فاز اول پروژه ی پایانی

پردازش زبان های طبیعی

نام:مهسا قادران

استاد: دکتر صالح اعتمادی

سرپرست گروه:زهرا سیدی

موضوع:طبقه‌بندی نوع اخبار موجود در سایت BBC فارسی:

هدف از این پروژه بررسی و مقایسه اخبار منتشر شده از این خبرگزاری بر اساس دسته‌بندی های انجام شده موجود در وبسایت رسمی خبرگزاری فوق است. با استفاده از تحلیل متن می‌توان مقایسه انواع کلمه های استفاده شده و نقش کلمات در جمله, میزان احساسات و انواع دیگر تسک ها موجود در پردازش زبان طبیعی را در انواع خبر ها را با هم مقایسه کرد.

جمع‌آوری اطلاعات

در قدم اول برای وبسایت BBC یک crawler نوشتم که در قالب یک ماژول نوشته شده است. خروجی این قسمت از کد به این صورت از که تمام اخبار در موضوع‌های مورد نظر را از خبرگزاری BBC استخراج میکند و در قالب یک فایل csv به صورت خام ذخیره میکند. با توجه به این که سایت BBC زبان فارسی به صورت روزانه آپدیت میشود در هربار از crawl آن اطلاعات جدیدی وجود دارند و هر بار این مقدار بیشتر میشود.

برای crawl کردن صفحات از ماژول BeautifulSoup در پایتون و همچنین ارسال request استفاده کردم. برای این کار در اولین قدم لازم است که فایل های HTML وبسایت BBC بررسی شوند تا بتوان دقیقا اطلاعات اخبار به درستی استخراج شوند و همه‌ی خبرهای در دیتاست موجود باشند همچنین در صورتی که تبلیغی در وبسایت موجود است, جز متن خبر وارد دیتاست نشود.

معماری کد به این گونه است که یک فایل کانفیگ و سه کلاس به ترتیب برای crawl کردن صفحه اصلی و پیدا کردن آدرس هر کلاس, کلاس دیگر برای crawl کردن صفحه ای که شامل لیست خبرها ست و نهایت صفحه ای که متن اصلی خبر در آن است. در ادامه به توضیح مفصل تر کلاس صفحه پرداخته میشود.

Config.yaml

در این فایل اطلاعات اولیه نظیر url سایت bbc و همچنین کلاس هایی که برای این پروژه در نظر گفته شده است را در آن وارد میکنیم. مورد دیگری که در این فایل که از نوع yaml است تعداد صفحاتی از هر موضوع هست که قصد داریم اخبار آن ها crawl شود.

Main_page.py

در این کلاس تنها قسمت header صفحه اصلی جهت پیدا کردن لینک های صفحات مرتبط با هر خبر بررسی میشوند و درصورتی که موضوع خبر مطابق با آنچه در فایل config خواسته شده است باشد, کلاسی از نوع لیست خبرهای فیلد مورد نظر میسازد.

Fields_page.py

نمایی کلی از این صفحات به صورت زیر است. در این صفحه هات ابتدا خبرهای از سایر قسمت ها جدا میشوند. و به طور کلی دو نوع خبر داریم. خبرهایی که کل متن خبر در همین صفحه موجود هستند و دسته دوم خبرهایی هستند که متن اصلی خبر به صفحه دیگری ارجاع داده شده است. این کلاس با این دو مدل خبر به دو گونه متفاوت رفتار میکند. و هدف آن ساختن لیستی از کلاس News است که اطلاعات مربوط به یک خبر در آن درج شده است.

نهایتا پس از crawl کامل یک صفحه و به دست آمدن تمام خبرهای آن تمام خبرها در فرمت فایل csv ذخیره میشوند. عمل ذخیره‌سازی برای جلوگیری از دست رفتن اطلاعات در صورت بروز هرگونه مشکل در میان

صفحات crawl شده در حین استخراج اطلاعات ذخیره میشود و همچنین برای بالا رفتن سرعت پس از استخراج تمام خبرهای هر صفحه انجام میشود.

News_page.py

در این کلاس اطلاعات مربوط به یک خبر در صورت نیاز از وب گرفته شده و در قدم بعدی از فایل html خام آن تیتر خبر, موضوع آن, آدرس لازم برای دیدن خبر و همچنین متن خبر به دست می‌آید. با توجه به این که این کلاس دو نوع خبر را استخراج میکند به گونه ای طراحی شده است که به درستی اخبار استخراج شوند.

Statistics

همه نمودارهای آماری موجود در این داکیومنت و سایرین که ارجاع داده شده اند در این پوشه موجود میباشد.

Documents

داکیومنت های نوشته شده برای هرفاز از پروژه در این پوشه موجود است.

پیش پردازش اطلاعات

در این مرحله دیتای خام استخراج شده در مرحله قبل پردازش میشود و نهایتا در فایلی به نام dataset-prepcossed ذخیره شده است. با توجه به این که متن اخبار بعضا ممکن است خیلی طولانی باشند و همیشه مهم ترین قسمت خبر در ابتدای آن بیان میشود, پس از جداسازی جملات متن هر خبر ماکزیمم ۱۰ جمله از هر خبر نگهداری میشود.

در زبان فارسی جملات با استفاده از علامت نگارشی نقطه از یکدیگر جدا میشوند به جز این مورد هیچ کدام از علایم نگارشی باید از جملات حذف شوند. هنگام crawl بعضی از علامت های نگارشی به کار رفته به درستی نمایش داده نشده اند. در نتیجه علاوه بر حذف علائم نگارشی از جملات این علایمی که به درستی انتقال نیافته اند نیز از داده حذف گردیدند.

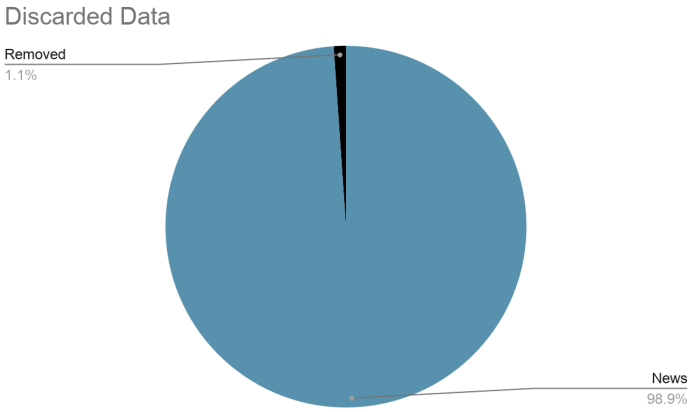
تعدادی از کلمات در زبان فارسی بار معنایی خاصی ندارند. به این کلمات در حیطه پردازش زبانهای طبیعی stop words گفته میشود با توجه به کاربرد در این زمینه و کلمات پرتکرار در ادبیات این وبسایت لیستی از stop words ها تهیه کردم و از کلمات دیتاست حذف کردم.

با توجه به این که دیتاست از یک وبسایت رسمی جمع آوری شده است همانطور که در پروپوزال پیش بینی میشد کمتر اطلاعات نامناسب در دیتاست یافت میشوند و نیازی به پیش پرداز پیچیده در این مرحله وجود ندارد.

یکی از راه هایی که برای تمیز کردن دیتا میتوان استفاده کرد tokenize کردن و همین طور lemmatize کردن کلمات است که سعی میشود کلماتی که هم ریشه و مشابه هم هستند به عنوان یک کلمه در نظر

گرفته شوند. با توجه این که در این مرحله از پروژه هدف بررسی کلمات به کار رفته در اخبار است استفاده از lemmatize منطقی به نظر نمیرسد. درصورتی که بخواهیم عملیات lemmatize را در زبان فارسی انجام دهیم از بهتری کتابخانه هایی که برای زبان فارسی موجود است stanza است که با ایجاد یک pipeline و استفاده از دستور lemma کلمات را lemmatize کرد.

بعضی صفحات نیز ممکن است در طول استخراج و ارسال و دریافت مخدوش شده باشند که این ردیف ها نیز از دیتاست حذف میشوند. در بعضی اخبار کاملا به صورت تصویر یا فیلم بوده اند و متن مرتبطی نداشته اند. این اخبار نیز در حالت preprocess حذف شده و نهایتا ۳۸۰۷ تعداد اخباری است که در دیتاست باقی میماند. تعداد اولیه خبر های موجود در دیتاست برابر ۳۸۴۹ بوده است. در نتیجه در طی preprocess به مقدار یک درصد از دیتا از دست میرود.



بهترین ماژول موجود در زبان فارسی stanza میباشد. این ماژول نمیتواند جملات را به خوبی جدا کند. با توجه به نوع دیتا میتوان آن را از روی علائم نگارشی ',' جدا کرد. اما برای جداسازی کلمات به خوبی عمل میکند. برای جداسازی کلمات و جملات جهت جلوگیری از تولید دیتاهای شبیه به هم درکد این جداسازی صورت میگرد و درجایی ذخیره نمی‌شود.

داده‌های آماری

در این بخش به بیان بعضی از آمارهای موجود در دیتاست جمع‌آوری شده پرداخته میشود. همان طور که اشاره شد منبعی که از آن دیتاست جمع‌اوری شده است روزانه تغییر میکند در نتیجه این آماربرای این نسخه از داده ها صحت دارد. لازم به ذکر است که انتظار می‌شود دربازه‌های زمانی کوتاه این نسبت تغییر نکند اما در بازه‌های زمانی بلندتر قابل پیشبینی نیست.
نکته قابل توجه این است که در دیتاست جمع‌اوری شده هم متن خبر وهم تیتر آن موجود است. برای ارزیابی کلمات و جمله ها این دو مورد یکسان در نظر گرفته شده اند.

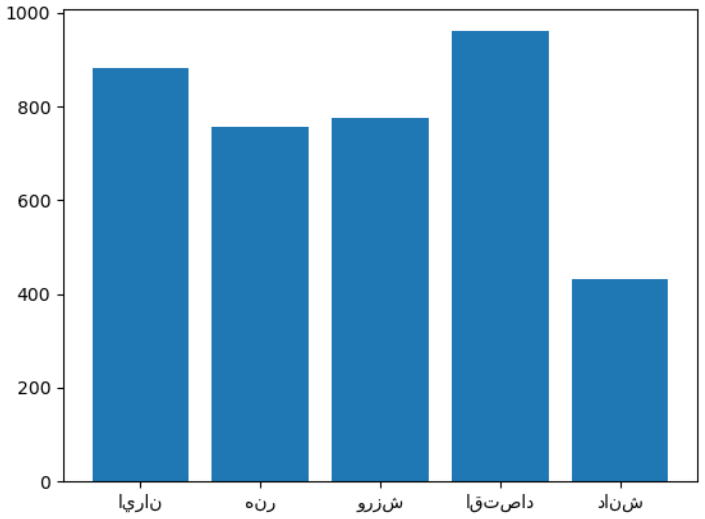
همان طور که در قسمت قبل تر اشاره شد هنگام حذف علائم نگارشی علامت نقطه که در زبان فارسی برای جداسازی جملات به کار میرود در متن نگهداری شده اند.

یکی از مزایای زبان فارسی به نسبت برخی زبان ها این است که بیشتر کلمات از یکدیگر با استفاده از فاصله میشوند. پس از حذف stop wordsها می توان کلمات را با استفاده از دستورهای زبان پایتون با فاصله از یکدیگر جدا کرد.
برای انجام داده های آماری ابتدا یک دیکشنری از تمام لغات موجود در دیتاست و تعداد تکرار آن ها تهیه کردم. با توجه به این که ۵ دسته از کلمات موجود هستند من به ازای هر کلاس دیکشنری لغات و تعداد تکرار آن ها را در فرمت فایل csv ذخیره کردم و یک فایل هم وجود دارد که تعداد تکرار کلمات در کل دیتاست preprocessed شده است.

تعداد داده

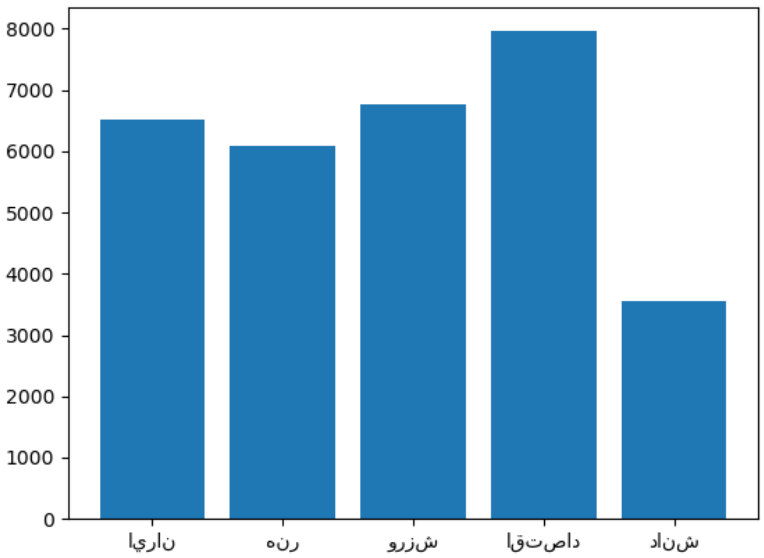
3807 تعداد کل اخبار استخراج شده.

داده ها به تفکیک هرکلاس به صورت زیر است.

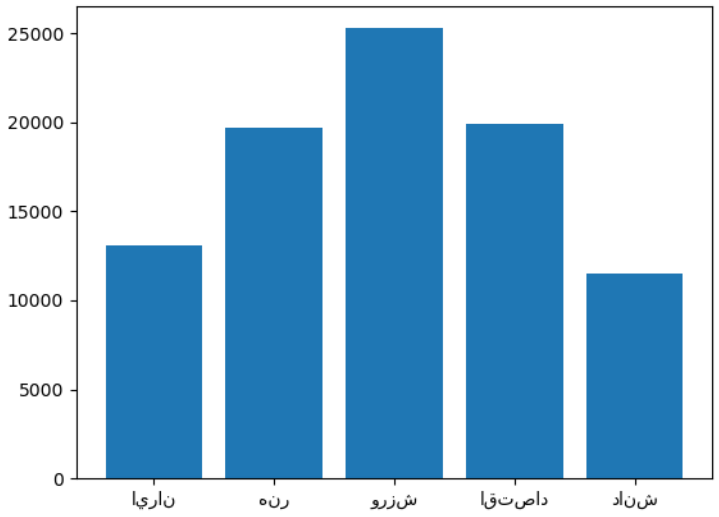


تعداد جملات

با توجه به آنچه در پیش پردازش انجام شد برای همسطح شدن اخبار ماکزیمم ۱۰ جمله اولیه هر خبر در دیتاست پیش پردازش شده موجود است. در نتیجه تعداد جملات حدودا مشابه هم هستند و مجموعا 30887 جمله در دیتاست موجود است.

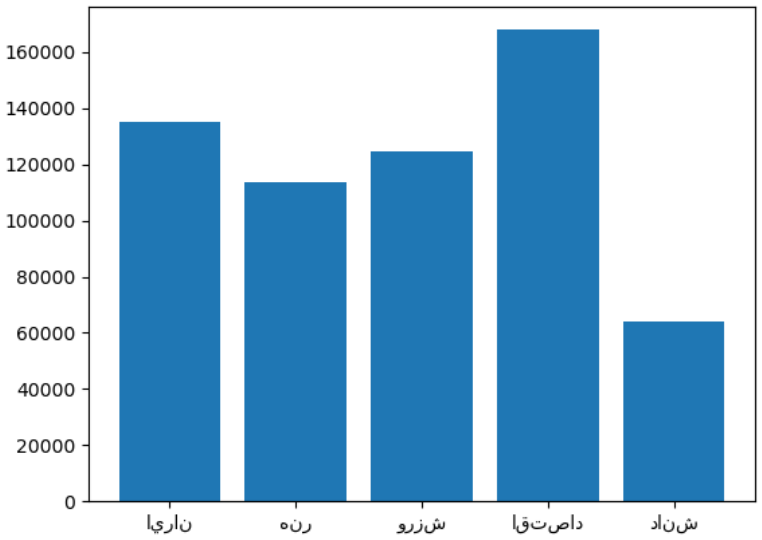


در صورتی که تعداد جملات برای دیتاست با دیتای خام را بررسی کنیم. جمله های هر کلاس به صورت زیر می شود. با توجه به این که تعداد داده ها در کلاس 'ورزش' تفاوت چندانی با دسته های ایران و هنر و اقتصاد ندارد اما تعداد جمله های در ورزش بیشتر است. میتوان نتیجه گرفته که اخبار ورزشی , اخبار طولانی تری به نسبت سایرین هستند.



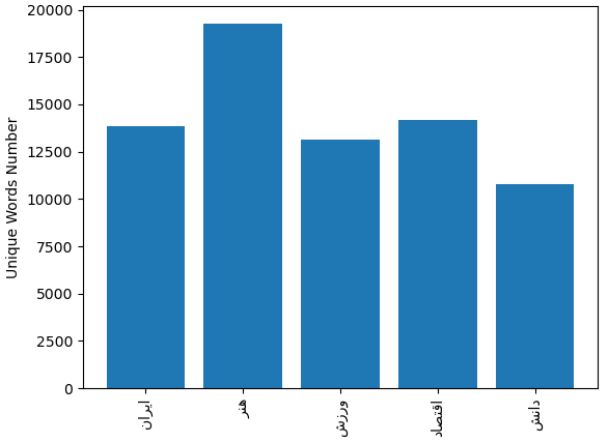
تعداد کلمات

تعداد کل کلمات برابر 605572 است. و تعداد کلمات در هر کلاس در نمودار زیر آمده است.



در جدول زیر تعداد کلمات هر کلاس به صورت دقیق بیان شده است.

تعداد کلمات منحصر به فرد

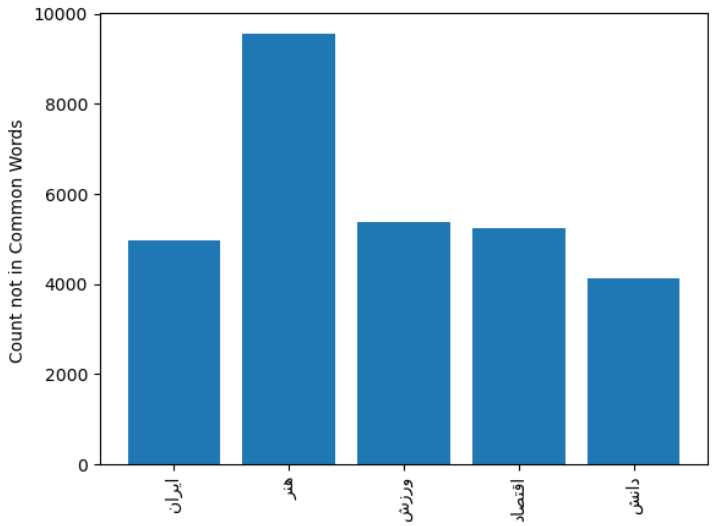


در جدول پیشرو خلاصه ای از چهار مورد بالا به دقت بیان شده است.

Label	ایران	هنر	ورزش	اقتصاد	دانش	مجموع
sents	883	756	776	961	431	30887
words	137648	115715	126999	171410	65158	605572
Unique words	13849	19256	13144	14184	10793	78431

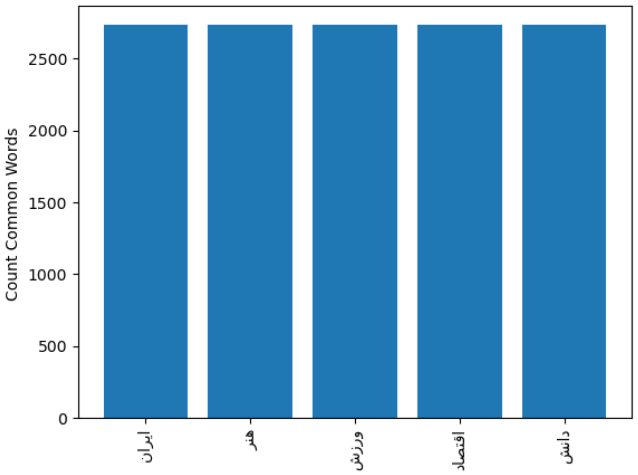
تعداد کلمات منحصر به فرد مشترک و غیرمشترک بین برچسپ‌ها

کلمات غیر مشترک



کلمات مشترک

با توجه به این که کلمات مشترک در تمامی کلاس ها درنظرگرفته شده اند. اندازه آن ها برای سه کلاس برابر است.



۱۰ کلمه پرتکرار غیر مشترک از هر برچسب

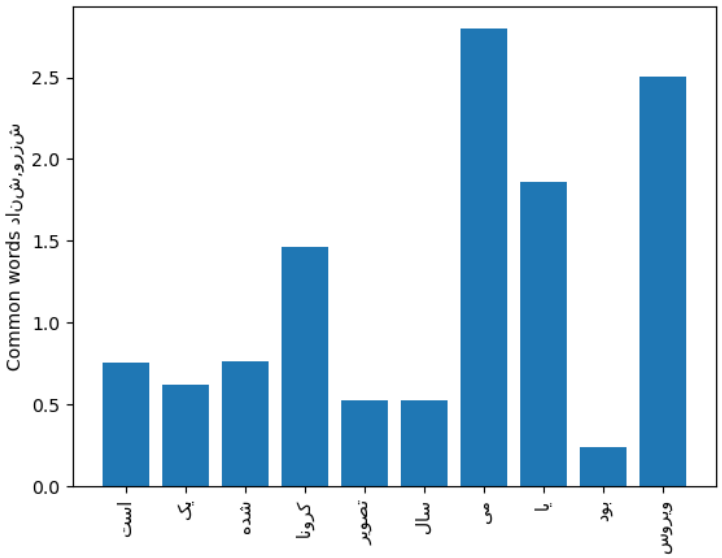
طلبان	اوین	نامداری	خاوران	فرونشست	خلوص	بورل	اطلاح	فایل	نوری	ایران

شجریان	خوانندگان	غزنی	منار	تاجیک	معارف	کانری	ولسوالی	اکران	پینگ	هنر
مربی ها	فدرر	ناتنهام	رونالدو	تورنومنت	مربیان	جوکوویچ	ویمبلدن	نادل	بازی	ورزش
واگن	صادرکننده	وانت بلند	ارزهای	شرکتهای	خودروسازان	لوکس	خودرو فولکس	خودروساز	شانسی	اقتصاد
کاوشگر	پادتن	پویایی	انتگرال	سفینه	سرفه	فضانوردان	والش	مرس	سارس	دانش

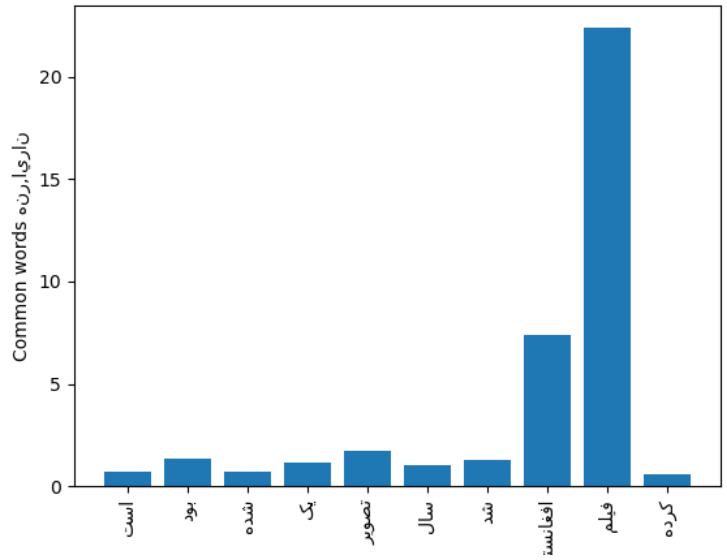
RelativeNormalizedFrequency

این پارامتر برای هر دو کلاس بررسی شده است. بعضی از نمونه های خروجی برای کلاس ها در ادامه آمده است سایرنتایج در گیت هاب موجود میباشند. برای مثال همان طور که در ادامه مشاهده میشود رابطه بین کلمات دانش و کلمات ورزشی بیشتر است از رابطه بین کلمات پر تکرار در هنور در برابر ایران یا اقتصاد.

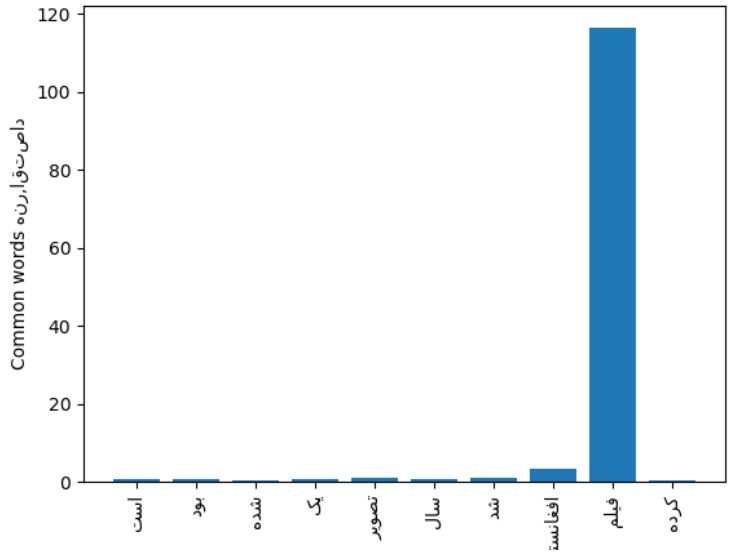
دانش-ورزش



هنر-ایران



هنر-اقتصاد

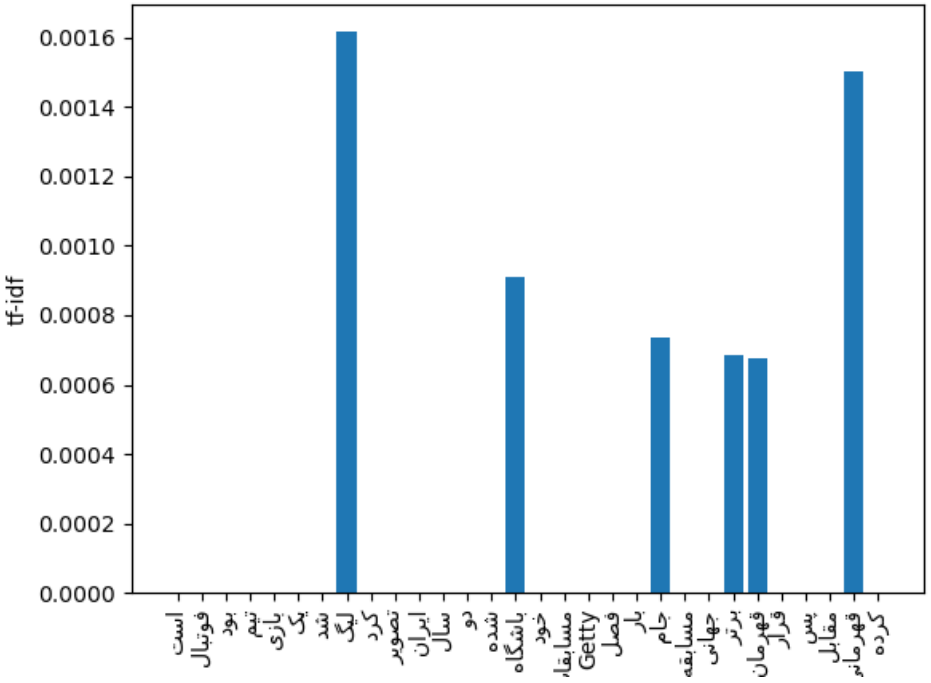


TF-IDF

با توجه به این که بسیاری از کلمات پر تکرار هرکلاس در کلاس های متفاوت مشترک هستند. در نتیجه idf که گلاریتم تعداد کل کلاس ها تقسیم تعداد کلاس هایی است که شامل این کلمه میباشند برابر صفر است. به همین دلیل tf-idf آن مقداری ندارد.

در ادامه نمونه ای از td-idf برای کلاس ورزشی آماده است. سایر تصاویر در گیتهاب موجود میباشند.

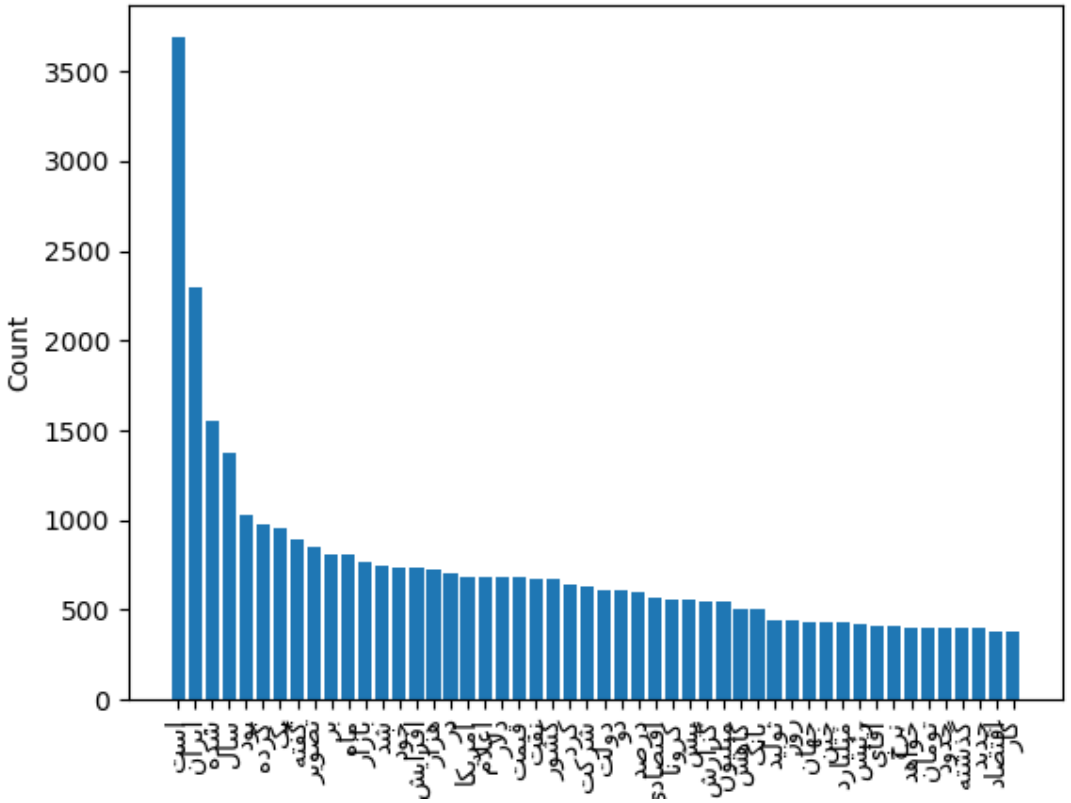
ورزشی

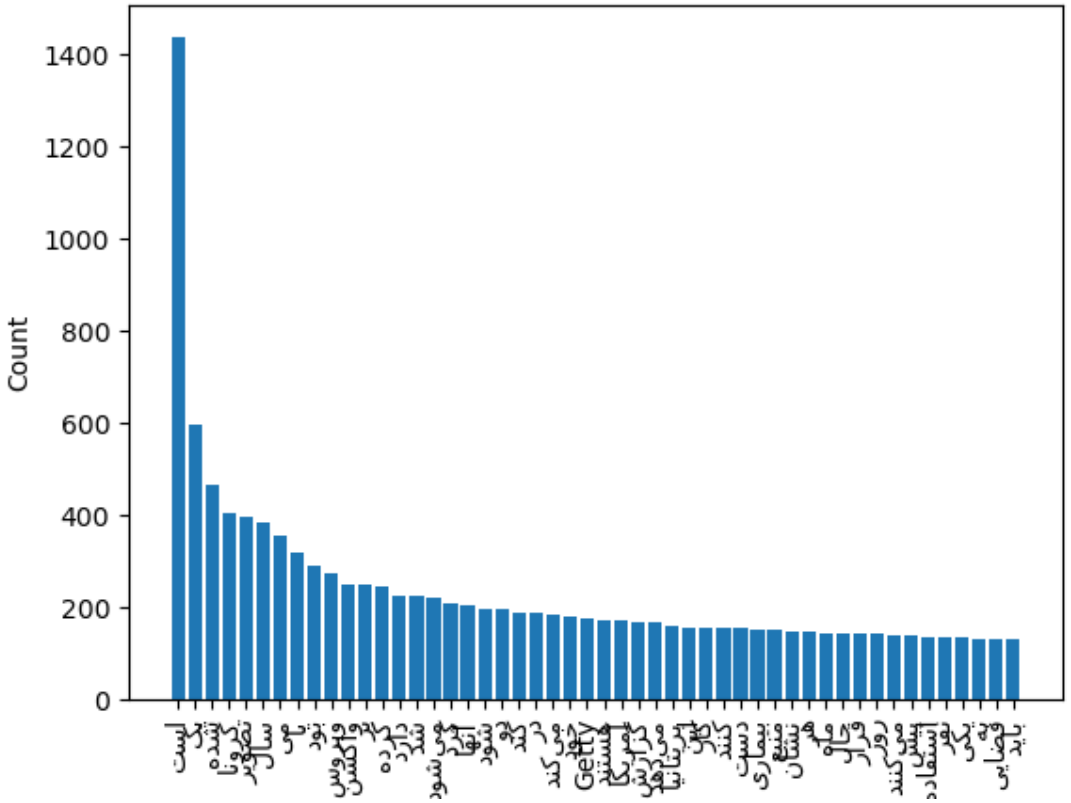


هیستوگرام

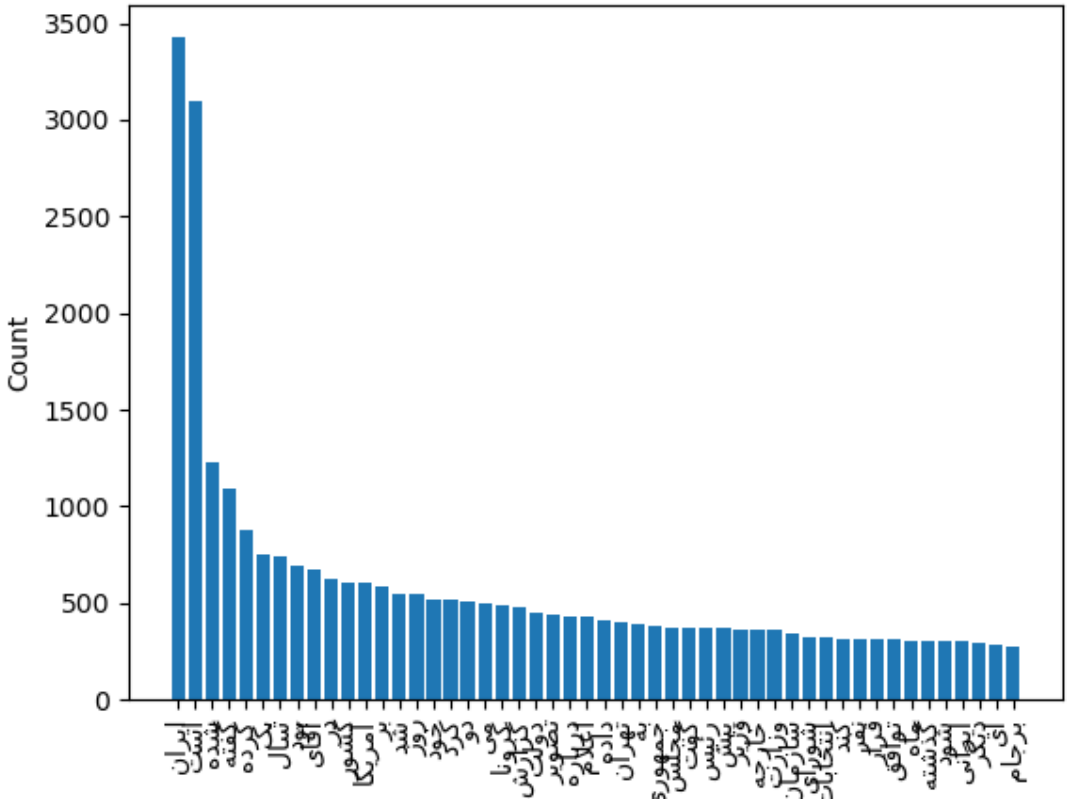
برای این که خوانایی بیشتری داشته باشد هیستوگرام کلمات در هر کلاس ۵۰ تایی که بیشترین تکرار را داشته اند آمده است.

اقتصاد

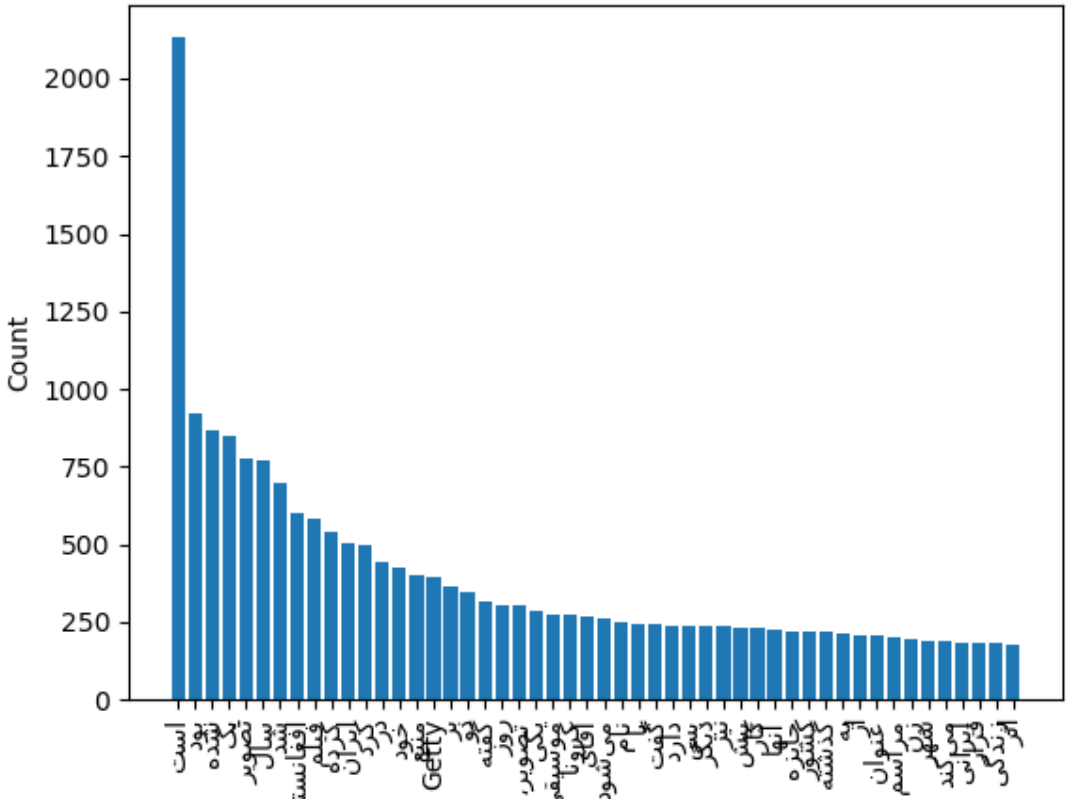



Count
0
200
400
600
800
1000
1200
1400

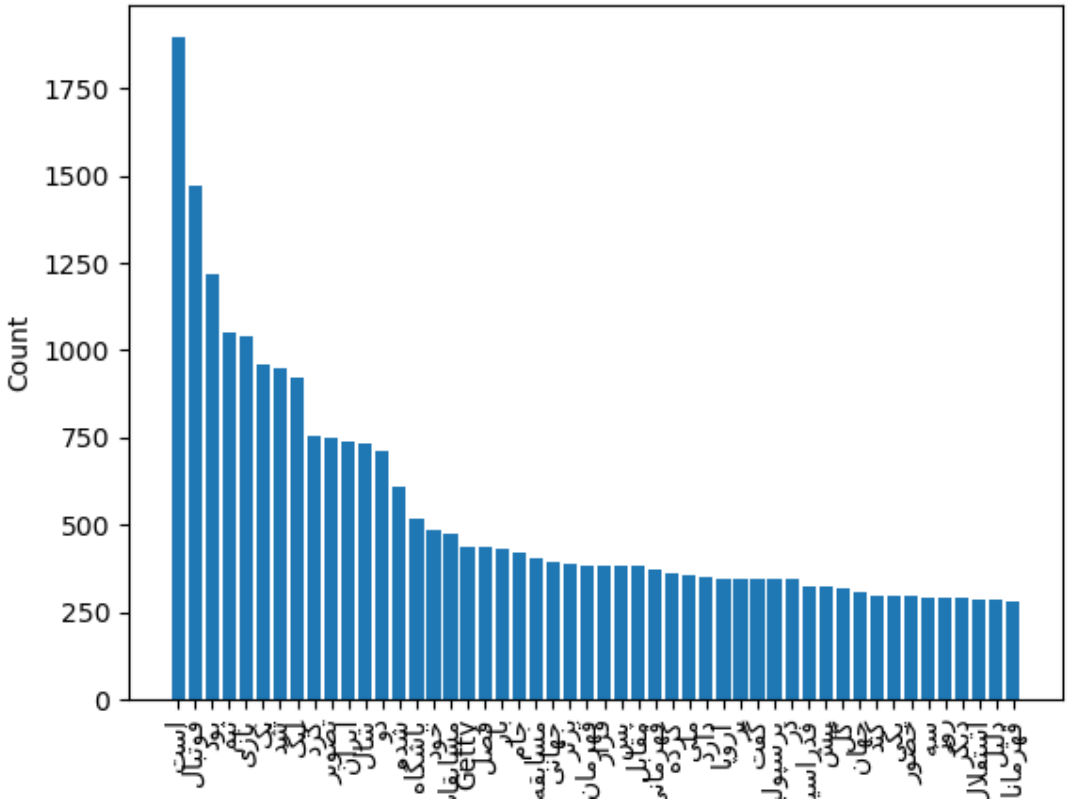
ایران



هنر



ورزش



اجرا

برای اجرای هر مرحله فایل های پایتون زیر اجرا شوند:

crawler فایل crawl.py

Preprocess فایل prepocess.py

Statistics فایل statistics.py

منابع:

[1] https://www.bbc.com/persian

[2] https://www.kaggle.com/c/learn-ai-bbc/data