعه ها کار میکند که شامل یک یا چند سند است.

برای اسناد درون یک مجموعه

نتایج برای هر سند درون مجموعه ابتدا به صورت عمومی نمایش داده می شود که تصویر بزرگی از اسناد را در مجموعه نشان میدهد. با کلیک بر روی صفحه، کاربران قادر به دسترسی به نتایج هستند. شرایط اولیه به صورت تعاملی در نظر گرفته می شود. در زیر ما به بررسی نتایج اسناد موجود در هر دو صفحه می پردازیم.

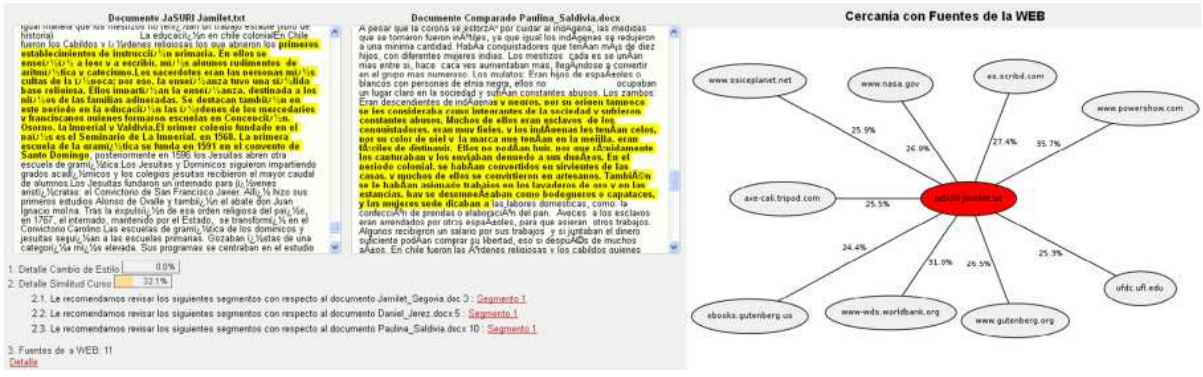
- تغییر سبک: در این صفحه عمومی،ما یک شاخصی را ارایه می کنیم که مقدار نسبی متون را در هر سند ارزیابی ارایه می کند. در نتایج دقیق،ما فهرستی را با متونی ارایه می کنیم که بر اساس همه الگوریتم ها هستند. با کلیک برروی هر صفحه برروی لیست، سیستم متون انتخاب شده را در اسناد برجسته می کند. این ویژگی بر اساس تغییر الگوریتم شناساگر سبک نگارشی است
- زنگ هشدار پیام مخفی: در صفحه کلی شاخصی را نشان می دهیم که زمانی روشن می شود که رفتار فریبانه را در سند شناسایی میکند. این ویژگی از شناساگر متن استفاده می کند
- شناسایی منبع وب: با استفاده از یک الگوریتم بازخورد سند وب سایت ما، منابع وب سایت مربوط به هر سند پردازش شده را جمع آوری می کنیم و آنها را شمارش می کنیم، و تعداد نتیجه را در صفحه کلی ارائه می کنیم. در صفحه دقیق یک سند خاص، لیستی از URL های هر وب سایت شناسایی شده را ارائه می دهیم. با کلیک کردن بر روی هر URL، گذرگاه های مشکوک مربوطه در سند تعاملی برجسته می شوند. علاوه بر این، همانطور که در شکل 4 نشان داده شده است، ما یک نمودار نشان می دهد که تمام منابع وب و میزان تشابه با سند نشان داده شده است.
- تشابه دوره اموزشی: در صفحه عمومی، شاخص شیوه استفاده از هر سند برای مجموعه نشان داده شده است. بر روی صفحه نتایج، فهرستی از همه اسناد ارایه شده است. با کلیک کردن، ما فهرستی از متون را ارایه می کنیم. در عین حال سیستم قادر به برجسته سازی زمان و نیز اسناد موجود در سرقت ادبی است که در شکل 4 نشان داده شده است. همه نتایج توسط الگوریتم شناساگر سند استفاده شده اند
- نقل قول های استخراج شده: در صفحه جزییات برای سندف فهرستی با همه نقل قول ها ارایه می شود. با این ویژگی، معلمان باید چک کنند که ایا دانش اموز رفرنس را ارایه کرده است یا خیر. لازم به ذکر است که کتاب شناسی برای تکالیف در نظر گرفته شده است .



شکل 2: ساختار الگوریتم

Nombre del Documento	Cambio de Estilo	Similitud Curso	Similitud Web	Opciones
JaSURI Jamilet.txt	0.0%	32.1%	11	Informe
test3.txt	0.0%	51.6%	11	Informe
Jamilet_Segovia.doc	0.0%	99.4%	11	Informe
Danilo_Gonzalez.docx	0.0%	45.0%	11	Informe
Daniel_Jerez.docx	0.0%	53.3%	11	Informe
test2.txt	0.0%	57.8%	11	Informe
Cristobal_Morales.docx	0.0%	68.7%	11	Informe
Valentina_Vega.docx	0.0%	0.0%	11	Informe
test1.doc	0.0%	0.0%	11	Informe
Paulina_Saldivia.docx	0.0%	34.7%	11	Informe

شکل 3: صفحه نشان دهنده نتایج عمومی برای مجموعه پردازش شده

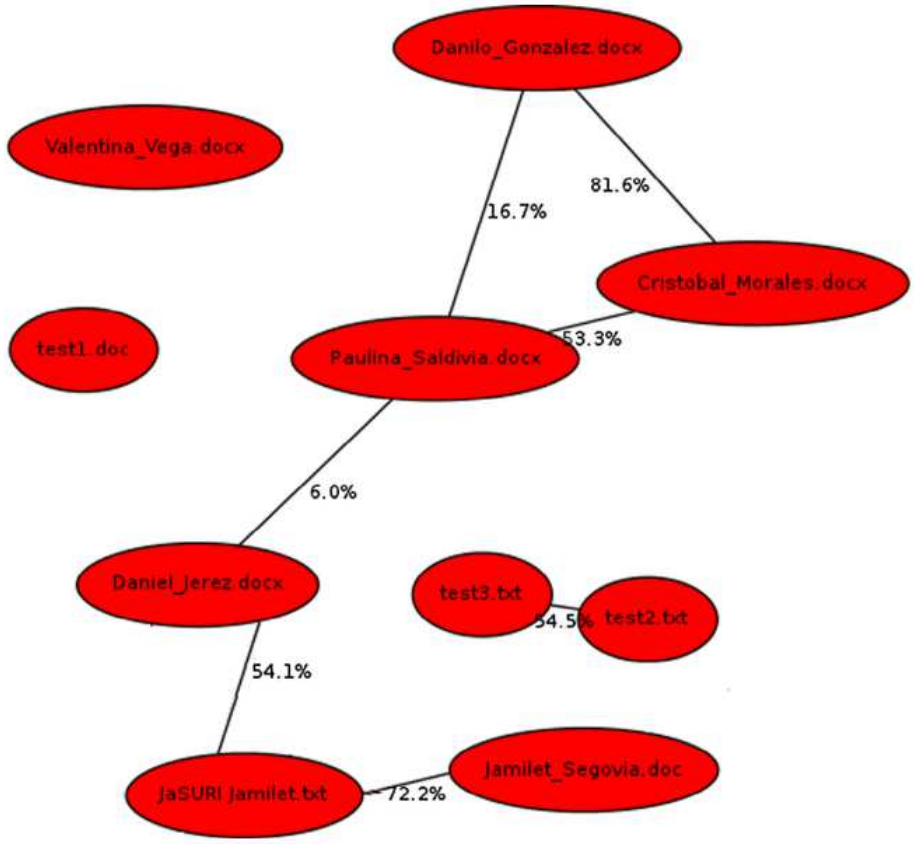


شکل 4: اسناد تعاملی نشان دهنده صفحات مشکوک در اسناد منبع و اصلی

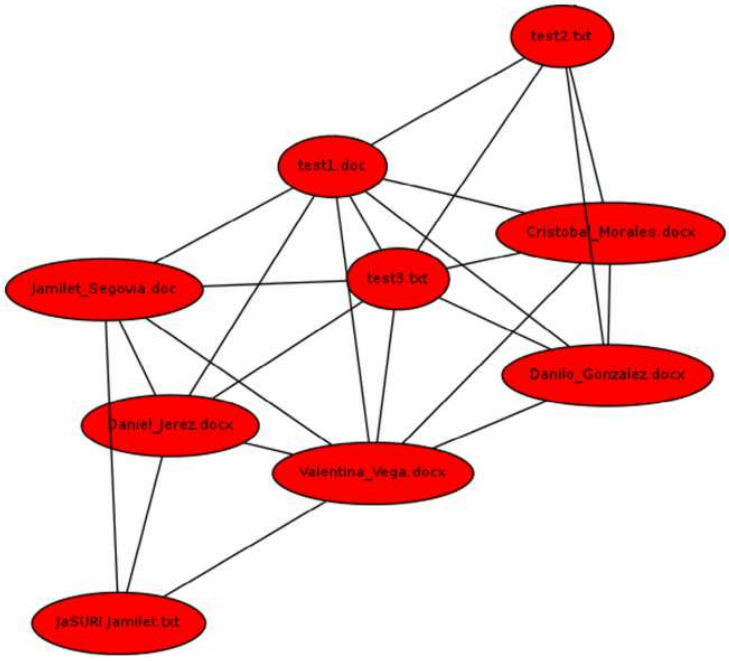
2-5 برای همه اسناد درون مجموعه

علاوهبر نتایج مربوط به اسناد، نتایچ پردازش همه اسناد درون مجموعه در نظر گرتفه می شود. خروجی تحلیل، نمودار های متفاوت برای خلاصه سازی وضعیت کلی اسناد است. این نمودار ها در صفحه عمومی نشان داده شده است

- تشابه دوره: با استفاده از FASTCODE، ما اسناد درون مجموعه را مقایسه کرده و این نمودار نتایج را به خوبی با درجه تشابه نشان میدهد. درجه تشابه بر اساس تعداد متون است که درصدی از طول کل اسناد است
- تحلیل موضوعی: این نمودار، بر اساس نتایج الکوریتم انالیزور موضوعی چند سندی برای نشان دادن تشابه در موضوعات بین اسناد است.



شکل 5: نمودار تشابه دوره که روابط میان اسناد مجموعه را نشان می دهد



شکل 6: نمودار تحلیل موضوعی نشان دهنده روابط بین اسناد

با این حال چون اسناد ارایه سرقت ادبی دارای متون مشترکی هستند، آن ها تعداد زیادی از موضوعات را نشان می دهد. از این روی با تشخیص گروهی از اسناد که دارای رفتار های نامنظم هستند، امکان کشف موارد مربوط به سرقت ادبی وجود دارد.

-تحلیل وب منبع: بر اساس نتایج بازیاب سند وب مشابه، منابع مختلف را انتخاب کرده و نموداری را برای نشان دادن آن ها ایجاد می کنیم. این شکل بسیار مشابه بانموداری است که منابع وب را برای هر سند نشان می دهد

DOCODE 3-5

در نهایت، ما DOCODE را ارایه می کنیم که در آن سرویس رایگان قادر به تحلیل اسناد با مقایسه محتویبا منابع وب مختلف می باشد. این سرویس خاص از طریق وب ایجاد می شود و توسط کاربر قابل دسترس است.این نرم افزار دارای کاربرد ها و عملکرد های متفاوتی است. طراحی DOCODE از ایده های مشابه بحث شده پیروی می می کند. اساسا ما یک حساب ازاد را ایجاد می کنیم که دارای فایل کم تر از 3 مگ بوده و از پلتفرم وب استفاده می کنیم. ما یک امتیازی را برای اندازه احتمال ارایه می کنیم.

نتیجه گیری

در این مقاله، ما به معرفی DOCODE پرداختیم که یک موتور تشخیص سرقت ادبی بوده و به مربیان و پرفسور ها یک مجموعه ای از ابزار را از طریق هم جوشی و ترکیب اطلاعات از منابع داده های مختلف می دهد. سرقت ادبی اشاره به فرایند ارایه کلمات، افکار و ایده های افراد دیگر به صورت کلمات، افکار و ایده های خود بدون رفرنس دادن به منابع آن ها دارد. رشد ازمایشی منابع اسناد دیجیتالی مختلف موجود در اینترنت موجب تسهیل توسعه این عمل شده و در نهایت موجب شده است تا تشخیص دقیق آن به یک فرایند مهم برای سازمان های اموزشی تبدیل شود.در این مقاله، DOCODE 3.0 که یک سیستم اینترنتی برای موسسات اموزشی جهت تحلیل مقادیر زیادی از اسناد دیجیتال در رابطه با درجه اصلیت است بررسی می شود. چون سرقت ادبی یک مسئله پیچیده است، سیستم ما از الگوریتم هایی برای فرایند تلفیق اطلاعات از منابع چند دادهای به همه این سطوح استفاده میکند. این الگوریتم ها به طور موفق در جامعه علمی درحل مسائلی نظیر شناسایی متن های سرقت شده و بازیابی کاندید های منبع از اینترنت استفاده شده اند.ما این الکوریتم ها را به معماری JEE جند لایه ای و قوی تلفیق کرده و به مشتریان مختلف با نیاز های مختلف امکان می دهیم تا خدمات ما را مصرف کنند. برای کاربران، DOCODE تولید گزارشاتی می کند که معلمان و پرفسور ها امکان دست یابی به اطلاعات در خصوص اصلیت اسناد را می دهد. تجربه ما مربوط به کشور شیلی با زبان اسپانیایی است که راه حل هایی را برای اسناد اموزشی شیمی را در هر یک از محیط های یادگیری مجازی مطلوب ارایه می کیند. با اینحال،DOCODE به اسانی قادر به افزایش پوشش زبان است همان طور که دیدیم هیچ یک از الگوریتم های پینشهادی در این مقاله قادر به حل مسئله با تشخیص پارافریز نبوده اند. با در نظر گرفتن این که پارا فریز دارای اشکال متعددی است، تشخیص توسط DOCODE در برخی از ماژول ها صورت می گیرد. برای مثال FASTDOCODE بستگی به n– گرم به عنوان یک ماژول دارد. ویژگی های دیگر ارایه شده توسط DOCODE یک انالیزور موضوعی از مجموعه زیادی از اسناد می باشد که امکان ترسیم نمودار موضوعی اسناد را می دهد. به عبارت دیگر، تحلیل موضوعی به تعیین این که آیا اسناد تحلیل شده درااای محتوای مشابه است یا خیر کمک می کند. این روش به تحلیل روابط بین یک مجموعه ای از اسناد واصطلاحات برای ایجاد مفاهیم جدید می پردازد. برای انجام این کار، این الکوریتم فرض می کند که از نظر معنی مشابه با قطعات متن

است.رویکرد ما از مطالعه(11) الهام گرفته است که درآن محققان روش ها و فنون تشخیص سرقت ادبی را بر اساس LSAپیشنهاد می کنند. به بیان ساده تر، ما ماتریس Tf-idf را برای اسناد محاسبه می کنیم و این موضوع را در نظر می گیریم که همه پارامترها قابل تحلیل هستند. svd شامل مدل جدیدی از فضای ویژگی است که در آن رابطه معنایی بین این دو در نظر گرفته شده است. وقتی که از svd در همه ماتریس ها استفاده شد،مجموعه ای از ماتریس های اسناد بعدی را میتوان داشت. چون ستون های M مدل های فضای مفهومی از هر سند می باشند، تشابه بین اسناد و ستون ها را میتوان با استفاده از شاخص های مختلف محاسبه کرد.(5).در پایان می توان گفت که DOCODE یک ابزار کاملی است که به معلمان و اساتید برای حل مسئله سرقت ادبی در موسسات اموزشی کمک می کند. بر اساس استفاده از سطوح رابطت مورد استفاده و در نظرگرفتن الگوریتم های فوق، سیستم قادر به پشتیبانی از فرایند های تصمیمگیری درمواجهه با با سرقت ادبی است. هدف نهایی بهبود یادگیری و افزایش کیفیت اموزش و در نظر گرفتن شرایط کشور شیلی به صورت یک مورد است. در کار های اینده، هدف ما اجرای یک سری از تحلیل های کیفی پلتفرم بر اساس بازخورد کاربران می باشد. یک مجموعه ای از ابزار های جدید بر اساس RIA ارایه شده است. با این حال امکان استفاده از ویژگیهای دیگر در عین حال برای استفاده از DOCODEبرای اماده سازی سیستم های فیزیکی برای مقیاس های بزرگ تر وجود دارد. در نهایت، با رشد دیتابیس های درونی، هدف ما طراحی الگوریتم های بهتر برای بازیابی اسناد داخلی و ترکیب و تلفیق الگوریتم در رابط های کاربری میباشد.

ترجمه فا
TarjomeFa.Com

این مقاله، از سری مقالات ترجمه شده رایگان سایت ترجمه فا میباشد که با فرمت PDF در اختیار شما عزیزان قرار گرفته است. در صورت تمایل میتوانید با کلیک بر روی دکمه های زیر از سایر مقالات نیز استفاده نمایید:

✔ لیست مقالات ترجمه شده

✔ لیست مقالات ترجمه شده رایگان

✔ لیست جدیدترین مقالات انگلیسی ISI

سایت ترجمه فا ؛ مرجع جدیدترین مقالات ترجمه شده از نشریات معتبر خارجی