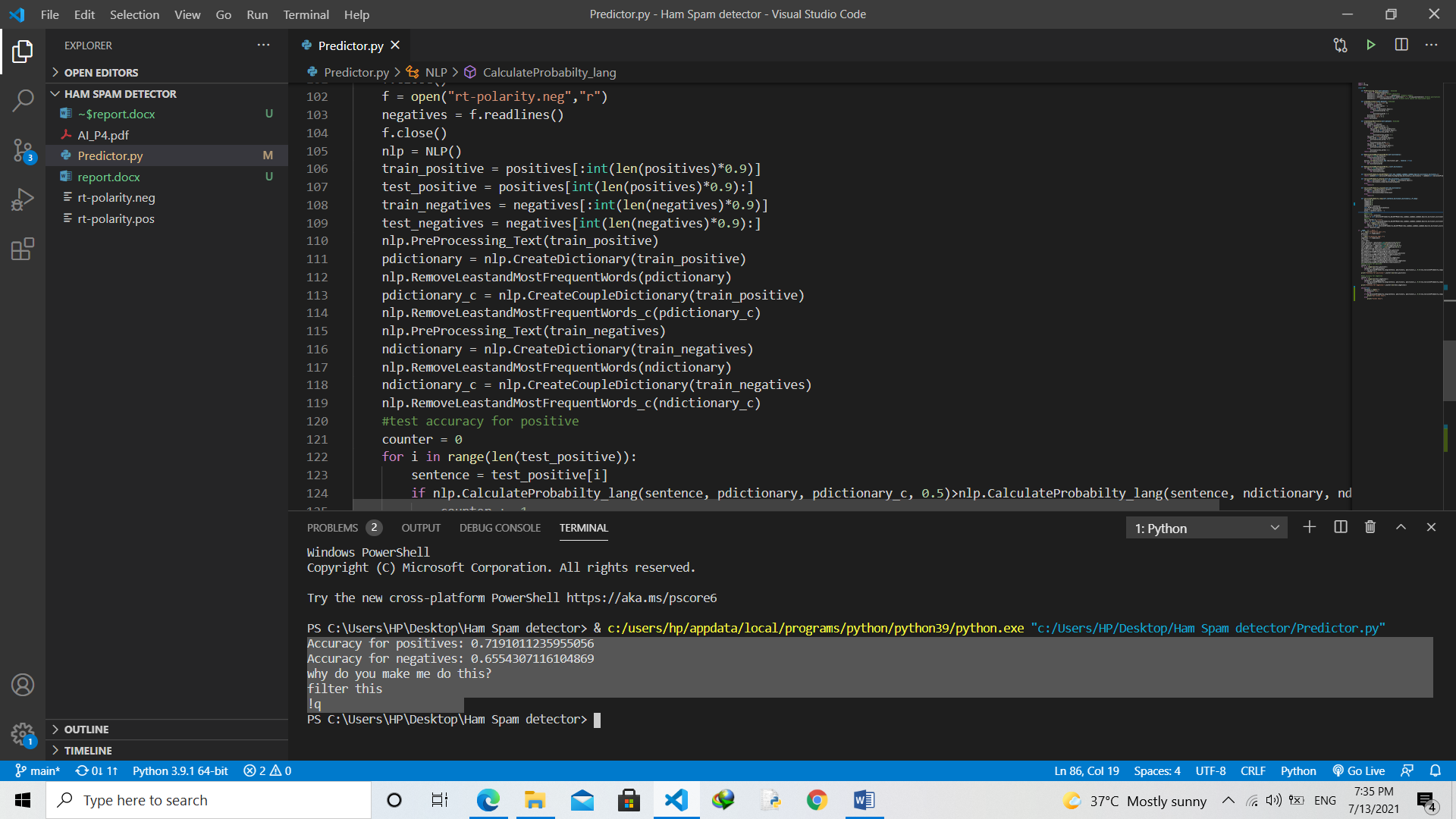
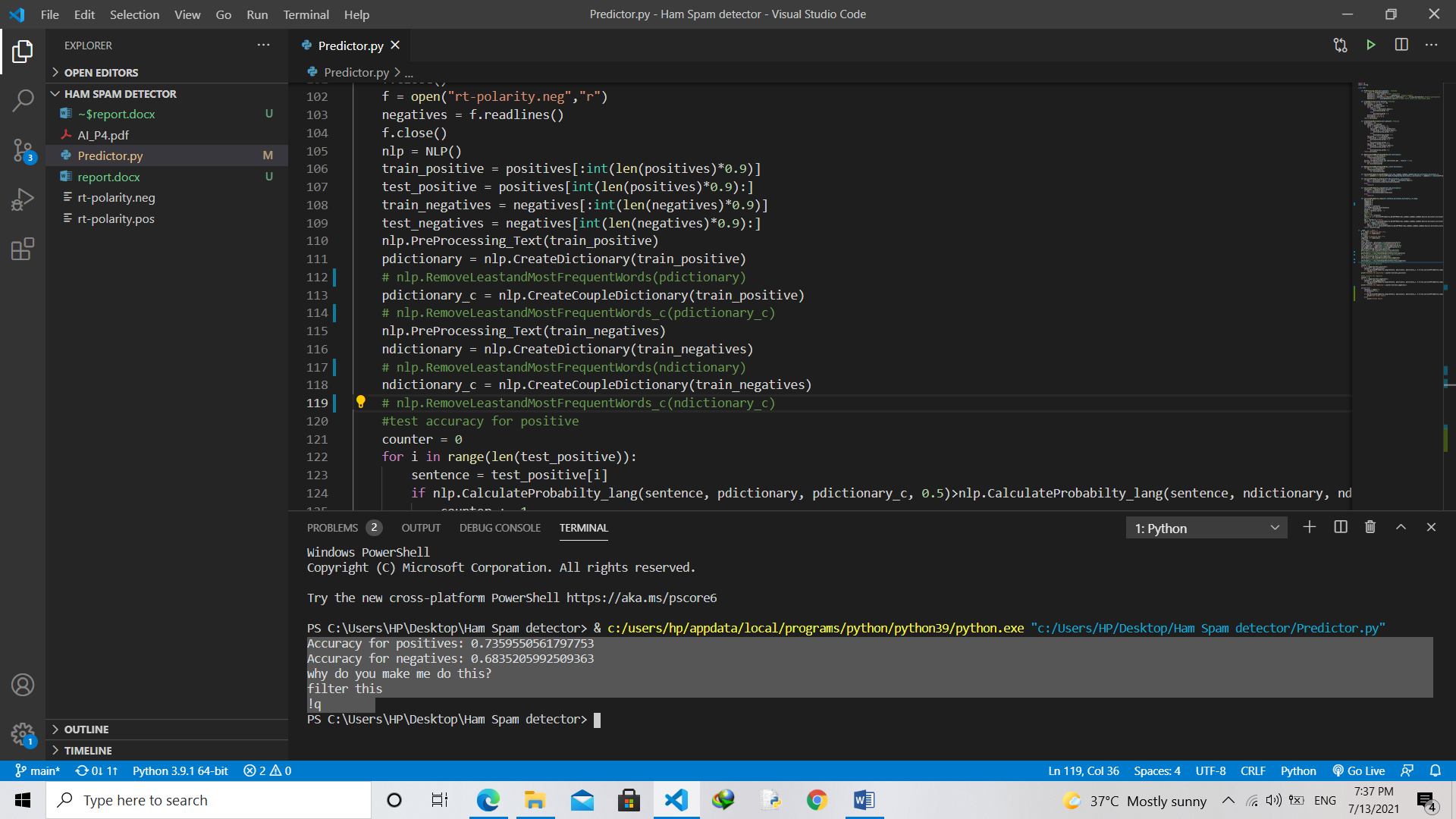
برای بررسی تاثیر کلمات پرتکرار و کم تکرار 3 سناریو را اجرا کردیم:

سناریو 1 : حذف کلمات کم تکرار و پرتکرار برای هر دو دیتاست :



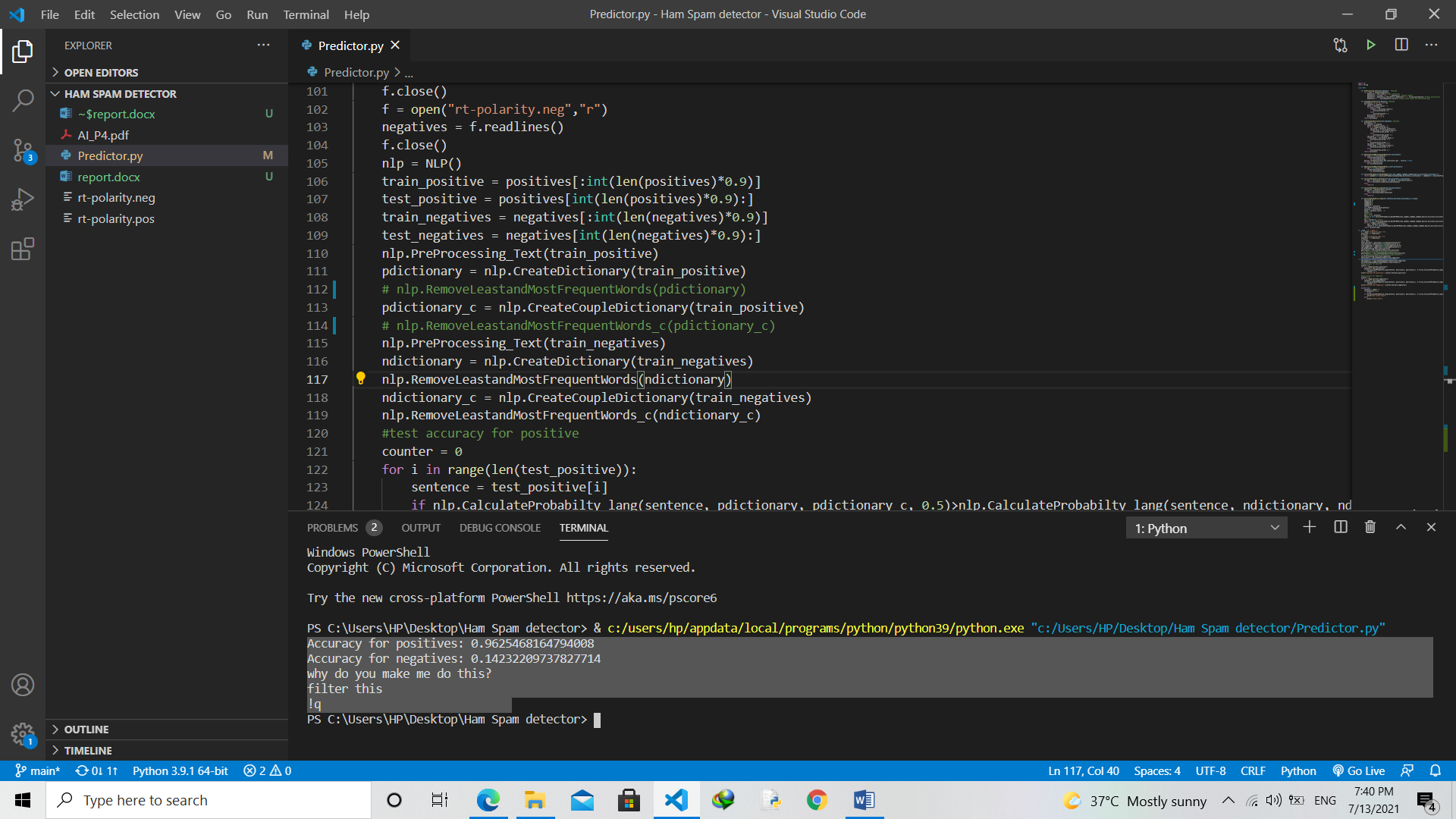
در این سناریو به صحت 72 درصد برای نظرات مثبت و 65 درصد برای کلمات منفی رسیده ایم.

سناریو دوم : حذف نکردن کلمات کم تکرار و پرتکرار برای هر دو دیتاست:



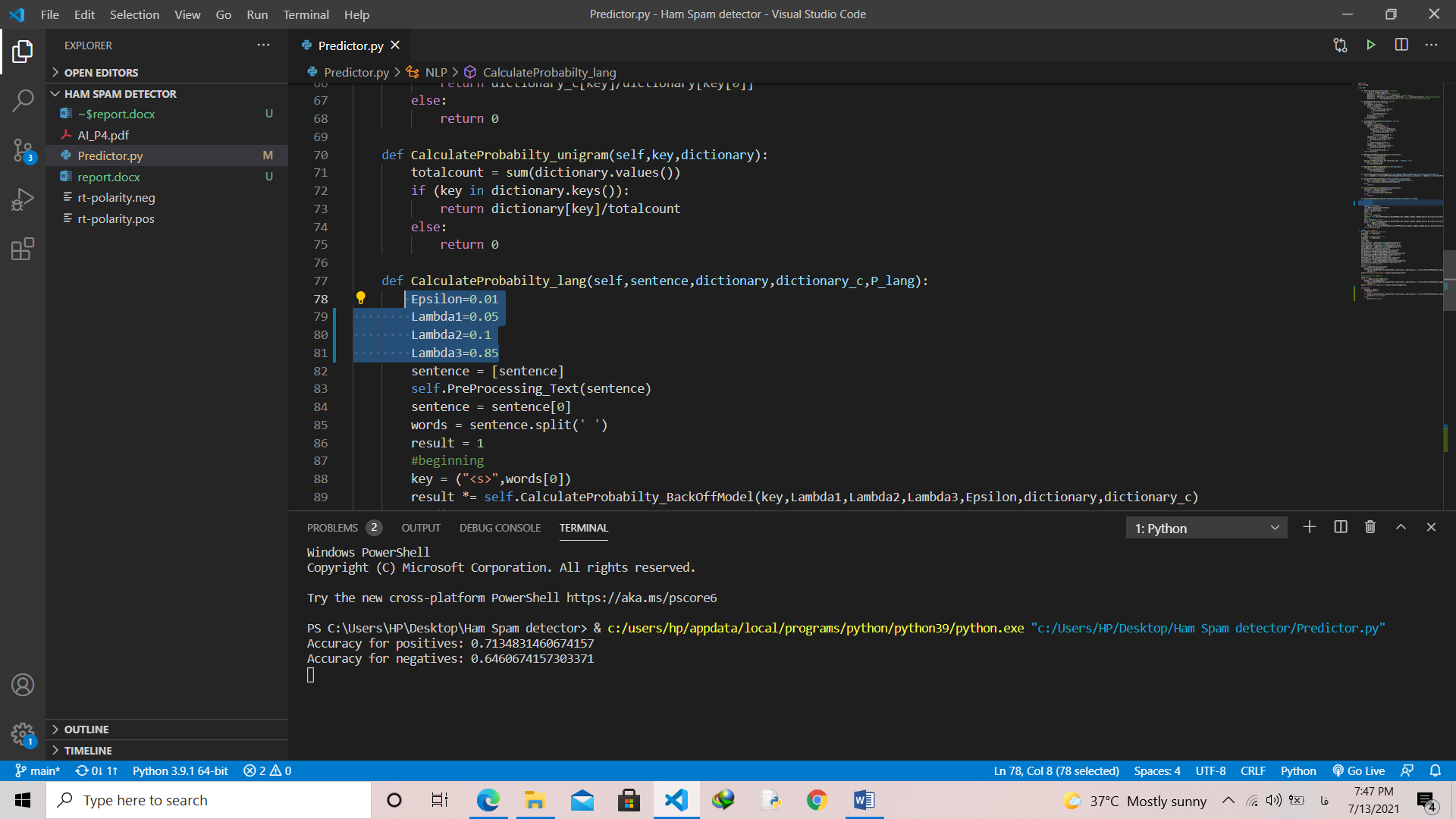
همانطور که مشخص است دقت ما از 72 به 73 و از 65 به 68 رسیده است. دقت ما بیشتر شده است اما از طرفی برنامه بسیار کند تر اجرا می شود و دقت اضافه شده مقدار چشم گیری نیست در نتیجه متوجه می شویم کلمات پرتکرار و کم تکرار تاثیر زیادی بر یادگیری ندارند .

سناریو سوم : حذف تنها برای یکی از دیتاست ها

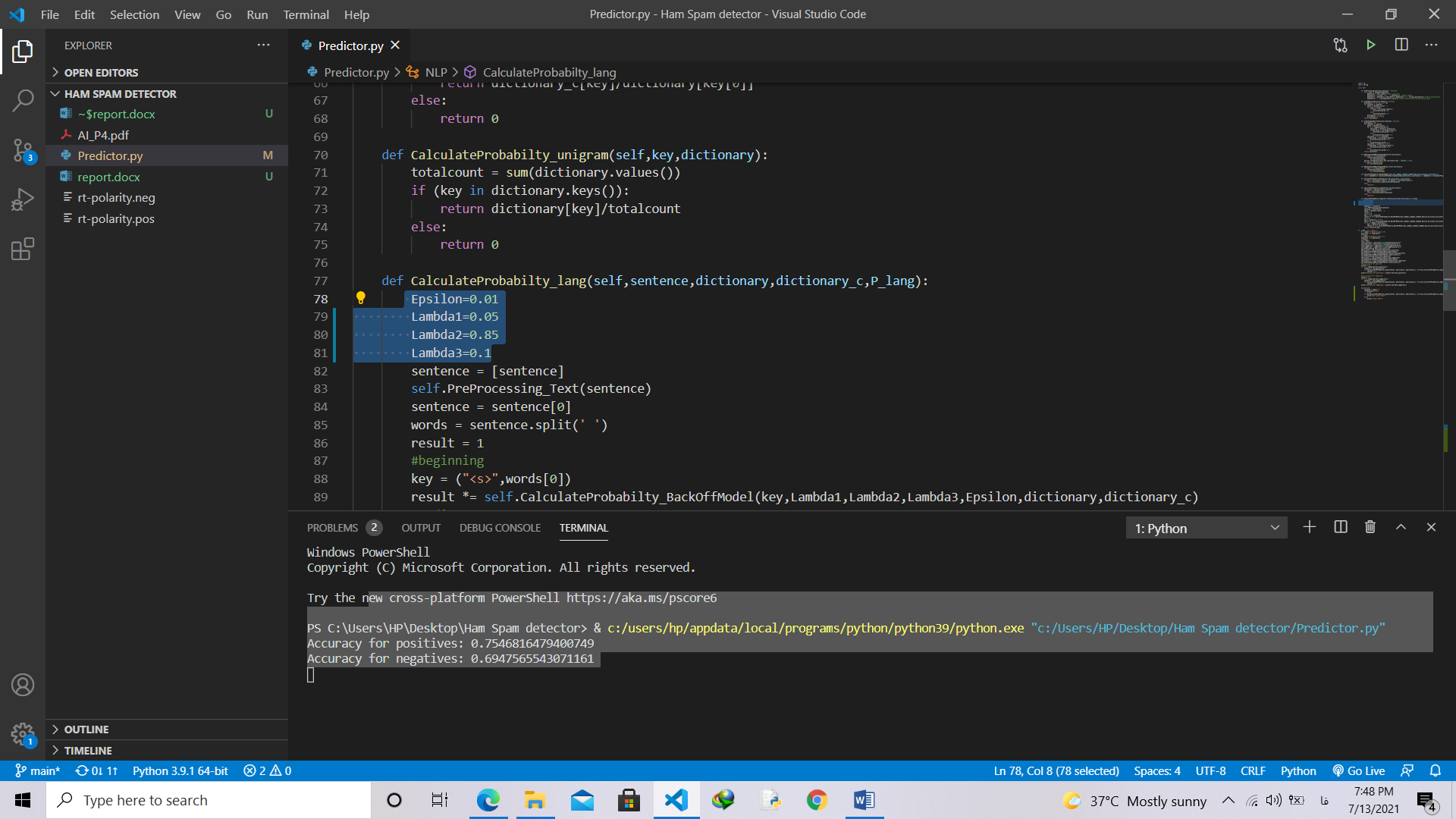


حال اگر تنها برای یک دیتاست حذف کنیم به شدت دقتمون دچار خطا می شود و دیتا ست ای که کلمات پرتکرار و کم تکرار در آن حذف شده به شدت دچار خطا می شود. ( 65 به 14 )

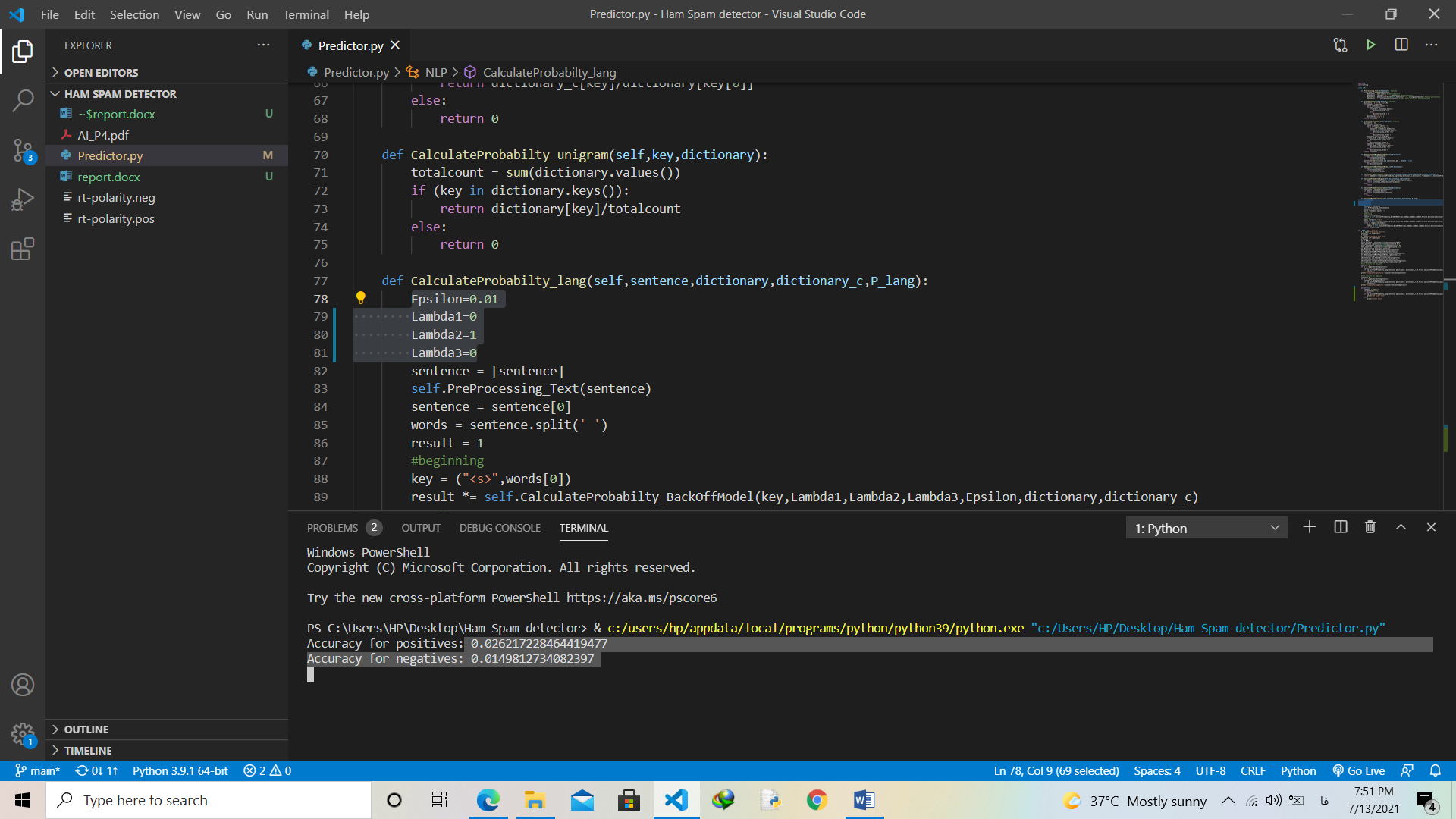
برای بررسی تاثیر مقادیر لاندا و اپسیلون چند سناریو را اجرا می کنیم :



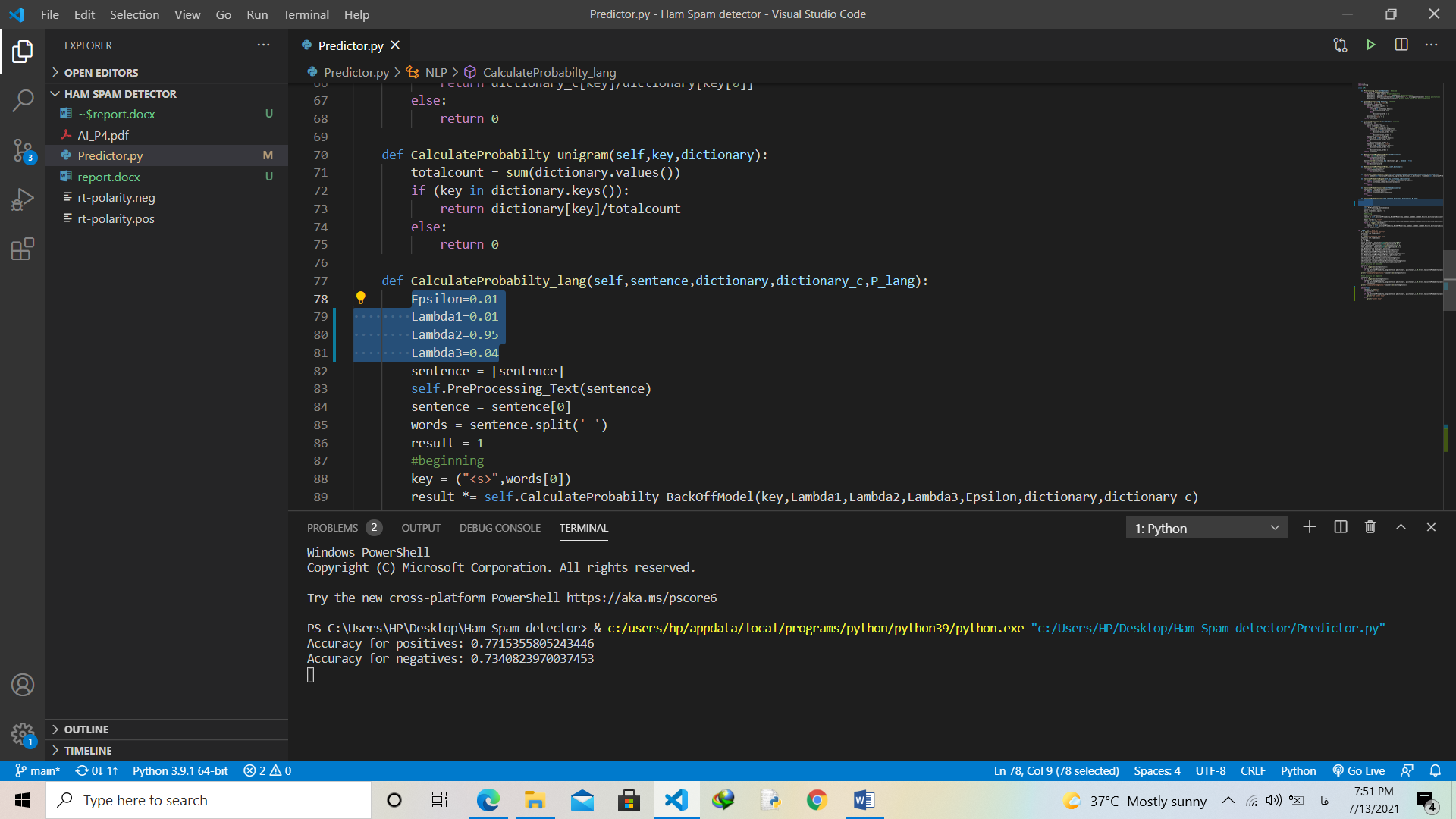
در سناریو اول تاثیر بایگرام بیشتر از یونیگرام و اپسیلون هست و به دقت 71 و 64 رسیده ایم.



با کم کردن تاثیر بایگرام و افزایش تاثیر یونیگرام دقت ما بیشتر شده است و این مسئله نشان می دهد که احتمال یونیگرام برای این دیتاست خاص بهتر عمل می کند و زبان بیشتر وابسته به خود کلمات می باشد و نه به ترتیب کلمات. دقت در این حالت 75 و 70 می باشد.



اما اگر به طور کل تنها از یونیگرام استفاده کنیم و تاثیر اپسیلون را حذف کنیم دقت مدل به شدت کاهش می یابد و به 2 و 1 درصد می رسد که این نشان می دهد که به کلمات جدید لازم است حساس باشیم و دقت کنیم و دیتاست ما همه کلمات را شامل نمی شود.



بهترین دقت زمانی اتفاق می افتد که تاثیر مدل یونیگرام از همه بیشتر باشد اما نه به طوری که به اپسیلون(برای کلمات جدید) و بایگرام دقت نکنیم چون همانطور که در بخش قبل دیدیم با حذف این دو مورد دقت مدل به شدت کاهش پیدا میکند.

و این مسئله نشان می دهد که برای زبان مورد نظر ما مدل یونیگرام بهتر است و زبان بیشتر به خود کلمات وابسته است تا به ترتیب کلمات .