*921531110*

**علی فاضلی**

**جناب آقای دکتر فیضی درخشی**

|  |
| --- |
|  |
| گزارش پروژه چهارم درس مبانی پردازش زبان گفتار |
|  |
|  |

**استاد محترم:**

گزارش پروژه چهارم درس مبانی پردازش زبان گفتار

علی فاضلی

**هدف پروژه :**

هدف این پروژه پیاده سازی ماتریس Bigram و تعیین جدول احتمالات آن می باشد.

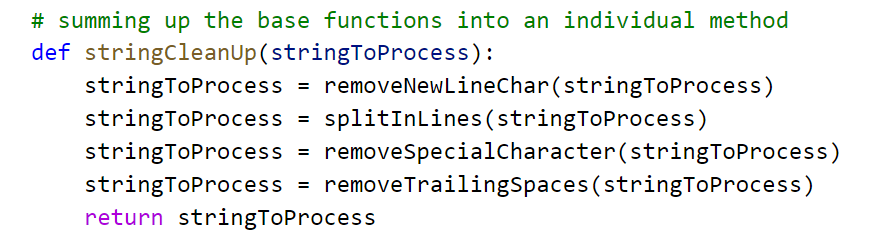
برای پیاده سازی از پایتون نسخه سوم استفاده شده است. پیاده سازی دارای دو بخش پیش پردازش داده و انجام محاسبات برای تشکیل ماتریس بایگرام می باشد که هردو به تفضیل شرح داده خواهند شد:

**مرحله اول، پیش پردازش داده:**

این فاز از پروژه توسط اسکریپت preprocessdata.py انجام می پذیرد. این اسکریپت وظیفه پیش پردازش داده ها را دارد.در این کلاس متد یا فانکشن های پایه ای مورد نیاز پیاده سازی شده اند که کارکرد هریک از آنان در جدول زیر آورده شده است:

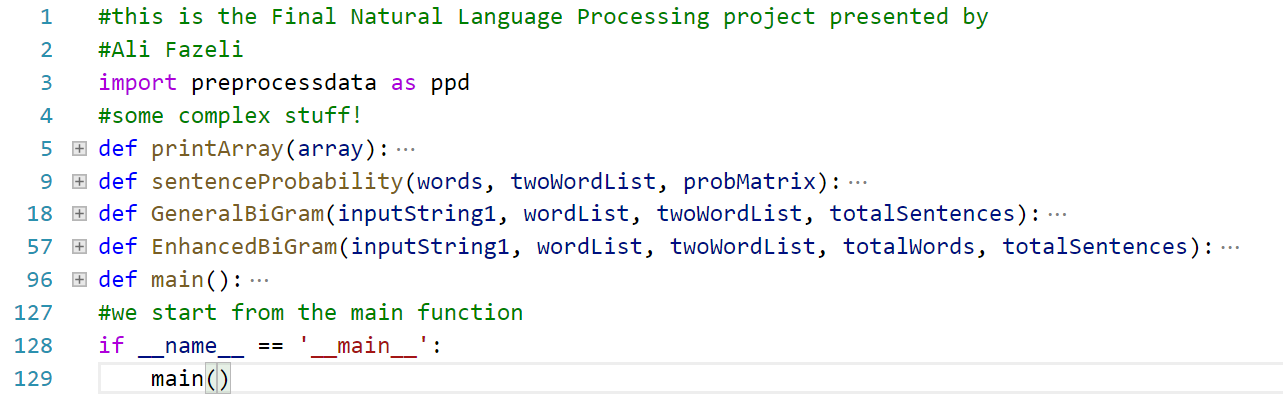
|  |  |
| --- | --- |
| حذف کردن کاراکتر"\n" به منظور جلوگیری از اشتباه متد splitInLines() | removeNewLineChar(stringToProcess) |
| جدا کردن متن با استفاده از کاراکتر توقف نقطه | splitInLines(stringToProcess) |
| حذف کردن کاراکتر های سمبل | removeSpecialCharacter(stringToProcess) |
| حذف فاصله های پشت سر هم در رشته ورودی | removeTrailingSpaces(stringToProcess) |
| خواندن ورودی الگوریتم (Corpus.txt) | getStringFromFile() |
| ایجاد لیست کلمات، محاسبه عداد آنها و طول رشته | getEachWordFrequency(stringToProcess) |
| محاسبه تعداد دفعات تکرار Bigram | getGivenPreviousWordNewWordFrequency(stringToProcess) |

در آخر این بخش متد stringCleanUp() را داریم که توابع بالا را در یک متد جمع آوری کرده است. یک رشته را به عنوان ورودی گرفته و در آخر رشته پردازش شده را به خروجی می برد (return می کند)



**مرحله دوم، پیاده سازی الگوریتم Bi-Gram :**

شمای کلی اسکریپت به شکل زیر است:



همانطور که قابل ملاحظه است به غیز از تابع main سه تابع sentenceProbability ، GeneralBiGram و EnhancedBiGram پیاده سازی شده است که کارکرد آنها به اختصار بررسی می شود:

**تابع sentenceProbability :**

این تابع وظیفه جایگذاری احتمال هر کلمه در ماتریس Bi-Gram را بر عهده دارد که به ازای هر توکن در جمله این کار را انجام می دهد.

**تابع GeneralBiGram :**

این تابع پیاده سازی محاسبه ی ماتریس احتمالات Bi-Gram را به روش اولیه تدریس شده توسط استاد می باشد. بطوریکه احتمالات تکرار نشدن را صفر در نظر می گیریم.

**تابع EnhancedBiGram :**

این تابع وظیفه محاسبه ماتریس احتمالات Bi-Gram را برعهده دارد. کلیات این روش مشابه روش معمول GeneralBiGram می باشد. اما با این تفاوت که به مقادیر جدول احتمالات مقدار ثابتی اضافه میشود.

به علت ذیق وقت قابلیت حدس کلمات بعدی و قابلیت ذخیره در فایل متنی پیاده سازی پیاده سازی نگردید. امید است تا زمان تحویل پروژه حضوری این قابلیت ها به پروژه اضافه گردد.