گزارش تحلیل احساسی

موضوع پروژه‌ی sentiment analysis من تحلیل توییت‌هایی است که با هشتگ worldcup# نوشته شده‌اند. برای گرفتن توییت‌ها از api توییتر با استفاده از ماژول tweepy در پایتون استفاده کردیم. یک نکته درباره‌ی این api این است که در صورت نداشتن توییت آیدی نمی‌توان توییت‌های قدیمی‌تر از ۱۰ روز پیش را fetch کرد به همین دلیل ۴۵۰۰۰ توییت از توییت‌های ۱۰ روز اخیر را گرفتیم (در هر روز ۵۰۰۰ توییت) و روی آن‌ها تحلیل احساسی را استفاده کردیم.
دیتا را با استفاده از اسکریپت data.py گرفتیم و آن را در داخل یک فایل csv با دو ستون یکی برای تاریخ و یکی برای متن توییت ریختیم.(worldCup.csv)

برای قسمت تحلیل احساسی ابتدا باید توییت را tokenize می‌کردیم برای این کار از word_tokenize در ماژول nltk استفاده کردیم و سپس آن‌ها را با استفاده از PorterStemmer در همین ماژول nltk نیز stem کردیم سپس برای هر کدام از کلمات به دست آمده یک polarity پیدا کردیم و برای به دست آوردن polarity کل جمله polarity همه‌ی کلمات موجود در جمله را با هم جمع کردیم؛ برای به دست آوردن polarity کلمات از دیتاست afinn استفاده کردیم. این دیتاست به این گونه است که به هر کلمه یک عدد از -۵ تا ۵ اختصاص می‌دهد و منفی بودن به معنای بد بودن کلمه است مثلا کلمه‌ی nigger عدد -۵ است کلمه‌ی outstanding عدد ۵+ است.
ما برای به دست آوردن polarity به این ترتیب عمل کردیم که ابتدا آیا notای در جمله وجود داشته است که قبل از آن "," یا "." نیامده باشد اگر وجود داشت polarity آن کلمه را برعکس کردیم و اگر not در جمله نبود ولی کلمه‌ی مورد نظر در دیتاست ما وجود نداشت نگاه می‌کنیم اگر در قبل یا بعد از این کلمه یک and آمده بود polarity کلمه برابر با polarity قبلی یا بعدی در نظر گرفته می‌شود و اگر not وجود نداشت و کلمه در دیتاست موجود بود نیز polarity آن را از دیتاست پیدا می‌کنیم و استفاده می‌کنیم و در غیر این صورت آن را صفر در نظر می‌گیریم.
برای به دست آوردن precision و recall از یک sample با ۱۰۰ عدد توییت استفاده کردیم، البته recall در این حالت معنی نمی‌دهد و فقط precision با معنا است.
در خروجی برنامه دو نمودار کشیده می‌شود که نمودار بالایی برای هر توییت کشیده شده است و نمودار پایینی میانگین polarity تمام توییت‌ها در آن روز است. محور x نمودار مربوط به روزها می‌باشد (۱۰ روز همانطور که گفتیم) و محور y مربوط به polarity می‌باشد. همانطور که می‌بینید تمامی میانگین‌های هر روز مثبت هستند و می‌توان گفت به طور میانگین افراد درباره‌ی worldCup توییت‌های خوشحال می‌گذارند چون بالاخره یک طرف بازی را برده‌اند و از وضع موجود راضی هستند و ظاهرا افراد خوشحال بیش‌تر توییت می‌گذارند و توییت‌ها در کل مثبت شده‌اند.

خروجی:
(precision = 0.9063643603782977)