چالش های روش های خوشه بندی به غیر از زبان انگلیسی

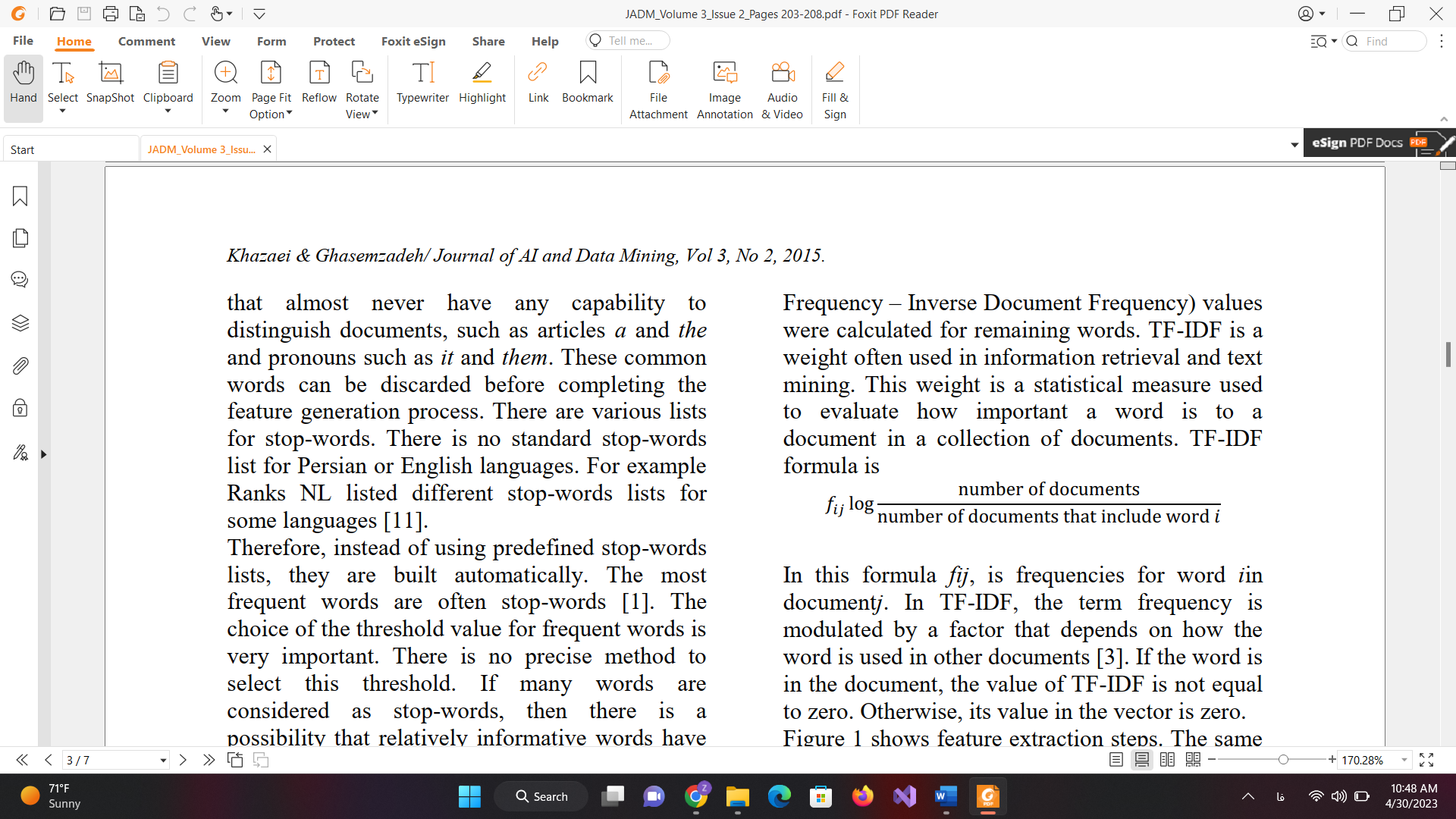
به دلیل اینکه زبان های مختلف میتواند ساختارهای vocabularyوmorphologyو grammarو syntactic متفاوتی داشته باشد ممکن است نیاز به الگوریتم های خوشه بندی موجود یا توسعه الگوریتم های جدید باشد.

کاربرد خوشه بندی اسناد

خوشه بندی اسناد کاربردهای عملی زیادی در زمینه های مختلف پردازش زبان طبیعی مانند بازیابی اطلاعات و خلاصه سازی خودکار متن دارد. به عنوان مثال، خوشه بندی اسناد می تواند دقت بازیابی اطلاعات را در موتورهای جستجو بهبود بخشد. هدف از خوشه‌بندی اسناد، گروه‌بندی اسناد بر اساس محتوای آنها است که می‌تواند به کاربران کمک کند تا اسناد مربوطه را سریع شناسایی کنند و اطلاعات مفیدی را از مجموعه‌های بزرگ متن استخراج کنند.

Tf\_idf

پس از استخراج کلمه، ریشه یابی و حذف کلمات پرتکرار و بسیار کمیاب، مقادیر TF-IDF برای کلمات باقی مانده محاسبه شد. TF-IDF وزنی است که اغلب در بازیابی اطلاعات و متن کاوی استفاده می شود. این وزن یک معیار آماری است که برای ارزیابی اهمیت یک کلمه برای یک سند در مجموعه ای از اسناد استفاده می شود. فرمول محاسبه TF-IDF به صورت زیر است:



در این فرمول،اگر کلمه در سند باشد، مقدار TF-IDF برابر با صفر نیست. در غیر این صورت مقدار آن در بردار صفر است.

بعد از اینکه روی متن tf\_idf اعمال شد. چندین مرحله وجود دارد که بسته به استفاده خاص میتوانیم از آن استفاده کنیم.

1-Clustering: اگر مجموعه بزرگی ازاسناد دارید و می خواهید اسناد مشابه را با هم گروه بندی کنید، می توانید از الگوریتم های خوشه بندی که در یک فصل جداگانه روش های مختلف آن را بررسی میکنیم استفاده کنید. امتیازات TF-IDF برای هر سند می تواند به عنوان ویژگی های ورودی برای الگوریتم خوشه بندی استفاده شود.

2- Classification: اگر می خواهید اسناد را به دسته ها یا برچسب های خاصی اختصاص دهید، می توانید از یک الگوریتم یادگیری نظارت شده مانند logistic regression یا decision tree استفاده کنید. دوباره، نمرات TF-IDF را می توان به عنوان ویژگی های ورودی برای الگوریتم استفاده کرد.

3- Information retrieval: فرآیند جستجو و بازیابی اطلاعات مرتبط از مجموعه ای از داده ها، معمولاً اسناد متنی یا چند رسانه ای است.

سیستم های بازیابی اطلاعات معمولاً از ترکیبی از تکنیک های ranking وindexing وquerying به این هدف استفاده میشود.

فصل خوشه بندی اسناد

خوشه بندی یک تکنیک قدرتمند در تجزیه و تحلیل داده ها است که اشیاء مشابه را بر اساس ویژگی های آنها گروه بندی می کند. در خوشه بندی اسناد فارسی، هدف گروه بندی اسناد فارسی زبان بر اساس محتوای آنهاست. در اسناد فارسی روش های مختلفی برای خوشه بندی وجود دارد که هر کدام مزایا و معایبی دارند. در این گزارش چندین مورد از این روش ها را بررسی می کنیم و بهترین روش موجود در حال حاضر را ارائه می دهیم.

روش خوشه بندی k\_means(کامل نیست)

خوشه بندی K-means یک روش خوشه بندی رایج و اساسی است که در داده کاوی از جمله خوشه بندی متن استفاده می شود. هدف از خوشه‌بندی k-means تقسیم n نمونه به k خوشه است، جایی که هر نمونه به خوشه‌ای با نزدیک‌ترین میانگین تعلق دارد. خوشه های k نهایی باید مجموع مربع های درون خوشه ای را به حداقل برسانند. میانگین مجموع مربع ها معمولاً معیاری برای مقایسه خوشه ها است.

الگوریتم k-means به صورت زیر عمل می کند:

1. تعداد خوشه هایی را که می خواهید ایجاد کنید (k) انتخاب کنید.

2. به طور تصادفی k مرکز خوشه را مقداردهی کنید.

3. هر نمونه را بر اساس فاصله اقلیدسی به نزدیکترین مرکز قرار دهید.

4. مرکز هر خوشه را بر اساس نمونه های اختصاص داده شده مجدداً محاسبه کنید.

5. مراحل 3 و 4 را تا زمان همگرایی تکرار کنید (یعنی زمانی که هیچ تغییر دیگری رخ نمی دهد).

در مرحله 3، هر نمونه بر اساس فاصله اقلیدسی بین ویژگی های آن و ویژگی های مرکز به نزدیکترین مرکز اختصاص داده می شود. در مرحله 4، مرکز جدید برای هر خوشه به عنوان میانگین تمام نمونه های اختصاص داده شده به آن خوشه محاسبه می شود.

یافتن مقدار مناسب برای k می‌تواند چالشی در خوشه‌بندی باشد و روش‌های زیادی برای رسیدگی به این موضوع پیشنهاد شده‌اند.

روش‌های مختلفی وجود دارد که می‌توان برای تعیین مقدار مناسب k در خوشه‌بندی k-means، صرف نظر از زبان مورد استفاده در داده‌ها، استفاده کرد. برخی از روش های رایج عبارتند از:

1-Elbow method : مجموع مجذور فواصل بین هر نمونه و نزدیکترین مرکز آن را برای مقادیر مختلف k رسم کنید. نقطه "elbow" در نمودار نشان دهنده یک مبادله خوب بین به حداقل رساندن واریانس درون خوشه ای و به حداقل رساندن تعداد خوشه ها است.

2-Silhouette method : این روش شامل محاسبه ضریب Silhouette برای هر نقطه داده در مجموعه داده است. ضریب Silhouette اندازه‌گیری می‌کند که یک نقطه داده چقدر به خوشه اختصاص داده شده آن نسبت به خوشه‌های دیگر تعلق دارد. ضریب از 1- تا 1+ متغیر است و مقادیر بالاتر نشان دهنده تخصیص بهتر خوشه است.

3-: Gap statistic method sum of squares(مجموع مربع ها) برای مقادیر مختلف k را با یک توزیع مرجع صفر که توسط داده های تصادفی ایجاد می شود، مقایسه میشود. مقدار بهینه k جایی است که شکاف بین sum of squares مشاهده شده و مورد انتظار درون خوشه ای بزرگ ترین باشد.

Comparing k-means clusters on parallel Persian-English corpus