پروژه ی واژه نما

موضوع پروژه ی من مقایسه ی اخبار ورزشی و سیاسی می باشد یعنی اینکه تشخیص بدهد که آیا خبر داده شده سیاسی است یا ورزشی. برای این منظور عنوان اخبار و خلاصه ای از اخبار ورزشی و سیاسی از تاریخ 19 اسفند تا 25 اسفند (به مدت یک هفته) را از سایت باشگاه خبرنگارن جوان تهیه کردم و به عنوان داده ورودی به برنامه داده ام.

ابتدا متن اخبار را خط به خط از فایل txt مورد نظر خوانده می شود و جمله های آن را به وسیله ی کتابخانه ی NHazm و کلاسSentenceTokenizer جدا می شود و همرمان جمله ها نیز نرمال می شود.

پس ازآن کلمات را از متن جدا کرده و با توجه به لیستی از حروف اضافه زبان فارسی وافعال زبان فارسی که از قبل تهیه کرده ام؛ حروف اضافه و ضمایر و فعل ها را از کلمات حذف می کنیم. سپس با استفاده از کلاس lemmatizer در کتابخانه NHazm ریشه کلمات را بدست آورده و کلماتی که دارای یک ریشه هستند را یکی در نظر می گیریم و سپس شروع به شمارش تعداد کلمات می کنیم. نتایج و آمار را در wwordmapدر برنامه گنجانده شده است.







