گزارش پروژه ی پایانی پردازش زبان و گفتار

موضوع این پروژه تحلیل توجه به حضور زنان در اخبار اجتماعی و اقتصادی است. برای این کار داده های مورد نیاز از وبسایت روزنامه ی همشهری دریافت شده اند. که دو مجموعه داده شامل اخبار اقتصادی و اجتماعی سال های 85-86 و سال های 95-96 بوده است. داده های مذکور از طریق کد datacoll.py دریافت گشته اند.

برای آموزش با توجه به آنکه باید از روش های بدون ناظر استفاده شود، از روش محاسبه ی پولاریتی استفاده شده است. به این صورت که یک بار برای هر کدام از دو دیتاست یک polarity متوسط محاسبه شده است. در این مرحله برای دو کلمه ی مردان و زنان برای دیتاست های قدیم و جدید به ترتیب polarity 36/0- و 84/0- به دست آمد. که به این معناست که به صورت کلی حضور زنان در اخبار کمتر شده است. (این خلاف انتظار بود. چنانکه با این انتظار پیش رفتیم که حضور زنان به مرور بیشتر شده است.)

در مرحله ی بعد این قابلیت وجود دارد که برای هر خبر جدید polarity را محاسبه کنیم. به این صورت که polarity عبارات مهم خبر را با توجه به تعداد تکرار در مجموعه ی داده محاسبه می کنیم. عبارات مهم در این پروژه که به زبان فارسی است شامل مضاف و مضاف الیه و صفت و موصوف می باشد. برای نرمال کردن، توکن کردن و تشخیص نقش کلمات درجمله از ماژول هضم در پایتون استفاده شده است.

|  |  |
| --- | --- |
| Polarity |  |
| 36/0- | سال های 85-86 |
| 84/0- | سال های 95-96 |

نکته: با توجه به آنکه در جلسات پایانی کلاس اشاره شده بود که هدف از این پروژه تحلیل احساسات می باشد. (و نه تشخیص) در این پروژه بنده زمانی برای تست در نظر نگرفته بودم. بعد از این که متوجه شدم احتیاج به گزارش نتایج تست می باشد، سعی به نمونه گیری و تست پروژه نمودم که به دلیل حجم بالای داده و کمبود وقت نتیجه ای برای گزارش ندارم. اما کد این قسمت زده شده و در فایل اصلی موجود است.