گزارش کار تمرین عملی سری دوم درس پردازش زبان طبیعی

محمدرضا اصولی - 610395077

مراحل انجام کار:

ابتدا به دانلود corpusهای مورد نظر پرداختیم (movie-reviews و stopwords برای حذف این نوع کلمات از reviewها).
سپس برای به دست آوردن ویژگی‌های از روش bag of words استفاده کردیم، به این روش که ویژگی‌های هر review بر اساس کلمات موجود در آن شناخته می‌شود. پس ابتدا همه کلمات موجود در همه reviewها را استخراج کرده، سپس stopwords و علائم نگارشی را از آن حذف کرده (این کار باعث بالا رفتن دقت می‌شود، به این دلیل که این کلمات در تمامی متن‌ها موجود بوده و اطلاعات خاصی در مورد آن متن به ما نمی‌دهند، پس بهتر است که حذف شوند) و در نهایت 1000 کلمه‌ای که بیشتر از آن‌ها استفاده شده را به عنوان feature استخراج کردیم. (مقدار 1000 یه می‌تواند متغیر باشد ولی اینکه کلمات کم تکرار از ویژگی‌ها حذف شوند معمولا باعث بالا رفتن دقت در این روش می‌شود)
سپس برای هر review ویژگی‌هایش را استخراج کرده و در نهایت با روش k-fold داده‌ها را به 5 دسته تقسیم کرده و در هر مرحله یک دسته را به عنوان تست و 4 دسته دیگر را به عنوان داده آموزشی استفاده کردیم (پس 5 دسته‌بند متفاوت خواهیم داشت) که برای هر دسته‌بند مقادیر accuracy، recall، precision و F-measure را محاسبه کردیم.

تحلیل تاثیر پیش‌پردازش‌ها بر نتایج رده‌بندی:

همانطور که در بالا گفته شد حذف stopwordها و علائم نگارشی به علت نداشتن اطلاعات مضاعف باعث بالا رفتن مقدار دقت می‌شود.

بهترین رده‌بند، بعد از حذف	بهترین رده‌بند، قبل از حذف
os_precision: 0.846590909090 neg_precision: 0.78125 pos_recall: 0.752525252525 neg_recall: 0.866336633663 pos_f_measure: 0.796791443850 neg_f_measure: 0.821596244131 accuracy: 0.81	pos_precision: 0.7864583333333334 neg_precision: 0.7115384615384616 pos_recall: 0.7156398104265402 neg_recall: 0.783068783068783 pos_f_measure: 0.7493796526054591 neg_f_measure: 0.7455919395465995 accuracy: 0.7475

توضیح ویژگی‌ها و دلیل انتخاب آن‌ها:

در این روش یک ویژگی به معنی داشتن یا نداشتن یک کلمه خاص تلاقی می‌شود. این روش، روشی ساده اما کارآمد برای تحلیل احساسات در متون است، به این صورت که ترکیب کلمات با بارهای معنایی متفاوت می‌تواند تا حد خوبی نشان‌دهنده احساس متن باشد. همچنین صفاتی که معمولا در این چنین متن‌ها ذکر می‌شود نیز می‌تواند راهنمای خوبی برای تشخیص مثبت یا منفی بودن متن باشد (برای مثال کلماتی مانند عالی، افتضاح، زیبا و ... در زبان فارسی).

تحلیل نتایج رده‌بندی:

با توجه به نتایج استخراج شده می‌توان دید که دقت این روش چیزی در حدود 80 درصد برای تمامی حالات k-fold است که البته نداشتن تفاوت چندان در بین نتایج رده‌بندها قابل پیش‌بینی بود، زیرا در این تمرین ترتیب reviewها تفاوتی در تصمیم‌گیری برای رده‌بند ما ایجاد نمی‌کرد.

pos_precision: 0.815028901734 neg_precision: 0.757709251101 pos_recall: 0.719387755102 neg_recall: 0.843137254901 pos_f_measure: 0.764227642276 neg_f_measure: 0.798143851508 accuracy: 0.7825	pos_precision: 0.787234042553 neg_precision: 0.726415094339 pos_recall: 0.718446601941 neg_recall: 0.793814432989 pos_f_measure: 0.751269035532 neg_f_measure: 0.758620689655 accuracy: 0.755	pos_precision: 0.798969072164 neg_precision: 0.757281553398 pos_recall: 0.756097560975 neg_recall: 0.8 pos_f_measure: 0.776942355889 neg_f_measure: 0.778054862842 accuracy: 0.7775	pos_precision: 0.788359788359 neg_precision: 0.781990521327 pos_recall: 0.764102564102 neg_recall: 0.804878048780 pos_f_measure: 0.776041666666 neg_f_measure: 0.793269230769 accuracy: 0.785	pos_precision: 0.846590909090 neg_precision: 0.78125 pos_recall: 0.752525252525 neg_recall: 0.866336633663 pos_f_measure: 0.796791443850 neg_f_measure: 0.821596244131 accuracy: 0.81

همچنین در مورد موثرترین ویژگی‌های هر رده‌بند می‌توان به نتایج زیر اشاره کرد

ridiculous neg : pos 5.1 : 1.0 awful neg : pos 4.7 : 1.0 waste neg : pos 4.6 : 1.0 worst neg : pos 4.6 : 1.0 memorable pos : neg 4.3 : 1.0	waste neg : pos 5.3 : 1.0 ridiculous neg : pos 4.8 : 1.0 awful neg : pos 4.5 : 1.0 worst neg : pos 4.2 : 1.0 mess neg : pos 4.1 : 1.0	awful neg : pos 5.8 : 1.0 waste neg : pos 4.9 : 1.0 ridiculous neg : pos 4.3 : 1.0 dull neg : pos 4.1 : 1.0 worst neg : pos 3.9 : 1.0	waste neg : pos 4.9 : 1.0 awful neg : pos 4.8 : 1.0 ridiculous neg : pos 4.5 : 1.0 worst neg : pos 4.4 : 1.0 stupid neg : pos 4.3 : 1.0	awful neg : pos 7.6 : 1.0 ridiculous neg : pos 7.3 : 1.0 waste neg : pos 5.5 : 1.0 worst neg : pos 4.8 : 1.0 dull neg : pos 4.6 : 1.0

stupid neg : pos 3.9 : 1.0 dull neg : pos 3.7 : 1.0 subtle pos : neg 3.6 : 1.0 mess neg : pos 3.4 : 1.0 boring neg : pos 3.3 : 1.0	dull neg : pos 3.9 : 1.0 boring neg : pos 3.8 : 1.0 stupid neg : pos 3.8 : 1.0 spielberg pos : neg 3.7 : 1.0 excellent pos : neg 3.7 : 1.0	subtle pos : neg 3.8 : 1.0 terrible neg : pos 3.7 : 1.0 stupid neg : pos 3.7 : 1.0 memorable pos : neg 3.6 : 1.0 excellent pos : neg 3.6 : 1.0	mess neg : pos 3.8 : 1.0 boring neg : pos 3.8 : 1.0 memorable pos : neg 3.6 : 1.0 dull neg : pos 3.5 : 1.0 realistic pos : neg 3.4 : 1.0	mess neg : pos 4.2 : 1.0 stupid neg : pos 4.0 : 1.0 memorable pos : neg 3.6 : 1.0 perfectly pos : neg 3.5 : 1.0 subtle pos : neg 3.4 : 1.0

که به وضوح میزان تاثیرگذاری هر کلمه روی منفی یا مثبت بودن آن نظر مشهود است (علاوه بر کلماتی که به طور کلی صفت‌های مثبت یا منفی هستند، در این لیست کلماتی مانند اسپیلبرگ را مشاهده می‌کنیم که باعث مثبت بودن نظر کاربر شده و دلیل آن محبوبیت این کارگردان در بین مردم است.)