پروژه پایانی درس مبانی بازیابی اطلاعات و جستجوی وب

استخراج کلمات کلیدی پرتکرار از یک صفحه خبری در Twitter

محمد رضا دلیری (۹۳۱۲۴۳۰۲۳۳)
مهدی اکبریان رستاقی (۹۳۱۲۴۳۰۴۳۷)

در این پروژه سعی شده تا با استفاده آخرین نسخه کتابخانه Lucene (نسخه 7.2.1) مجموعه‌ای از کلمات کلیدی پرتکرار (داغ) بکار رفته در یک صفحه خبری در شبکه اجتماعی Twitter در و حوزه اقتصادی و اجتماعی استخراج شده و پس از اندیس‌گذاری، مهم ترین آن‌ها بازیابی شود. در این بررسی از صفحه رسمی **خبرگزاری تسنیم** به آدرس **@Tasnimnews_Fa** به‌علت متنوع بودن خروجی خبری آن استفاده شده است.

۱ اجرای برنامه

اجرای پروژه سه مرحله اصلی دارد، ابتدا لازم است تا توئیت‌ها crawl شده و سپس اندیس‌گزاری و در نهایت بازیابی شود.

۱.۱ استخراج توئیت‌ها

عملیات crawling با استفاده از کلاس **Crawler** انجام می‌شود. نحوه استفاده از این کلاس به‌شرح زیر است:

```
java -cp ProjectNews.jar ir.ac.um.ce.projectnews.crawler.Crawler
   FLAGS
```

پرچم‌های مورد نیاز عبارتند از:

-i: مشخص‌کننده شناسه صفحه‌ی مورد نظر در توئیتر برای استخراج توییت‌ها

-s: مشخص‌کننده تاریخ شروع بازه زمانی مورد نظر برای دریافت توئیت‌ها (با فرمت YYYY-MM-DD)

-e: مشخص‌کننده تاریخ پایان بازه زمانی مورد نظر (با فرمت YYYY-MM-DD)

-m: محدود کننده حداکثر تعداد پیام های استخراج‌شده در بازه زمانی مورد نظر

-p: مسیر ذخیره فایل CSV محتوی توئیت‌های استخراج‌شده

-n: نام انتخابی برای فایل CSV

به‌عنوان مثال برای استخراج توئیت‌های صفحه‌ی **Tasnimnews_Fa** در ژوئن ۲۰۱۸ و ذخیره آن در مسیر **out** ورودی باید به‌صورت زیر باشد:

```
java -cp ProjectNews.jar ir.ac.um.ce.projectnews.crawler.Crawler
   -i Tasnimnews_Fa -s 2018-06-01 -e 2018-06-30 -p out/
```

۲.۱ اندیس‌گذاری

با توجه به خام بودن داده‌های برنامه (توئیت‌ها) ابتدا باید پیش‌پردازش روی داده‌ها انجام شود.

در فرآیند پیش‌پردازش ازکلاس **PersianAnalyzer** کتابخانه Lucene و لیست stop words ارائه‌شده توسط *آقای وحید خرازی* [1] استفاده شده است.

با استفاده از کلاس **Indexer** می‌توانیم روی داده‌های ورودی فرآیند پیش‌پردازش را انجام داده و سپس آن‌ها را اندیس‌گذاری کنیم. نحوه اجرای این برنامه به‌صورت زیر است:

```
java -cp ProjectNews.jar ir.ac.um.ce.projectnews.search.Indexer
   CORPUS_FILE [INDEX_DIR]
```

CORPUS_FILE: مسیر فایل corpus با فرمت CSV که شامل توئیت‌ها با فرمت زیر است:

```
ID, Permalink, Text, Date, Retweets, Favorites
```

INDEX_DIR: مسیر مورد نظر برای ذخیره‌سازی index ساخته شده توسط برنامه را مشخص می‌کند. این پارامتر اختیاری است و برنامه به‌صورت پیش‌فرض از پوشه indices (درکنار فایل اجرایی برنامه) استفاده می‌کند.

[1] Persian (Farsi) Stop Words List

۳.۱ دسته‌بندی توئیت‌ها و یافتن کلمات کلیدی پرتکرار

به‌علت مشخص نبودن موضوع توئیت‌های استخراج شده، ابتدا باید توئیت‌ها دسته‌بندی شده و پیام‌های اضافه حذف شوند. برای مشخص کردن دسته‌بندی هر توئیت، از تکنیک جستجو استفاده شده است. بدین صورت که برای هر دسته (اقتصادی و اجتماعی) یک لیست از کلمات کلیدی مرتبط تعریف شده و با جستجوی یک پرس‌وجو شامل این کلمات، پیام‌های مرتبط بر اساس مقدار threshold شناسایی می‌شوند. سپس این پیام‌ها به نمایه‌ی جدیدی منتقل شده و فرآیندهای آتی روی نمایه جدید انجام می‌شود. مقدار threshold از فرمول زیر بدست می‌آید:

$$\frac{minScore + maxScore}{۲}$$

لیست کلمات کلیدی مورد استفاده برای دو بخش **اقتصادی** و **اجتماعی** در جدول ۱ آمده است.

پس از تهیه یک نمایه جدید به‌ازای هر دسته از کلمات که فقط شامل توئیت‌های مرتبط با آن دسته است، فرآیند جستجو برای یافتن پرتکرارترین کلمات (terms) انجام می‌شود. پس از پایان این جستجو، خروجی برنامه به‌صورت یک لیست متنی از کلمات به‌ترتیب تکرار (نزولی) همراه با تعداد تکرار هر کلمه در خروجی استاندارد چاپ می‌شود. این لیست پس از بررسی از نظر ارتباط معنایی با دسته‌بندی، به‌عنوان خروجی نهایی ارائه می‌شود.

چگونگی فراخوانی این برنامه به شرح زیر است:

```
java -cp ProjectNews.jar ir.ac.um.ce.projectnews.search.
   Classifier [INDEX_DIR] QUERIES_FILE [RESULTS_COUNT]
```

INDEX_DIR: به محل ذخیره‌سازی index ساخته شده در مرحله قبل اشاره می‌کند. مانند مرحله پیشین، این پارامتر اختیاری است و برنامه به‌صورت پیش‌فرض از پوشه indices (در کنار فایل اجرایی برنامه) استفاده می‌کند.

توجه: در صورت تغییر مقدار این پارامتر، لازم است پارامتر سوم (RESULTS_COUNT) هم مقداردهی شود.

QUERIES_FILE: مسیر یک فایل متنی ساده حاوی کلمات کلیدی مشخص‌کننده را مشخص می‌کند. این فایل یک سطر دارد.

RESULTS_COUNT: تعداد توئیت‌های بررسی شده را مشخص می‌کند. این پارامتر نیز اختیاری بوده و مقدار پیش‌فرض آن، **صد هزار** است.

توجه: در صورت تغییر مقدار این پارامتر، لازم است برای پارامتر اول (INDEX_DIR) هم مقداری مشخص شود.

جدول ۱: لیست کلمات کلیدی مورد استفاده برای دسته‌بندی توئیت‌ها

دسته‌بندی اقتصادی	سود، ارز، یارانه، واردات، تومان، نزول، سکه، قیمت، بازرگانی، طلا، نرخ، مسکن، رهن، اجاره، مستاجران، برجام، تجاری، صنایع، بازرگانی، بها، کاهش، دلار، افزایش، افزایشی، کلان، اقتصاد، بازار، تک‌نرخی، حقوق، اونس، بهای، سرمایه، تحریم، سرمایه‌گذار، واردکنندگان، خودرو، نرخ، بانک، تجارت، توسعه، سازمان، فروش، هزینه، میانگین، تورم، جواهر، اتحادیه، رانت، اقلام، کالا، میلیون، هزار، رکورد، خودرو، محصولات، نمایندگی، قاچاق، قرارداد، لابی، یارانه، نفت، نفتی، نقدینگی، نقدی، نقد، خانه، رقم، مالیات، مالیاتی، مالک، مالکانی، پول، صادرکنندگان، صادرات، خصوصی، هواپیما، مصرف‌کنندگان، مصرف، سرمایه‌گذاری، گران، گرانی، آماری، آمار، امار، محصول، کاهش، ایرباس، توتال، فساد، بودجه، شاخص، بورس، رشد، تحریم، اوپک، مدیریت، تولید، واردات، وام، بدهکار، نوساز، یورو، پوند، ریال، برنز، نقره، الومینیوم، آلومینیوم، مس، آهن، اهن، فولاد، بیکار، کارمندان، بهره، قسط، تحریم، میلیارد، اشتغال، تخلف، سفارش، سرمایه، معاملات، معامله، اسکناس، موجودی، مصرف‌کننده، دلال، دلالان، دزد
دسته‌بندی اجتماعی	اجتماعی، فیلم، مردم، دستگیری، مترو، اتهام، فرهنگیان، قتل، آزادی، فیلترینگ، فیلتر، مصرف، پلیس، ایران، دریاچه، ازدواج، طلاق، مهریه، خانواده، خرید، آموزش، پزشکی، پزشک، محرم، مهمانی، ترافیک، روابط، ادارات، اداره، کارگر، مرگ، مترو، سکانس، بهداشت، تحصیل، پلیس، آب، پزشکی، دارو، سلامت، کشور، حج، حاجی، بیمه، دانش‌آموزان، دانش‌آموز، مهمان، تابستان، تعطیلات، محاکمه، قرص، روانگردان، اعتیاد، معتاد، جسد، میوه، بیماری، تهران، مدرسه، آتش‌سوزی، مستمری، سفر، مسافرت، بازنشستگان، فرهنگستان، تصادف، صنایع‌دستی، کارت، ملی، مخدر، نوجوان، استخدام، ایران، معلم، معلمان، صندوق، شهر، استان، زلزله، سیل، جنگل، عید، مسافرت، مسافر، هتل، رستوران، تفریح

۲ کامپایل برنامه

سورس برنامه در پوشه `src` قابل مشاهده است. در صورتی تمایل برای کامپایل سورس‌کد، لازم است از ابزار Maven استفاده کرده و فایل `pom.xml` به‌عنوان فایل تنظیمات به این ابزار داده شود.

لازم به ذکر است که دریافت توئیت‌ها از کتابخانه GetOldTweets-java (نسخه 1.2.0) و برای ذخیره آن‌ها از استاندارد CSV با کمک کتابخانه OpenCSV استفاده شده است.

۳ سورس‌کد برنامه

این برنامه شامل سه کلاس اصلی زیر می‌باشد:

کلاس `Crawler`: این کلاس همان‌طور که بیان شد برای دریافت و استخراج توئیت‌ها تهیه شده است. با توجه به محدودیت‌های API رسمی Twitter، کتابخانه GetOldTweets-java مورد استفاده قرار گرفته است.

کلاس `Indexer` و کلاس `Writer`: این دو کلاس در مجموع برای پیش‌پردازش، ساخت index و ذخیره‌سازی آن در محل مورد نظر کاربر استفاده می‌شود..

کلاس `Classifier` و کلاس `Searcher`: وظیفه‌ی این کلاس‌ها درکل جداسازی توئیت‌های مرتبط با هر دسته‌بندی و سپس مشخص‌کردن کلمات کلیدی پرتکرار آن دسته‌بندی می‌باشد. برای تشخیص توئیت‌های مرتبط با هر دسته‌بندی و لیست کلمات آن، از الگوریتم Okapi BM25 استفاده شده است.

۴ نتایج بدست آمده

پس از اجرای فرآیند گفته شده روی صفحه خبرگزاری تسنیم در بازه زمانی سه ماه (ماه‌های آوریل، مه و ژوئن ۲۰۱۸) برای دو دسته اقتصادی و اجتماعی، کلمات کلیدی پرتکرار در هر ماه مشخص شد که لیست تفکیکی و مرتب‌شده آن‌ها (به‌صورت نزولی) در جداول ۲ و ۳ ارائه شده است.

جدول ۲: پرتکرارترین کلمات کلیدی در خبرهای اقتصادی

رتبه	آوریل ۲۰۱۸	مه ۲۰۱۸	ژوئن ۲۰۱۸
۱	تومان	ایران	بازار
۲	قیمت	تحریم	ارز
۳	دلار	کشور	ایران
۴	کشور	نفت	سرمایه
۵	ارز	برجام	قیمت
۶	افزایش	خصوصی	دلار
۷	بازار	واگن	افزایش
۸	مصرف	اروپایی	تحریم
۹	تحریم	امریکا	نرخ
۱۰	خودرو	بانکی	مصرف
۱۱	ریال	تولید	نفت
۱۲	سکه	روحانی	اروپا
۱۳	فروش	سرمایه	امریکا
۱۴	نرخ	سوئیس	اهن
۱۵	۴۲۰۰	شرکت	دولت

جدول ۳: پرتکرارترین کلمات کلیدی در خبرهای اجتماعی

رتبه	آوریل ۲۰۱۸	مه ۲۰۱۸	ژوئن ۲۰۱۸
۱	اب	دانش اموزان	اموزش
۲	مترو	زندان	مدرسه
۳	مصرف	اتهام	پخش
۴	ایران	ازاد	دانش اموزان
۵	تجریش	اموزش	ایران
۶	تهران	تهران	فوتبال
۷	دریاچه	حمید صفت	پلیس
۸	عذرخواهی	قتل عمد	اذیت
۹	فروشی	ایران	تجاوز
۱۰	پلیس	بازداشت	مدارس