گزارش پروژه ی دوم درس NLP فاز دوم

قسمت اول

الف. پیش پردازش ها

HTML cleanup:

در این مرحله داده های لازم را از آرشیو سایت خبرنگاران جوان جمع آوری کرده ام و چون در آرشیو، داده ها مرتب شده و منظم بود نیازی به روش های استخراج متن از وبسایت نبود و به طوری دستی از سایت داده ها را کپی کردم.

Sentence Breaking:

برای جدا کردن جمله ها، از تابع GetSentences در کلاس wordmap استفاده می کنم که در این تابع به کمک کلاسSentenceTokenizer از کتابخانه ی Nhazm جمله های متن را تشخیص داده و در data structure دیکشنری ذخیره می کنم و به عنوان خروجی تابع برمی گردانم.

Number/Date/... Handling and Tokenization:

در کلاس های wordmap و naiveBayes متدی را تعریف کردم که متن ورودی را بر اساس جدا کننده های زیر به صورت آرایه ای از tokenها در خروجی بر می گرداند.

```
char[] delimiters = new char[] {
                ',', '"', ')', '(', ';', '.', '\n', '\r', '\t',
                '>', '<', ':', '=', '\'', '[', ']',' ','!','#',
                '؛','،','-','$','%','&','\'', '+','/','?','؟',
                '{','}','»','«','1','2','3','4','5','6','7','8',
                '9','0','۱','۲', '۳','۴','۵','۶','۷','۸','۹','۰'};
```

Normalization:

در کلاس naiveBayes متد normalize را تعریف کردیم که در آن از کلاس normalizer کتابخانه ی Nhazm برای نرمال کردن متن ورودی استفاده می کنیم.

Test/Train Split up:

داده های استخراج شده از سایت برای خبر های ورزشی و سیاسی حدود 35 هزار کلمه بوده است که 10 در صد را برای تست از هر دو کلاس جداکردیم یعنی حدود 3500 کلمه که 10 فایل تست برای کلاس خبر های ورزشی و 10 فایل تست هم برای کلاس خبر های سیاسی ایجاد کرده ایم. برای هر یک از کلاس ها یک فایل train که حدود 32000 کلمه است را درنظر گرفته ایم. فایل ها در فولدر data قرار داده شده است.

ب. Naive Bayes Text Classifier:

کلاس بندی فایل های تست در تابع naivebayes و در متد classifier انجام می شود. به این صورت که Token های منحصر به فرد هر دو کلاس را از داده های train با متد uniqueToken بدست می آوریم و سپس لگاریتم احتمال هر کلمه را در هر یک از کلاس ها با smoothing با متدCalulateWordLogProbs بدست می آوریم. احتمال هر یک از کلاس ها یک دوم است زیرا تعداد فایل های train برای هر یک از کلاس ها یکی است. سپس هر یک از فایل های تست را می خوانیم و tokenize می کنیم و احتمال آن فایل تست را برای هر یک از کلاس ها محاسبه می کنیم. احتمال هر کدام از کلاس ها که بیشتر شد فایل تست متعلق به آن کلاس خواهد بود.

ج. محاسبه و گزارش Precision/Recall د. محاسبه و گزارش موثرترین feature ها

politicsTest1.txt is class 2
	Class1:-2475.77431052837	Class2:-2179.50119890199
	TopFeature: اقتصادی ملت استیضاح ارز دولت
politicsTest10.txt is class 2
	Class1:-3013.05512368508	Class2:-2689.60900532443
	TopFeature: ملت تولید دولت روحانی ظریف
politicsTest2.txt is class 2
	Class1:-2576.58276599254	Class2:-2364.96406167692
	TopFeature: شرق نیروهای ملت تولید دولت
politicsTest3.txt is class 2
	Class1:-2544.72954465364	Class2:-2248.3650599466
	TopFeature: اقتصادی ملت دولت روحانی برجام
politicsTest4.txt is class 2
	Class1:-2566.53301632354	Class2:-2305.7469719914
	TopFeature: ترامپ ارز دولت ظریف برجام
politicsTest5.txt is class 2
	Class1:-2610.86497305499	Class2:-2279.70995203301
	TopFeature: احزاب نظامی ملت دولت برجام
politicsTest6.txt is class 2
	Class1:-3092.69326852935	Class2:-2751.95122650628
	TopFeature: تولید دولت ارتش ظریف برجام
politicsTest7.txt is class 2
	Class1:-2742.34891260277	Class2:-2413.23990972845
	TopFeature: اقتصادی ارز تولید دولت ظریف
politicsTest8.txt is class 2
	Class1:-2790.75873235624	Class2:-2449.43317979359
	TopFeature: ترامپ تولید دولت ارتش ظریف
politicsTest9.txt is class 2
	Class1:-2686.65016884675	Class2:-2349.6608279241
	TopFeature: دولت روحانی ارتش ظریف برجام
sportsTest1.txt is class 1

Class1:-1771.73946702294 Class2:-2157.47379613525
TopFeature: فوتبال گل جام خلاصه لیگ
sportsTest10.txt is class 1
Class1:-1711.72609587551 Class2:-2039.42488371979
TopFeature: پرسپولیس کشتی فوتبال جام لیگ
sportsTest2.txt is class 1
Class1:-1673.66897340513 Class2:-1956.85290407381
TopFeature: شفر کی فوتبال گل لیگ
sportsTest3.txt is class 1
Class1:-1957.99027825152 Class2:-2234.7699220747
TopFeature: صعود شاگردان گل خلاصه لیگ
sportsTest4.txt is class 1
Class1:-1770.73493804825 Class2:-2084.98179946806
TopFeature: شفر کی فوتبال جام لیگ
sportsTest5.txt is class 1
Class1:-1620.74334633916 Class2:-1873.78135915537
TopFeature: صفر کشتی فوتبال صعود جام
sportsTest6.txt is class 1
Class1:-1590.18155395262 Class2:-1881.7446617745
TopFeature: ارتش صعود جام خلاصه لیگ
sportsTest7.txt is class 1
Class1:-1597.30245039159 Class2:-1887.59389530777
TopFeature: کشتی فوتبال صعود خلاصه لیگ
sportsTest8.txt is class 1
Class1:-1660.61560148293 Class2:-1918.58502623859
TopFeature: کشتی فوتبال گل خلاصه لیگ
sportsTest9.txt is class 1
Class1:-1814.8324673958 Class2:-2146.1205570694
TopFeature: فوتبال گل جام خلاصه لیگ

همان طور که نتایج را مشاهده می کنید تمامی تست فایل ها را بدرستی تشخیص داده است. بنابراین precision و recall برای هر دو کلاس برابر یک است.

قسمت دوم

الف. آماده سازی داده برای ورودی VW

برای اینکه داده ها را برای VW آماده کنیم در کلاس vowpalWabbit در متد convertToVW داده های train هر دو کلاس ورزشی و سیاسی را به صورت یک درمیان و همراه با label در یک فایل تست به نام inputVW.txt ذخیره می کنیم.

ب. تولید و اجرای دستور مناسب VW برای آموزش و امتحان.