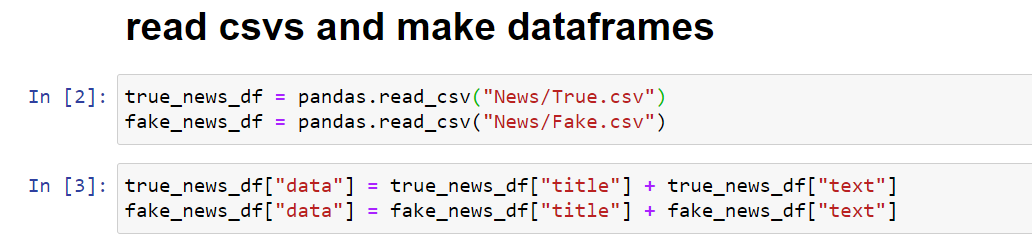
# تمرین ششم پردازش زبان‌های طبیعی

# امیررضا صدیقین

# 993614024

# مراحل 1تا5

در این بخش فایل‌های csv به وسیله‌ی کتابخانه‌ی pandas خوانده شده و داخل یک دیتافریم ریخته شده. همچنین ستونی به اسم data ساخته شد که از بهم پیوستن title و text هر خبر بهم تشکیل شده است.



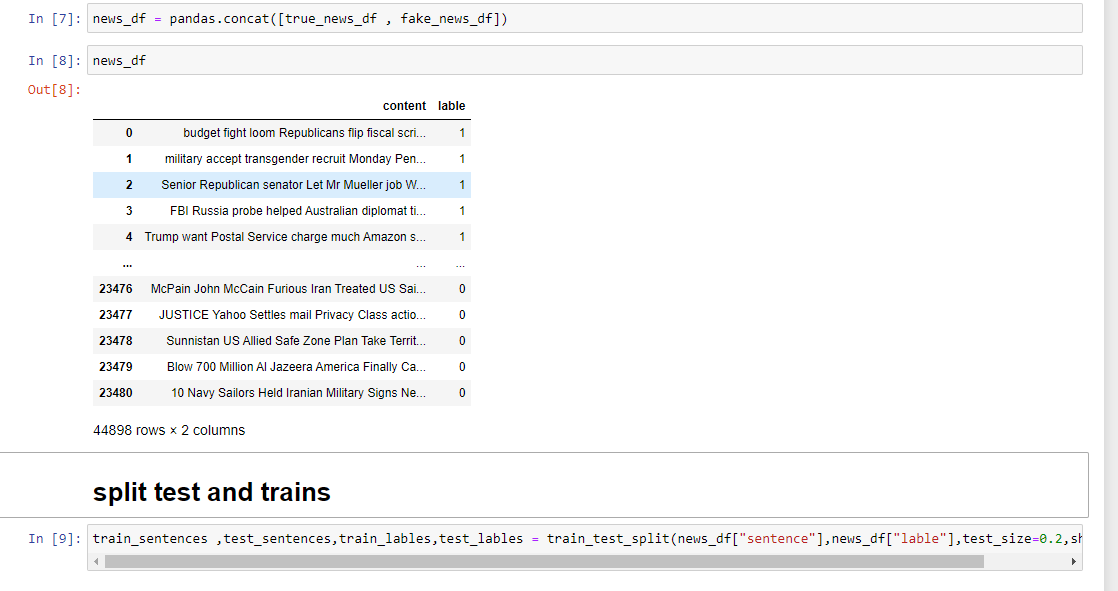
سپس هر داده پیش پردازش شده علائم نگارشی و stop\_words و کلمات تک کلمه‌ای حذف شدند و عملیات lemmatize برروی کلمات اعمال شد و متن‌های پیش‌پردازش شده به دست آمده است. برای‌هر یک از داده‌ی درست و داده‌های غلط به همراه برچسب دیتافریمی تشکیل شده است.

برچسب 1 برای داده‌های true\_news و برچسب 0 برای داده‌های fake\_news.

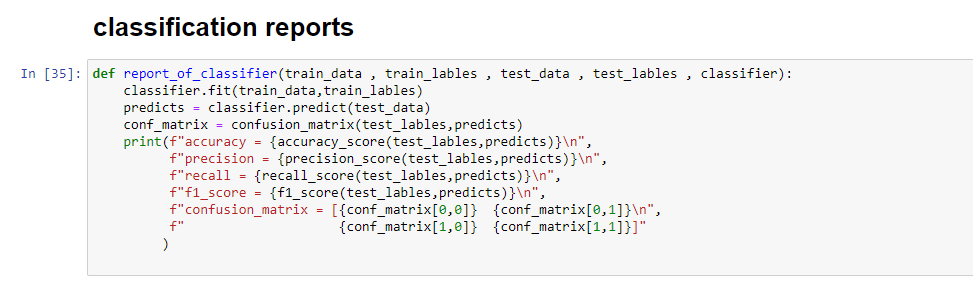


# مراحل 6 تا 13

برای استفاده ‌از داده‌ها داده‌های درست و غلط را با هم ادغام شده و سپس از روی‌ آن‌ها داده‌های آموزش و تست جدا شده است. (20 درصد داده‌ها برای تست و مابقی برای آموزش است.)

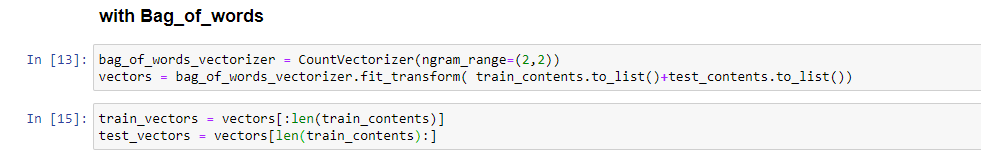


تابعی نوشته شد که الگوریتم دسته‌بندی و داده‌های آموزش و آزمایش را گرفته و مدل را با استفاده از داده‌های آموزشی آموزش می‌دهد و برچسب داده‌های آزمایش را پیش‌بینی می‌کند و صحت، دقت، بازیابی، f1\_score و ماتریس درهم‌ریختگی را نمایش می‌دهد.

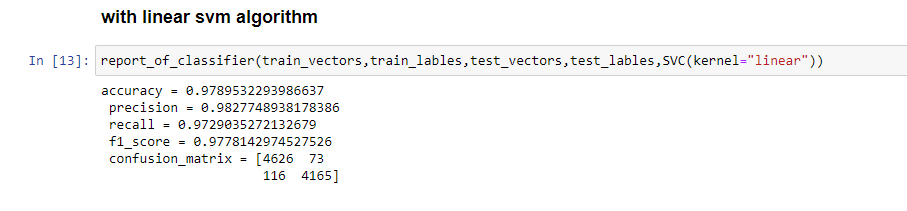


با استفاده از روش bag\_of\_words جملات ( هم آموزش ،هم آزمایش) به بردار‌ها تبدیل شده است. (برای تبدیل جملات به بردار باید همزمان هم داده‌های آموزش هم آزمایش به الگوریتم‌داده شود و سپس باز به دو دسته‌ی آموزش و آزمایش تبدیل شود.)

در روش bag\_of\_words به ازای هر مدل از دنباله‌کلمه (مثل unigram یا bigram که در ورودی کلاس تعیین می‌شود) یک ستون و برای هر متن بودن و نبودن آن با 0 و 1 نمایش داده‌میشود.

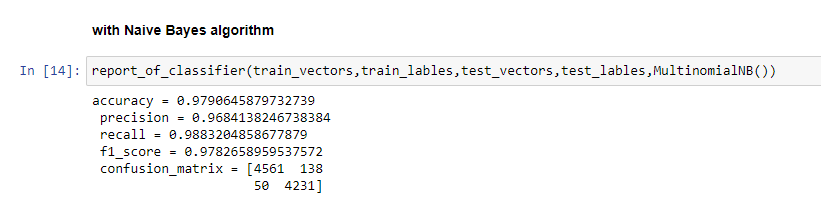


با استفاده از الگوریتم SVM خطی و روش bag\_of\_words عملیات آموزش و آزمایش انجام شد که به صورت زیر است.



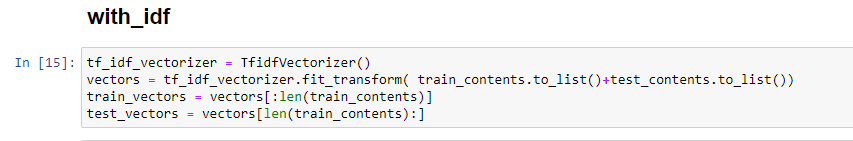
که نشان می‌دهد دقت خوبی داشته است و همه‌ی scoreهای آن در یک حدود است که بیان می‌کند در تشخیص اخبار درست و غلط به خوبی عمل کرده و همچنین با یک هسته‌ی خطی به خوبی عمل کرده است. این الگوریتم بسیار زمانبر است.

همچنین با استفاده از الگوریتم naïve\_base مدل MultinomialNB و روش bag\_of\_words عملیات آموزش و آزمایش انجام شد که به صورت زیر است.

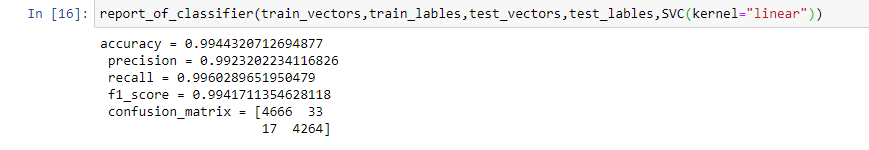


در این روش نیز عملکرد خوبی داشته است که در تشخیص اخبار درست بهتر عمل کرده است ولی در تشخیص اخبار غلط مقدار خیلی ناچیزی ضعیف‌تر از SVM است. (البته به خاطر تصادفی بودن تقسیم بندی‌ داده‌های آزمایش و آموزش این مقادیر را می‌توان صرف نظر کرد.) این الگوریتم بسیار سریع‌تر از الگوریتم SVM است و کمی نتیجه‌ی بهتری نسبت به SVM نیز داشته است.

با استفاده از روش tf\_idf جملات ( هم آموزش ،هم آزمایش) به بردار‌ها تبدیل شده است. (برای تبدیل جملات به بردار باید همزمان هم داده‌های آموزش هم آزمایش به الگوریتم‌داده شود و سپس باز به دو دسته‌ی آموزش و آزمایش تبدیل شود.) در این روش برای هر کلمه میزان تکرار در آن متن نگه داشته می‌شود. البته می‌توان وزن هر کلمه نیز مشخص شود که برای هر کلمه متناسب با میزان کلیدی بودن آن وزن تخصیص داده شود.

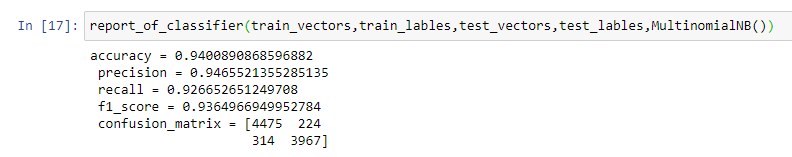


با استفاده از الگوریتم SVM خطی و روش tf\_idf عملیات آموزش و آزمایش انجام شد که به صورت زیر است.



در این روش الگوریتم SVM خطی عمکرد فوق العاده‌ای داشته است. به دلیل آن که بردار‌ها بعد زیادی دارند و دارای مقادیر غیر صفر و یک هستند این الگوریتم توانسته به دسته‌ها را از هم تمایز دهد و ساختار با هسته‌ی خطی داشته باشد. ولی همچنان این الگوریتم کند است.

با استفاده از الگوریتم naïve\_base مدل MultinomialNB و روش tf\_idf عملیات آموزش و آزمایش انجام شد که به صورت زیر است.



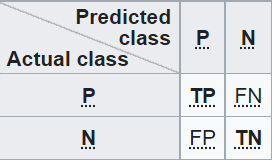
در این الگوریتم بخاطر سادگی بسیار سریع است و دقت خیلی خوبی داشته است ولی به دلیل تنوع در مقادیر مولفه‌های بردار‌ها نتوانسته به خوبی SVM عمل کند.

پس به ترتیب عملکرد به صورت زیر است. (همه‌ی دسته‌بند‌ها در یک حدود عملکردند ولی هر کدام در بخشی بهترند که در برایند کلی و صحت به صورت زیر هستند.

* SVM خطی با روش tf\_idf
* Niave bayes با روش BOW
* SVM با روش BOW
* Nive bayes با روش tf\_idf

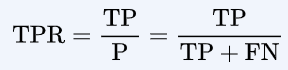
# روابط بین معیار‌ها:

ماتریس confusion به صورت زیرتعریف می‌شود که در آن P (positive) برای برچسب 1 و N (negative) برای برچسب 0 است.



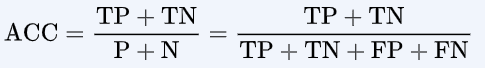
همچنین recall از فرمول زیر بدست می‌آید:

(میزان درستی پیش‌بینی کلاس Positive)



Accuracy نیز به صورت زیر است.

(میزان درستی پیش‌بینی هر داده)



Precision نیز به صورت زیر است.

(میزان درستی حدس زدن کلاس Positive در بین همه‌ی حدس‌های کلاس Positive)

