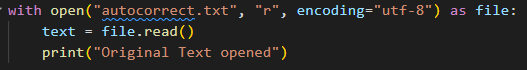
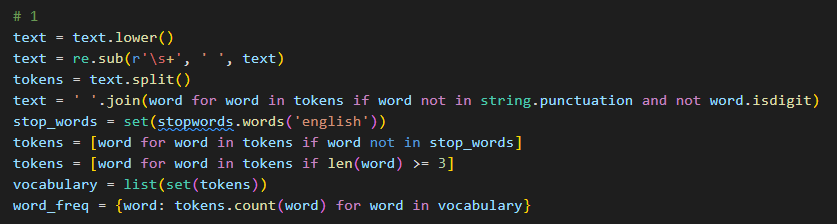
گزارش تمرین اول علی آرمان

نکته: در کد ارسال شده پرینت هر مرحله و نمایش پلات ها کامنت شده و برای نمایش خروجی نیاز است تا پرینت مربوطه از کامنت خارج شود.

ابتدا فایل تکست موجود را باز میکنیم:



مرحله 1 پیش پردازش



جهت انجام پیش پردازش ابتدا تمام متن لور کیس شده، با استفاده از regex بیش از یک فاصله حذف شده.

در مرحله بعدی با یک if توکن هایی که علائم نگارشی و یا اعداد نیستند به عنوان متن استخراج شده و در متغیر txt ذخیره شده اند.

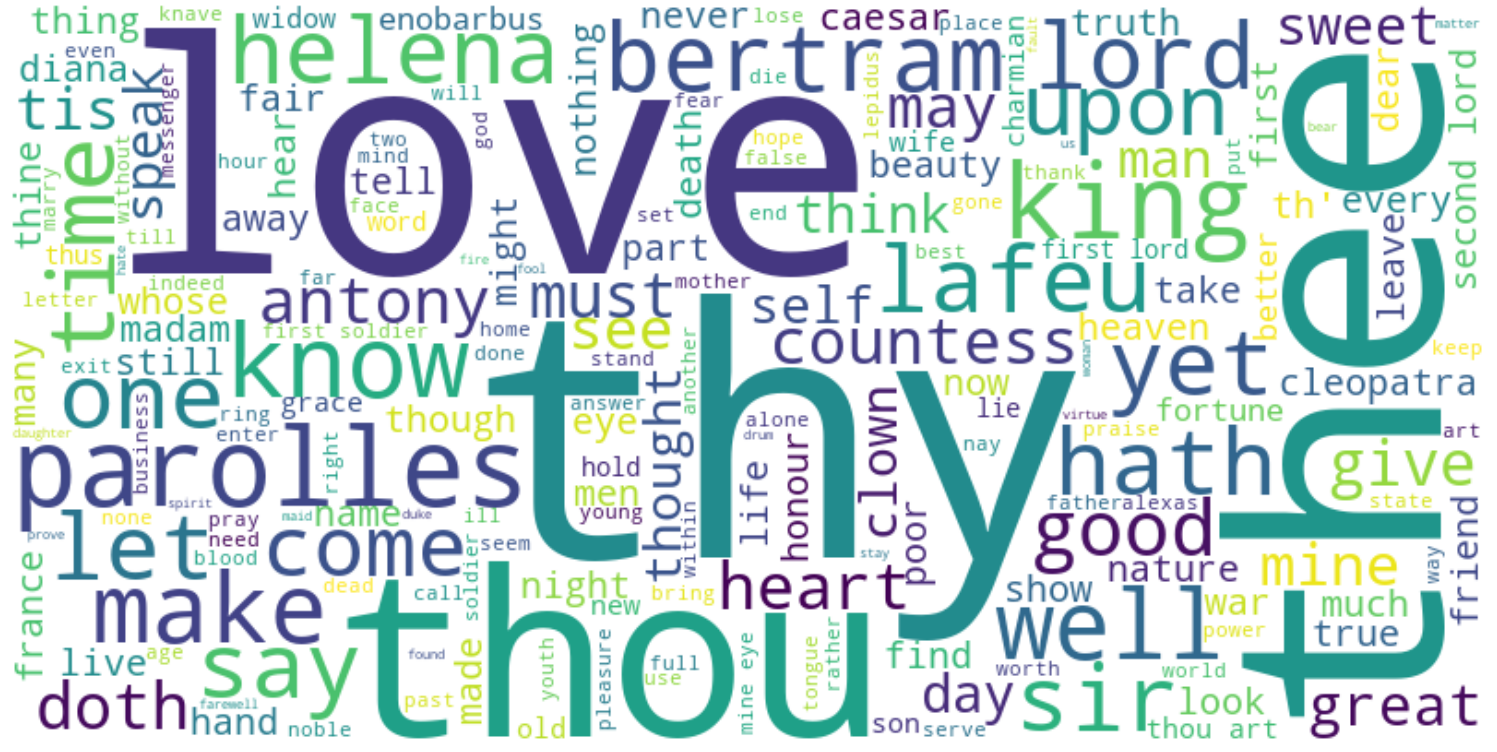
در دو مرحله لیست توکن ها با حذف استاپ وارد ها و کلمات کمتر از 3 کارکتر اپدیت شده.

دلیل حذف کلمات کمتر از 3 کارکتر می تواند به دلیل مخفف بودن، احتمال کم، اشتباه املایی باشد.

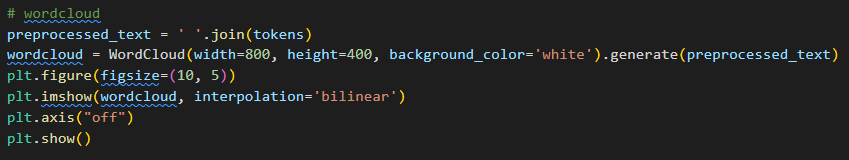
در مرحله بعد با گرفتن ست توکن ها، دایره لغات ایجاد میشود.

و در دیکشنری word\_freq کلید میشود کلمه و تکرار میشود مقدار.

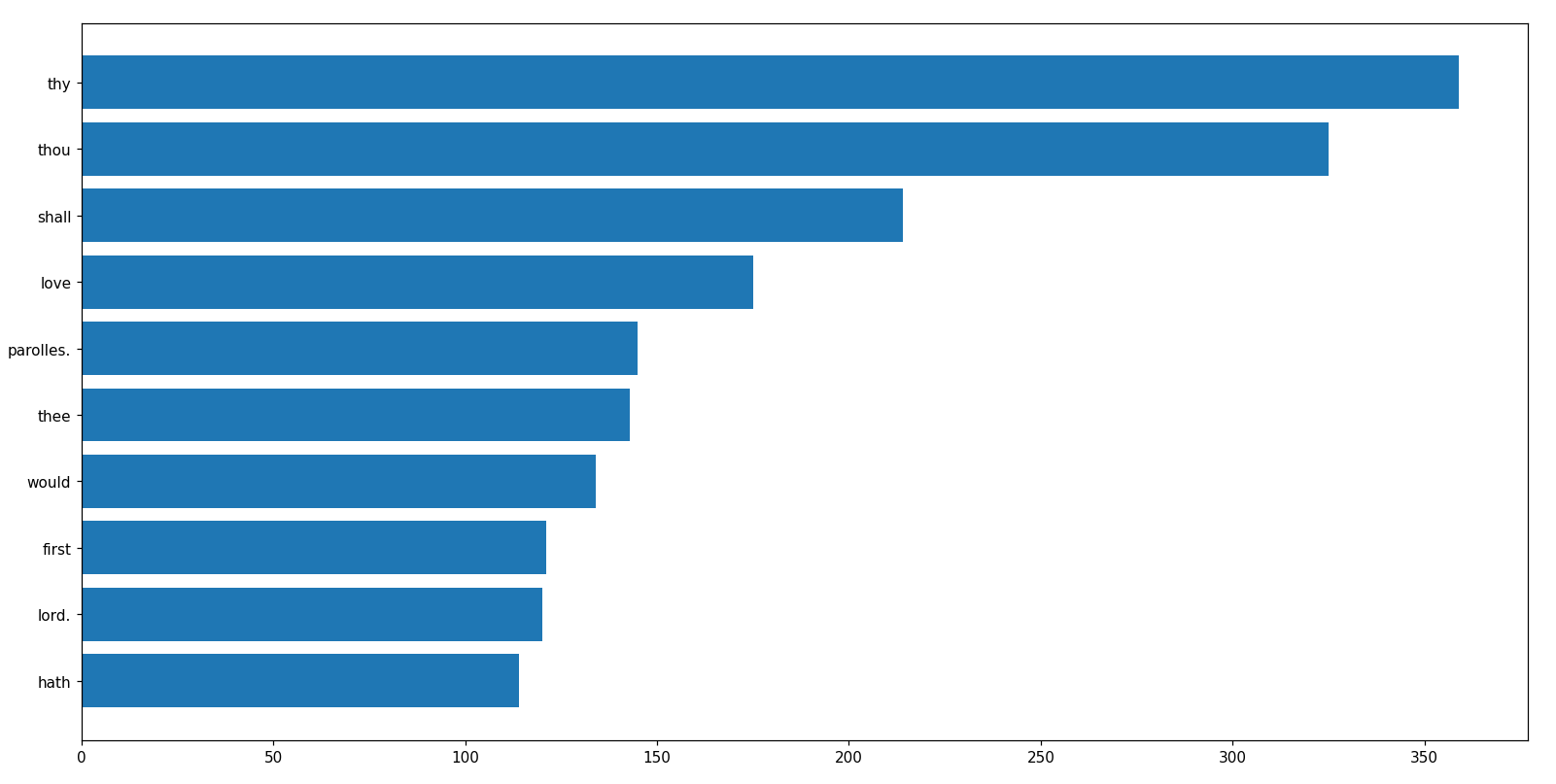
نمایش wordcloud



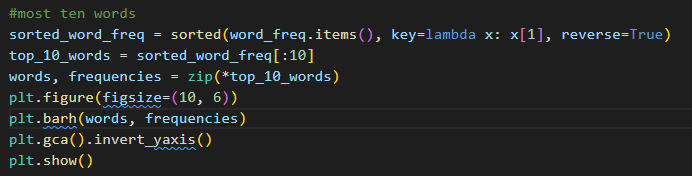
جهت نمایش ابر کلمات از کد زیر استفاده شده:



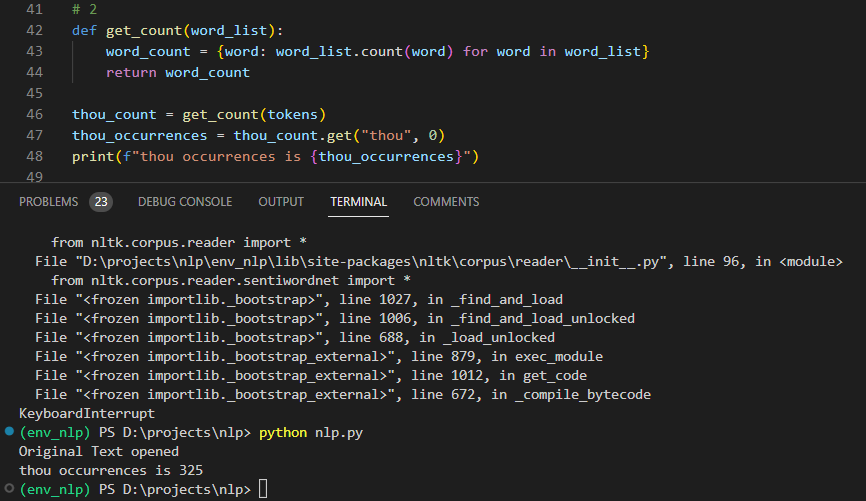
نمایش نمودار



جهت نمایش نمودار از کد زیر استفاده شد:

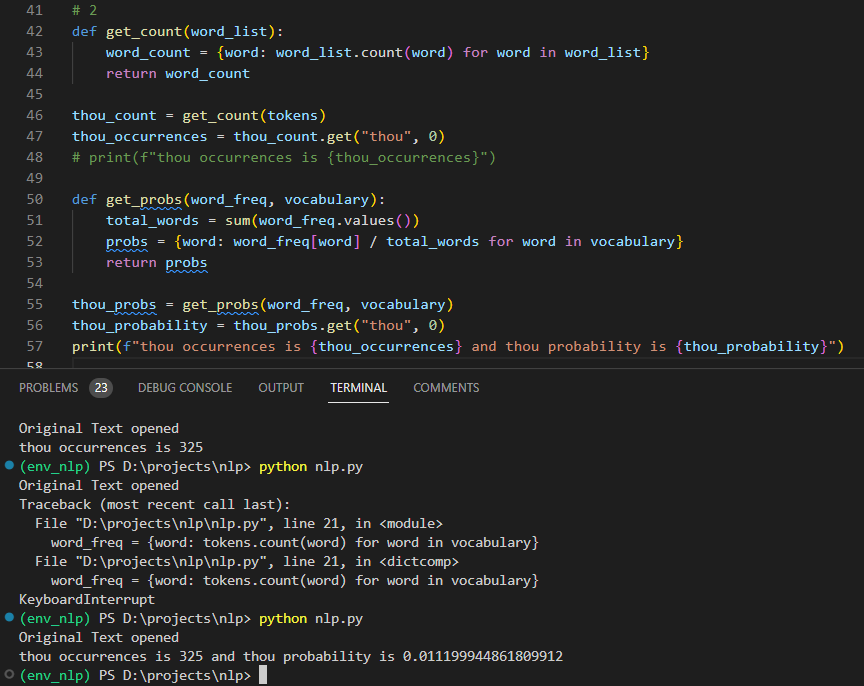


مرحله دو محاسبه احتمالات



همان طور که پیداست تعداد تکرار کلمه thou برابر با 325 می باشد.

همچنین احتمال وقوع



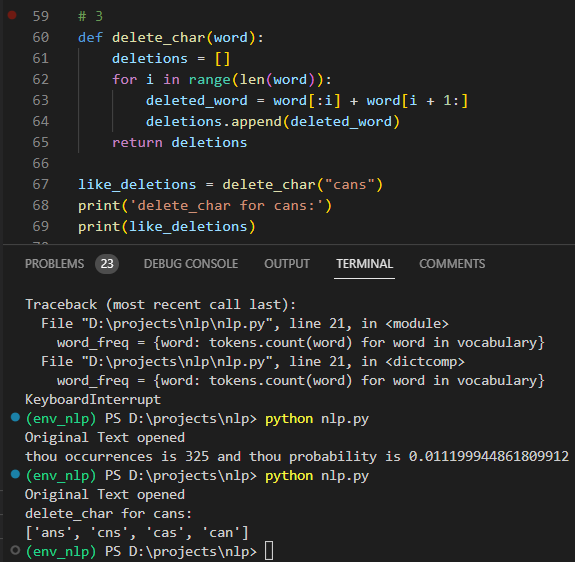
همان طور که در تصویر مشاهده میشود امکان وقوع برابر است با 0.011199944861809912

مرحله 3 دستکاری رشته ها

تابع delete\_char

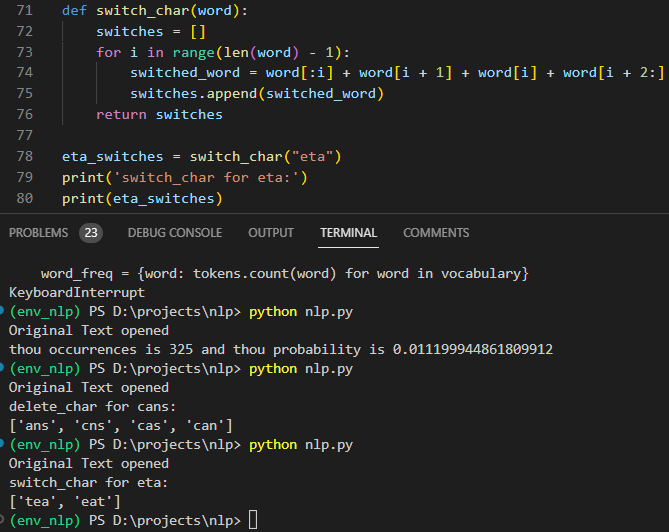
در پیاده سازی این تابع با استفاده از یک for تمام حالات نبودن یک کارکتر پیاده سازی شده.

و در انتها تمام حالات برای کلمه cans پرینت شده:



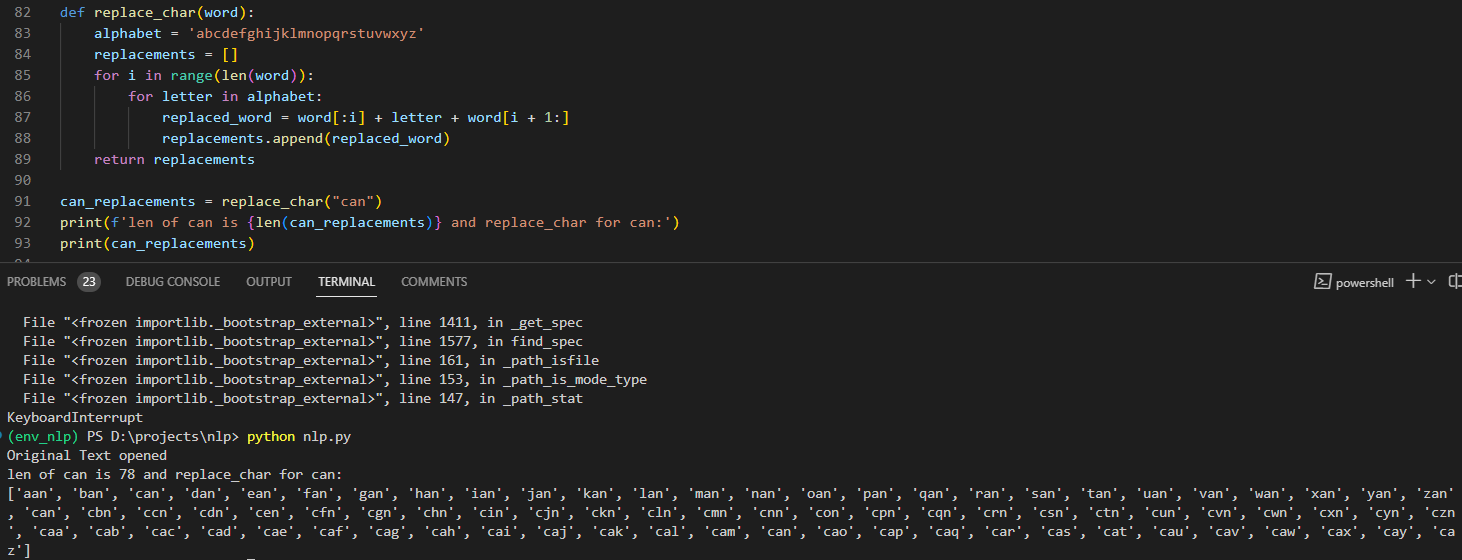
تابع switch\_char

پیاده سازی تابع switch\_char مانند تابع قبلی با استفاده از for می باشد:



تابع replace\_char

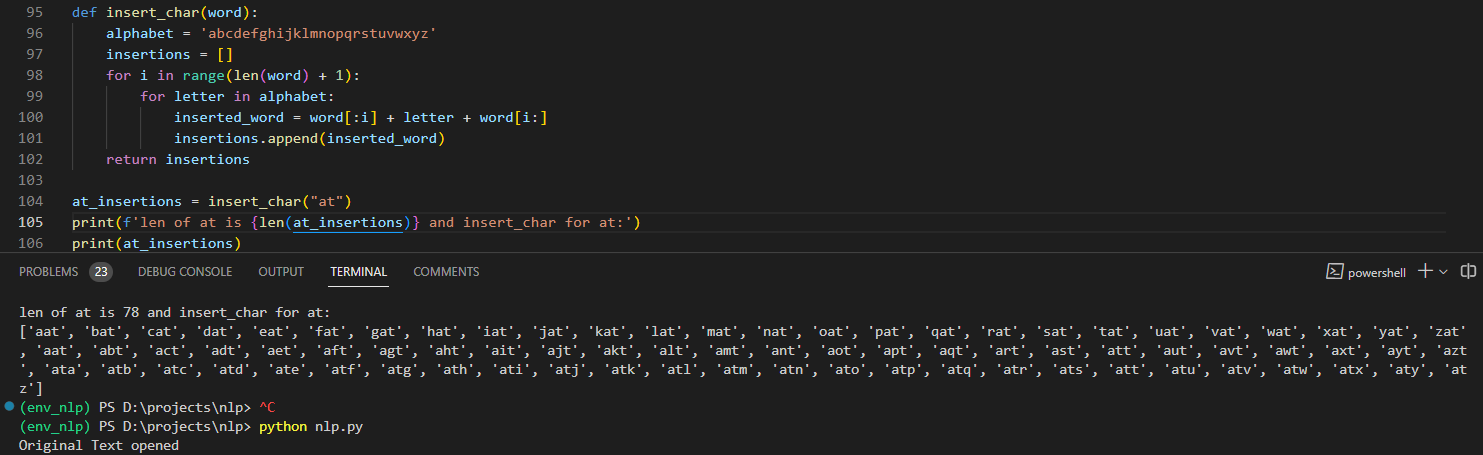
در پیاده سازی این تابع ابتدا حروف مشخص شده اند و با for تمام حالات رخداد به وجود آمده اند:



که همان طور که در تصویر میبیند تعداد برابر با 78 می باشد.

تابع insert\_char

این تابع هم به مشابه تابع قبل پیاده سازی شده:



که همان طور که در تصویر میبیند تعداد برابر با 78 می باشد.

تابع edit\_one\_char

در این تابع توابع پیاده سازی شده فراخوانی شده و در نهایت برای جلوگیری از تکرار ست گرفته شده:



همان طور که در تصویر میبیند تعداد برابر با 130 می باشد.

تابع edit\_two\_char

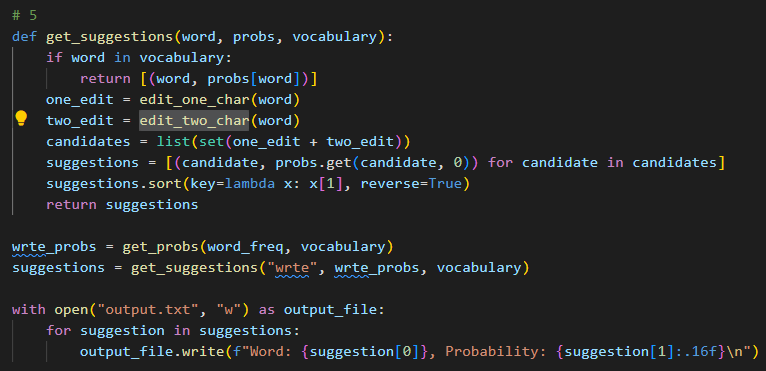
ابتدا یک بار edit\_one\_char فراخوانی شده و برای هر کلمه ایجاد شده مجدد با استفاده از حلقه فراخوانی edit\_one\_char انجام شده.

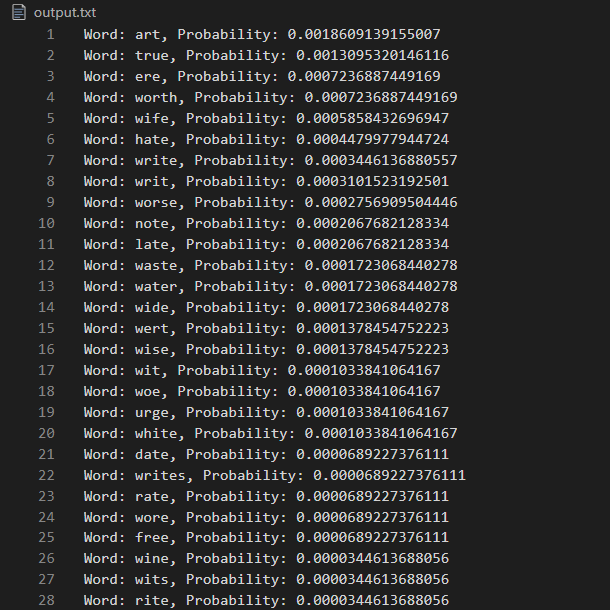


همان طور که در تصویر میبیند تعداد برابر با 20740 می باشد.

تابع get\_seggestions

برای پیاده سازی ابتدا کلمه، احتمال کلمه و لیست کلمات را دریافت کردیم، ابتدا اگر کلمه در لیست کلمات باشد همان کلمه با احتمالش برگردانده میشود. در غیر این صورت edit\_one\_char و edit\_two\_char فراخوانی شده و حاصل دو تابع ست شده و تاپلی از کلمات ایجاد شده که کلید کلمه و مقدار تکرار است و اگر تکراری نداشت مقدار 0 لحاظ می شود. در نهایت به ترتیب نزولی مرتب شده و در یک فایل ذخیره می شود.





همان طور که مشخص می باشد بیشترین احتمال برای کلمه art با احتمال 0.0018609139155007 می باشد.