محمد مهدی یادگار

96522051

موضوع این پروژه متون مربوط به شکایات و آرای قضایی در دو حوزه‌ی کیفری و حقوقی می باشد. برای بدست آوردن متن مورد نیاز، از سایت https://ara.jri.ac.ir/Judge/Index اقدام به کرال کردن کردم. در صفحه‌ی اصلی سایت مکان تفکیک بر اساس نوع شکایت حقوقی و کیفری وجود دارد که از آن ها استفاده شده است. برای کرال کردن از کتاب خانه‌ی Scrapy استفاده می‌شود. مشکل اصلی ای که در این بخش وجود داشت این بود که بدلیل اینکه طراحی سایت چند نوع است کار استخراج داده را دشوار ساخته است که مجبور شدم از چند روش برای گرفتن اطلاعات استفاده کنم. دو اسپایدر برای این کار ایجاد شده است که در پوشه‌ی src/ijra/ijra/spiders موجود می باشند.

بعد از کرال کردن و بدست آوردن اطلاعات کاری که باید انجام شود تمیز کردن داده است. دو اسپایدر ما دو فایل جیسون به ما دادند که باید آن ها را بخوانیم و اطلاعات لازم را بدست آوریم. مشکلاتی در متن ها وجود داشت مانند آنکه مثلا اسامی به صورت اختصاری آورده شده بودند مانند : الف.ب. . که اینکار بسیار مشکل ایجاد می کرد مخصوصا در زمان شکستن جملات. مشکل دیگر آورده شدن شماره ی قوانین بود که آن ها هم حذف شدند. حروف اضافه و اشاره نیز تا حدی پاک سازی شدند. نرمالایزر هزم روی دادگان اجرا شد و هم چنین از توکنایزر و sentence breaker هزم نیز استفاده شد. دادگانی که حذف شدند شامل 10 تا 20 عدد از داکیومنت ها بودند که عدد قابل توجهی نبود.

کلاس بندی دادگان بر اساس هر داکیومنت انجام شده است و در دو کلاس کیفری و حقوقی قرار می گیرند.

نکات آماری این دادگان نیز به این شرح است:

تعداد واحد داده : 6904

تعداد جملات : 46345

تعداد کلمات : 3592975

تعداد کلمات منحصر به فرد : 27458

تعداد کلمات منحصر به فرد مشترک بین برچسب ها : 10581

تعداد کلمات منحصر به فرد غیر مشترک بین برچسب ها : 16877

10 کلمه پرتکرار غیر مشترک کیفری : ('قصاص', 314), ('اولیای', 306), ('ارتشا', 97), ('هرمی', 92), ('تسخیری', 91), ('گلوله', 88), ('اشد', 86), ('زانوی', 86), ('مقتوله', 83), ('عضوگیری', 81)

10 کلمه پرتکرار غیر مشترک حقوقی : ('نحله', 335), ('پیشه', 264), ('خواهی_نموده', 251), ('العقد', 190), ('اجراییه', 186), ('وصیت', 173), ('بائن', 162), ('عسروحرج', 144), ('حسبی', 143), ('وقفیت', 132)

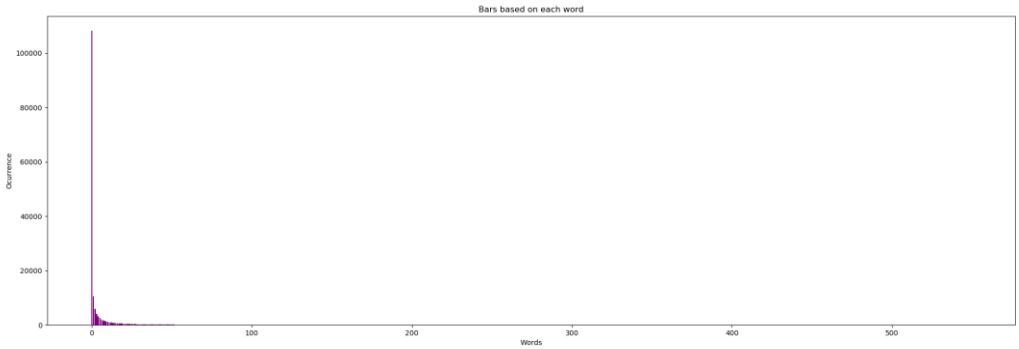
10 کلمه مشترک برتر کیفری بر اساس فرمول : ('رشوه', 276.28459756223634), ('شلیک', 266.88716227100383), ('کفیل', 236.81536933905974), ('ربایی', 233.99613875169), ('بزهکاری', 167.81134448629518), ('متهمان', 166.17798073329524), ('متهمه', 161.6358870091995), ('تخفیف', 152.94325936480942), ('شکات', 149.95621743295337), ('مرتکبین', 145.66024701410421)

10 کلمه مشترک برتر حقوقی بر اساس فرمول : ('فرجامی', 396.38474589713763), ('داوران', 375.1023434328618), ('فرجام', 267.03503314203886), ('خواهان\\u200cها', 261.7735503105929), ('داور', 240.13644113857916), ('حرج', 198.7244330101757), ('عسر', 197.92634291776537), ('خواهان\\u200cهای', 130.88677515529645), ('موقوفه', 113.86085318387578), ('بطلان', 97.31862217755227)

10 کلمه ی برتر TF_IDF کیفری : ['دادگاه', 'قانون', 'خواهان', 'دعوی', 'خوانده', 'ماده', 'دادرسی', 'آقای', 'دادنامه', 'شماره']

10 کلمه ی برتر TF_IDF حقوقی : ['دادگاه', 'قانون', 'آقای', 'ماده', 'متهم', 'الف', 'عمومی', 'پرونده', 'دادنامه', 'رأی']

هیستوگرام تکرار کلمات بدلیل آنکه تعداد تکرار بعضی کلمات خیلی زیاد بود بسیار بایاس می باشد و هم چنین چون تعداد کلمات زیاد است تعداد از آن ها پلات شده اند. برای دیدن بهتر کد را ران کنید تا بتوانید در تصویر زوم کنید.


Bars based on each word
Ocurrence
100000
80000
60000
40000
20000
0
0
100
200
300
400
500
Words