به نام وی

گزارش پروژه پایان ترم مبانی هوش مصنوعی

نام : سید سروش مرتضوی مقدم

شماره دانشجویی : 9631068

دی ۹۸

شرح مسئله :

بنا به پیاده سازی یک روش برای دسته بندی کردن متون داریم (به یکی از ۵ دسته : اقتصاد، ادب و هنر، اجتماعی، سیاسی ، ورزش)

الگوریتم مورد استفاده در این پروژه مدل سازی unigram و bigram متن ها است. و استخراج احتمال تعلق یک متن به هر یک از دسته بندی های موجود. به این منظور ابتدا از مجموعه کثیری از داده های مجموعه آموزش استفاده کرده و دو مدل طراحی می کنیم :

۱ مدل تک کلمه ای: به این منظور تعداد استفاده هر کلمه در هر یک از دسته بندی های ذکر شده را شمارش کرده و ذخیره می کنیم (مثلا میدانیم کلمه <جامعه> در هر یک از دسته بندی ها چند بار تکرار شده است)

۲ مدل دو کلمه ای: به این منظور به ازای هر دو کلمه ای که در همه متون پشت سر هم آمده اند ، تعداد این رخ داد (یعنی پشت سر هم آمدن آن دو کلمه) را در همه دسته بندی ها شمارش کرده و ذخیره می کنیم. (یعنی مثلا تعداد رخ داد عبارت < جامعه ما> را در همه دستبه بندی ها به صورت مجزا ذخیره می کنیم)

همچنین برای محاسبه p(topic) برای هر دسته بندی ، نسبت تعداد آن دسته بندی به کل دسته بندی ها در نظر گرفته می شود.

پس از برسی کلمات مجموعه آموزش ، زمان برسی دقت مدل است. برای این منظور از مجموعه تست استفاده می کنیم.

در مدل اول) به ازای هر کلمه در یک متن تست ، تعداد رخ داد آن در هر دسته بندی را تقسیم بر کل کلمات آن دسته بندی می کنیم (= احتمال unigram برای آن کلمه) . اگر رخ داد کلمه ای ۰ باشد به آن یک احتمال خیلی کم نسبت می دهیم. سپس از همه اعداد به دست آمده برای همین این کلمات لگاریتم گرفته و با هم جمع می کنیم + لگاریتم احتمال همان دسته بندی.

در نهایت عدد بدست آمده برای هر دست که بیشتر باشد ، متن احتمالا به همان دسته تعلق دارد.

در مدل دوم) به ازای هر دو کلمه متوالی در متن تست ، رخ داد این عبارت را در همه دسته بندی ها شمارش کرده و بر تعداد رخ داد کلمه اول عبارت ، تقسیم می کنیم اما این بار اگر رخ داد عبارت ۰ باشد از روش backOff استفاده می کنیم . یعنی دو عدد را محاسبه می کنیم :

اولی احتمال unigram کلمه دوم در عبارت (مشابه حالت قبل) و دومی هم یک عدد کوچک به ازای رخ داد کل عبارت. هر یک از این اعداد در پارامتری ضرب شده (lambda) به نحوی که جمع lambda ها ۱ شود. و حاصل به عنوان احتمال کل در نظر گرفته می شود. با امتحان کردن حدود این اعداد را به ترتیب ۰/۸ و ۰/۲ قرار می دهیم (نسبت با حالات دیگر بهتر بود) سپس از همه اعداد به دست آمده برای همین این کلمات لگاریتم گرفته و با هم جمع می کنیم + لگاریتم احتمال همان دسته بندی.

در نهایت عدد بدست آمده برای هر دست که بیشتر باشد ، متن احتمالا به همان دسته تعلق دارد.

برای هر دو مدل بالا همه متون تست پیش بینی شده اند و در نهایت همه اطلاعات خواسته شده در صورت تعریف پروژه به دست آمده و در زیر گزارش شده است:

```
_____________________________
unigram true predictions: 694\860
precision(UNIgram):
   0.978 : اقتصاد
   0.8 : ادب و هنر
   0.531 : اجتماعی
   0.895 : سیاسی
   1.0 : ورزش
recall(UNIgram):
   0.658 : اقتصاد
   0.965 : ادب و هنر
   0.926 : اجتماعی
   0.685 : سیاسی
   0.925 : ورزش
F1_score(UNIgram):
   0.787 : اقتصاد
   0.875 : ادب و هنر
   0.675 : اجتماعی
   0.776 : سیاسی
   0.961 : ورزش
```

```
_____________________________
bigram true predictions: 799\860
precision(BIgram):
   0.956 : اقتصاد
   1.0 : ادب و هنر
   0.84 : اجتماعی
   0.886 : سیاسی
   0.995 : ورزش
recall(BIgram):
   0.947 : اقتصاد
   0.844 : ادب و هنر
   0.901 : اجتماعی
   0.895 : سیاسی
   0.982 : ورزش
F1_score(BIgram):
   0.952 : اقتصاد
   0.915 : ادب و هنر
   0.869 : اجتماعی
   0.890 : سیاسی
   0.988 : ورزش
```