گزارش پروژه ناشناس کردن اطلاعات شخصی

نام خانوادگی اعضای گروه: نفیسه نیک اقبال- محمد هادی حاجی حسینی- الیاس اسماعیلی

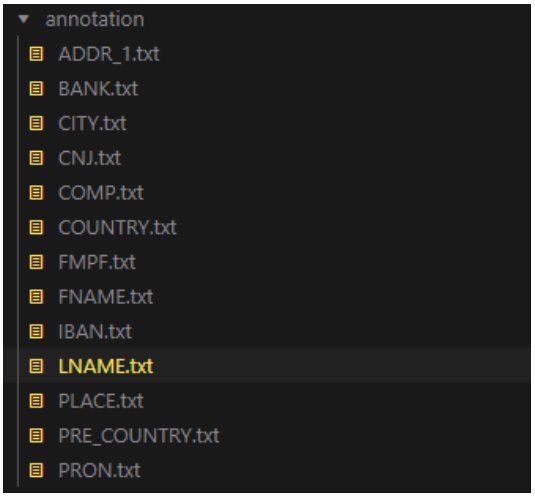
پوشه های اصلی پروژه :

پوشه annotation :

داخل این پوشه دیتاست های ما مانند لیست اسامی، لیست نام های خانوادگی، لیست نام شرکت ها، لیست اسامی شهرها و کشورها وجود دارد.

علاوه بر موارد ذکر شده واحدهای سازنده هرRule ها مانند کلمات خاصی که در آدرس‌ها استفاده می‌شوند مانند کلمات شهر، خیابان، جنب، روبه رو، علائم ربط مثل کاما و ویرگول، ضمایر و... آورده شده اند که ما از فایل‌های داخل این پوشه استفاده می‌کنیم تا Pattern ها را بسازیم که در حقیقت بخشی از واحدهای سازنده pattern ها این annotation ها می‌باشند.

لیست فایل‌های داخل پوشه annotation به شکل زیر است:



فایل ADDR_1 : کلمات کلیدی موجود در آدرس‌ها مانند خیابان، کوچه، جنب، بزرگراه و.... می باشد.

فایل BANK : شامل 6 شماره معتبر اول شماره‌های کارت می‌باشد.

فایل CITY : شامل اسامی تمامی شهرهای ایران و دنیا می‌باشد.

فایل COUNTRY : شامل اسامی تمام کشورها می‌باشد.

فایل PLACE : مکان های مختلف مثل سینما، قوه قضائیه، رستوران و ... می‌باشد.

فایل IBAN : شامل regex های معتبر کشورهای مختلف برای IBAN می‌باشد.

فایل COMP : نام کلیه شرکت هاهم انگلیسی و هم فارسی در این فایل موجود می‌باشد. که با crawl سایت دانشکار و jobinja بدست آمده‌اند.

فایل FNAME و LNAME : یک مجموعه از اسامی فارسی و فامیل های فارسی رایج می‌باشد که بخشی از آن ها با crawl سایت ویکی پدیا بدست آمده است.

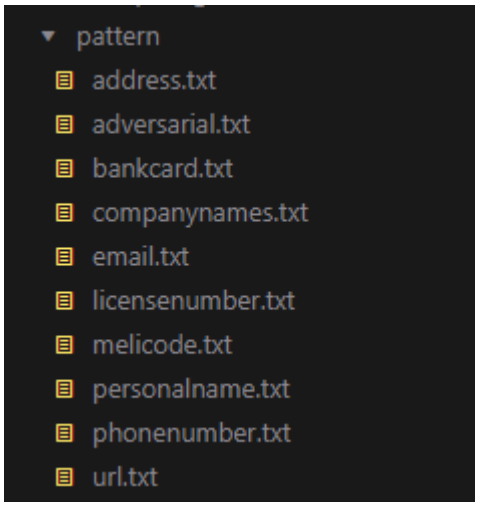
فایل PRON : شامل تمام ضمائر جدا و چسبیده فارسی می‌باشد.

فایل CONJ : شامل ویرگول و کاما و - می‌باشد.

فایل FMPF : این فایل تمامی پسوندهای معتبر برای فامیل مانند فر، نژادو... می‌باشد

Pattern :

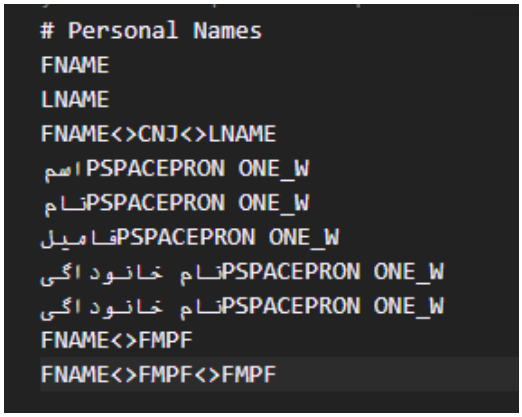
ما داخل این پوشه برای هر category از اطلاعات که می‌خواهیم اطلاعات آن را مخفی کنیم یک فایل در نظر می‌گیریم و در داخل pattern ها ما با استفاده از annotation ها قانون های خود را می‌نویسیم که در نهایت این ruleها تبدیل به regex می‌شوند.



به عنوان مثال فایل personalname شامل قانون هایی است که برای شناسایی نام افراد می‌خواهیم داشته باشیم. که هر خط از داخل این فایل به یک rule تبدیل می‌شود و داخل فایل‌هایی که در ادامه توضیح می‌دهم به regex تبدیل می‌شود.

در خط اول و دوم می‌گوید هر جا یکی از اسم ها یا فامیل‌های داخل فایل‌های FNAME یا LNAME آمد را در نظر بگیر و خط بعدی نیز می‌گوید بین اسم و فامیل می‌تواند علائم ربط مثل – نیز بیاید ودر خطوط بعدی می‌گوییم اگر کلمات اسم،نام،نام خانوگی و فامیل امدند و بعد آن‌ها یک ضمیر منفصل یا متصل آمد کلمه بعدی آن‌ها را نیز به عنوان اسم در نظر بگیرد و این باعث می‌شود که اگر جمله ای مانند" اسم من جیسون اکبری است" را کلمه جیسون را به عنوان اسم برگرداند و هم چنین FNAME<>FMPF برای ما از روی اسامی و پسوندهای تعریف شده فامیل می‌سازد و علامت<> یعنی صفر فاصله یا بیشتر می‌توانند داشته باشند. به

عنوان مثال نام محمد را داریم و از روی آن فامیل‌های محمدی، محمدی فر، محمدپور و.... را می‌سازد. که قانون‌های توضیح داده شده در شکل زیر آورده شده اند:



فایل pattern_to_regex :

در این فایل ما دو کلاس داریم :

1. کلاس Annotation :

تمام annotation ها را از روی فایل‌ها خوانده و یک دیکشنری ساخته که کلید آن برابر اسم فایل می‌باشد و value آن OR تمام اطلاعات خوانده شده از فایل‌ها می‌باشد و هم چنین اگر regexای هم بخواهیم اضافه کنیم ک داخل متن فایل‌های annotation نبوده و بخواهیم در داخل patternها از آن استفاده کنیم را مستقیم اینجا تعریف کرده و به دیکشنری خود اضافه می‌کنیم.

2. کلاس Pattern :

در این کلاس ما فایل‌های پوشه Pattern را خوانده و یک دیکشنری باید درست کنیم که کلید آن شامل نام فایل‌های پوشه Pattern است و value آن شامل تمام regex های قانون‌هایی است که در داخل این فایل‌ها نوشتیم و در حقیقت در داخل این کلاس با استفاده از دیکشنری ساخته شده در کلاس annotation یک دیکشنری جدید می‌سازیم که شامل تمام regex ها ما هست.

فایل Spans :

در این فایل کاری که انجام می‌دهد این است که با توجه به دیکشنری ساخته شده در مرحله قبل span تمام regex های نوشته شده را بدست می‌آورد و اگر دوتا pattern پیدا کرده باشد و بین آن‌ها فقط space باشد

آنگاه spanهای آن‌ها را با هم merge می‌کند. به عنوان مثال در جمله "علی رضا را در مدرسه دیدم" علی را داخل نام‌ها پیدا می‌کند رضا را هم پیدا می‌کند و بعد چون بین آن‌ها فقط space است آن‌ها را به هم می‌چسباند.

هم چنین چون اسم آدرس‌ها گاهی به عنوان اسم اشخاص در نظر گرفته می‌شود ما در این کلاس اگر یک اسم به عنوان آدرس تشخیص داده شده باشد دیگر به عنوان نام اشخاص در نظر نمی‌گیریم.

فایل anonymizer :

در این فایل یک کلاس Model تعریف می‌کنیم و در آن از کلاس pattern استفاده کرده و regex ها را ساخته و بعد از پکیج parstdex نیز استفاده می‌کنیم تا زمان‌ها و تاریخ را استخراج کنیم و بعد در نهایت تمام span ها در قالب یک دیکشنری برمی‌گردانیم و در کلاس example از مدل ساخته شده استفاده می‌کنیم.

ورودی :

```
sent = """
علی احمدی در شهرستان اصفهان شهر فولادشهر به دنیا آمد.
علی رضا و زهرا در کشور زیمباوه زندگی می‌کنند.
زندگی در کشور france به نظر سخت است.
کشور به منطقه‌ای گفته می‌شود که مرز آن با سیاست تعیین شده است.
من در ۱۶ بهمن ۱۳۷۵ به دنیا آمدم.
در تاریخ ۱۶ فروردین به مکه رفتم و از شرکت داده پردازش نیز دیدن کردم.
کدملی من ۱۱۳۰۳۹۴۷۸۹ است.
کد ملی من ۱۱۳۰۳۹۴۷۸۹ است.
شماره کارت اعتباری من 6104337958646987 است که از آن به حساب شما پول می‌ریزم. من سایت .
شماره شبای من IR069600000001032420000011 است.
شماره IBAN من KW81CBKU0000000000001234560101 است.
ساعت ۸ صبح پرواز به مقصد روستای احمدآباد را دارم.
من در اصفهان، کوشک، خیابان احمدی، کوچه شهید علی علیزاده پلاک ۱۴۳ زندگی میکنم.
این نشانی خیابان بهشتی است.
شماره تلفن من ۰۹۱۲۳۴۵۶۷۸۹ و شماره خانه علیرضا ۰۲۱۳۳۴۴۵۵۶۶ است.
خیابان اهواز شهر اهواز بسیار تمیز است.
"""
```

خروجی:

```
<#نام شخص#> در <#آدرس#> به دنیا آمد.
<#نام شخص#> و <#نام شخص#> در <#آدرس#> زندگی می‌کنند.
زندگی در <#آدرس#> به نظر سخت است.
کشور به منطقه‌ای گفته می‌شود که مرز آن با سیاست تعیین شده است.
من در <#تاریخ و ساعت#> <#تاریخ#> به دنیا آمدم.
در <#تاریخ و ساعت#> <#تاریخ#> به <#آدرس#> رفتم و از <#اسم شرکت#> نیز دیدن کردم.
کدملی من <#کد ملی#> است.
<#کد ملی#> کد ملی من است.
سایت <#آدرس سایت#> رو مشاهده کردم و ایمیل کارمند <#اسم شرکت#> <#ایمیل#> را برداشتم.
شماره <#تاریخ و ساعت#> <#تاریخ#> من <#اطلاعات حساب#> است.
من <#اطلاعات حساب#> است IBAN شماره.
تاریخ و ساعت <#زمان#> پرواز به مقصد <#آدرس#> را دارم.
من در <#آدرس#> زندگی میکنم.
این نشانی <#آدرس#> است.
شماره تلفن من <#شماره تماس#> و شماره خانه <#نام شخص#> <#شماره تماس#> است.
آدرس <#آدرس#> بسیار تمیز است.
```

ما از پایتون 3 استفاده کردیم و از کتابخانه re برای شناسایی regexها و از parstdex برای شناسایی زمان و تاریخ استفاده نمودیم.