به نام خدا

گزارش پروژه پیدا کردن شاعر

استاد درس: مهندس روشن فکر

میلاد اسرافیلیان
۹۷۳۱۰۰۷

- **اجرای برنامه به ازای مقادیر مختلف لاندا و اپسیلون:**

مقادیر اول:

Landa3 = 0.9, Landa2 = 0.09 , Landa1 = 0.01
Epsilon = 0.01

خروجی:

```
Ferdowsi Training is completed
Hafez Training is completed
Molavi Training is completed
Ferdowsi Accuracy is:  91.58110882956879
Hafez Accuracy is:  72.82608695652173
Molavi Accuracy is:  86.0576923076923
```

مقادیر دوم:

Landa3 = 0.2, Landa2 = 0.1 , Landa1 = 0.7
Epsilon = 0.01

خروجی:

```
Ferdowsi Training is completed
Hafez Training is completed
Molavi Training is completed
Ferdowsi Accuracy is:  84.68125594671741
Hafez Accuracy is:  68.74074074074073
Molavi Accuracy is:  82.8125
```

مقادیر سوم:

Landa3 = 0.2, Landa2 = 0.1 , Landa1 = 0.7
Epsilon = 0.001

خروجی:

```
Ferdowsi Training is completed
Hafez Training is completed
Molavi Training is completed
Ferdowsi Accuracy is:  89.50554994954591
Hafez Accuracy is:  70.40673211781207
Molavi Accuracy is:  84.32122370936902
```

مقادیر چهارم:

Landa3 = 0.2, Landa2 = 0.1 , Landa1 = 0.7
Epsilon = 0.1

خروجی:

```
Ferdowsi Training is completed
Hafez Training is completed
Molavi Training is completed
Ferdowsi Accuracy is:  79.48260481712757
Hafez Accuracy is:  67.66666666666666
Molavi Accuracy is:  79.98056365403305
```

- **شرح دلایل:**

همانگونه که در درس آموختیم استفاده از بایگرام جهت مدل سازی یک ترکیب دو کلمه ای و یونیگرام یک ترکیب یک کلمه ای است حال اگر در حین تشخیص کلمه ای در واژه نامه ما وجود نداشته باشد احتمال اختصاص داشتن به آن زبان صفر خواهد شد به همین دلیل از مدل بک آف استفاده می کنیم.
در مدل بک آف ضریب بایگرام نشان دهنده اهمیت به بایگرام و ضریب یونیگرام اهمیت به آن و لاندا ۱ و اپسیلون نشان دهنده اهمیت به کلمات جدیدی است که در لغت نامه ما وجود ندارند. از آنجا که تشخیص جمله با توجه به ترکیبات دو کلمه ای دقیق تر خواهد بود (در صورت وجود واژه های یکسان در زبان هر دو مدل) پس باید ضریب بیشتری داشته باشد. همانطور که مشاهده می شود مدل اول دقیق تر از بقیه است چرا که ضریب بایگرام عدد بزرگتری است و مقدار اپسیلون نیز کوچک است.
در مثال دوم وقتی مقدار بزرگتری به لاندا ۱ داده می شود دقت کاهش می یابد(باید توجه شود که وجود اپسیلون کوچک همچنان مقدار نهایی را برای کلمات ناشناخته برای زبان کم می کند و از کاهش زیاد دقت جلوگیری می کند)
در مثال سوم مشاهده می شود که با کاهش بیشتر اپسیلون تاثیر کلمات ناشناخته در تشخیص کم می شود و دقت کمی بیشتر می شود.
اما در مثال چهارم که هم لاندا ۱ زیاد است و هم اپسیلون مقدار نسبتا بزرگی دارد می توان کاهش دقت را مشاهده کرد. چرا که با دیدن عباراتی که در مدل وجود ندارد نیز با توجه به مقدار زیاد landa1 * epsilon احتمال بزرگی به آن نسبت داده می شود و همین سبب کاهش دقت و نسبت دادن مصراع به شاعر اشتباه می شود.

- **شرح کد:**

```
def readFromFile(path: str):
    'Reads all lines of specified file'
```

جهت خواندن جملات از فایل مشخص شده و افزودن </s> , <s> به ابتدا و انتهای جملات.

```
def removeSigns(sentences: list):
    'Remove all . , ? / ! signs from sentences'
```

جهت پاک کردن علائم نگارشی از جملات و افزایش دقت.

```
def createDictionary(sentences: list):
    'Find and return dictionary of all words'
```

جهت ایجاد دیکشنری از لغات و تعداد تکرار آن ها.

```
def findUnigram(dictionary: dict, total: int):
    'Finds and returns the probability of each key of dictionary'
```

جهت محاسبه احتمال رخداد هر کلمه در مدل زبانی و ایجاد یونیگرام.

```
def findBigram(sentences: list , dictionary: dict):
    'Finds and returns the probability of any two consecutive words'
```

جهت محاسبه احتمال هر دو کلمه متوالی و ایجاد بایگرام.

```
def backOffModel(twoWords: str , bigram: dict, unigram: dict):
    'Finds the probability of given string with landa and epsilon value'
```

جهت محاسبه بک آف برای هر دو کلمه ورودی در یونیگرام و بایگرام ورودی با استفاده از مقادیر لاندا و اپسیلون تعیین شده.

```
def readTestFile():
    'Reads and filters test file'
```

جهت خواندن جملات فایل تست افزودن </s> , <s> به ابتدا و انتهای آن و حذف علائم نگارشی.

```
def findAccuracy():
```

تابع اصلی جهت یافتن شاعر برای هر مصراع و چاپ دقت یافتن درست شاعر هر مصراع.