بسمه تعالی

Artificial Intelligence

Assignment 3

Mohammad Mahdi Islami
810195548

چارچوب کلی

ابتدا داده هارا به حالت دیتافریم با استفاده از کتابخانه Pandas تبدیل میکنیم. سپس مصرع های مربوط به هر کدام از شاعر ها را جدا میکنیم. و احتمال این که یک مصرع شاعرش حافظ باشد یا سعدی باشد را حساب میکنیم.P(c)
لغات این مصرع ها را جدا کرده و داخل یک دیتافریم نگه میداریم. سپس تعداد هر کدام از آن هارا حساب کرده و احتمال هر کدم به شرط اینکه بدانیم شاعر آن کیست را حساب میکنیم P(xlc)
حال برای این هر مصرع جدید بررسی میکنیم که آیا لغات این مصرع در داده های ما قرار دارد یا خیر.

۱-اگر موجود در تنها در اشعار حافظ موجود بود: احتمال لغت در سعدی را برابر صفر قرار میدهیم و بالعکس.
۲-اگر در هر دو موجود بود : احتمال همه ی لغات را در هم ضرب کرده و نگهداری میکنیم و در نهایت در P(c) ضرب میکنیم.
۳-اگر در هیچکدام موجود نبود آن را در نظر نمیگیریم.

$$P(c\,|\,x) = P(x\,|\,c) * P(c)$$

$$P(x\,|\,c) = P(x_1\,|\,c) * P(x_2\,|\,c) * \ldots * P(x_n\,|\,c)$$

پس از این که برای هر شاعر این این مقدار را حساب کردیم آن با یکدیگر مقایسه میکنیم. هر کدام بزرگتر بود آن مصرع را به آن شاعر نسبت میدهیم. در نهایت این مجموعه جواب را با مجموعه واقعی که ۲۰ درصد داده های انتهایی داده های train میباشد را مقایسه کرده و مقادیر خواسته شده زیر را محاسبه میکنیم:

$Accuracy = 74.5$
$Saadi.Recall = 80.8$ $Saadi.Precision = 76.14$
$Hafez.Recall = 64.5$ $Hafez.Precision = 70$

```
0.7407230069427819  :Accuracy
0.7000634115409005  :hafez Precision
0.6441073512252042  :Hafez Recall
0.7653846153846153  :Saadi Precision
0.8079577750710516  :   Saadi Recall
```

سوال‌های اضافی

۱- Precision دقت یک عامل را محاسبه میکند. این در حالیست که ما باید به دنبال میانگینی از دقت همه ی عامل ها باشیم تا بتوانیم دقت کل ماشین خود را مورد ارزیابی قرار دهیم. اگر تنها برای حافظ جواب های صحیح بدهد اما برای سعدی اینطور نباشد نمیتوانیم از صحت کارکرد ماشین خود اطمینان حاصل کنیم و باید مجموع سیستم را در نظر بگیریم.

یک سیستم چند عاملی. برای مثال به جز حالت دوتایی حافظ و سعدی شعرای دیگر نیز وجود داشتند و اگر قرار بود تنها دقت تشخیص سعدی بالا باشد اما برای شعرای دیگر جواب های غلط و درهم بدهد این سیستم سالم نیست و باید میانگین دقت همه ی این شعرا سنجیده شود و عوامل دیگر دخیل شوند.

۲- Accuracy برای این اگر تعداد یک عامل بسیار زیاد تر از عامل دیگر باشد این ماشین میتواند به دقت بالایی برسد بدون نیاز به تحلیل. فرض کنید از ۲۰۰۰۰ مصرع موجود ۱۹۰۰۰ مصرع متعلق به سعدی باشد و ۱۰۰۰ بیت متعلق به حافظ. مسلما ضریب تشخیص سعدی بالاتر میرود. برای ما مهم است که ۱۰۰۰ مصرع حافظ هم به درستی تشخیص داده شود و از دقت بالایی برخوردار باشد و بدانیم که آیا درست تشخیص داده شده است یا نه. اگر از ۱۹۰۰۰ مصرع سعدی ۱۴۰۰۰ را درست تشخیص دهد و از مصرع های حافظ ۵۰۰ تا را درست تشخیص دهد دقت ماشین حدود ۷۰ درصد است . فلذا اگر تنها ماشینی در نظر بگیریم که بر اساس تعداد قضاوت کند و بیشتر مصرع هارا به سعدی و مقدار کمی را حافظ تشخیص دهد نهایت دقت بالایی دارد چون نسبت مصرع های سعدی به حافظ بسیار زیاد است و تعداد درست ها به تعداد کل نسبت قابل ملاحظه ای میباشد. (حدود ۷۰ درصد) اما میدانیم تشخیص داده های حافظ از دقت خوبی برخوردار نیست(حدود ۵۰ درصد).
بنابراین باید برای تک تک عوامل دقتی در نظر بگیریم تا از صحت عملکرد اطمینان حاصل کنیم.

لاپلاس

برای این امر در قسمت Naive Base اگر داده در دایره لغات حافظ نبود اما در لغات سعدی یافت میشد احتمال آن داده را برای حافظ برابر صفر قرار میدادیم و در نهایت حاصل ضرب احتمال ها برابر صفر میشد.

اگر داده در هیچکدام از ابیات نبود یا لغت آ در کلمات سعدی نبود و لغت ب در لغات حافظ نبود باز حاصل این احتمال مقدار صفر می‌گرفت. ولو شواهد زیادی اختصاص این مصرع را به شاعری تایید کند.

برای حل این مساله راهکار پیشنهادی این است که اگر داده ای در هیچکدام از ابیات نبود آن را کلا در نظر نگیریم و در احتمال هیچکدام از شاعر ها آن را دخیل نکنیم که دقت کار بالاتر رود و در واقع این شانس را به لغات دیگر بدهیم تا اگر احتمالشان بالاتر رفت با یکدیگر مقایسه شوند و شواهد دیگر را دخیل کنیم.

دقت محاسبه بیز:

```
0.7407230069427819  :Accuracy
0.7000634115409005  :hafez Precision
0.6441073512252042  :Hafez Recall
0.7653846153846153  :Saadi Precision
0.8079577750710516  :   Saadi Recall
```

دقت محاسبه راهکار لاپلاس:

```
0.764184821642327   :Accuracy
0.7093624353819644  :hafez Precision
0.720536756126021   : Hafez Recall
0.8033661740558292  :Saadi Precision
0.7945594803085668$ :Saadi Recall
```