مجید ادیبیان 9631003 گزارش پروژه پایانی

توضیح کلی پروژه:

در این پروژه می خواهیم با استفاده از مفاهیم n-gram در پردازش زبان طبیعی به دسته بندی متن بپردازیم.

به این منظور دو فایل در اختیار داریم که یکی از آن ها فایل آموزشی است و دیگری فایلی برای تست مدل سازی انجام شده. در هر یک از این فایل ها مجموعه ای از متن ها با موضوع آن داده شده است.

ابتدا باید با استفاده از فایل آموزشی مدل زبانی را استخراج کنیم که مدل های بررسی شده ungram و bigram می باشند. مدل سازی زبان را در هر دوی این روش ها انجام می دهیم و در هر روش احتمالات کلمه ها را در هر یک از موضوعات با استفاده از فایل آموزشی به تفکیک بدست می آوریم و ذخیره می کنیم و سپس فایل تست را اجرا می کنیم و در هر بند آن احتمال هر یک از موضوعات را با استفاده از احتمالات بدست امده در مرحله قبل بدست می آوریم و هر یک احتمال بیشتری داشت احتمالا آن بند در مورد آن موضوع است.

سپس پاسخ های بدست آمده را با مقدار درستی آن ها مقایسه میکنیم و درست درستی را در دو معیار $precision$ و $recall$ بدست می آوریم.

نحوه پیاده سازی:

پروژه مورد نظر به زبان پایتون نوشته شده است. در ابتدا تابع get_training_data() فراخوانی می شود و در این تابع فایل آموزشی فراخوانی میشود و حلقه ای روی تمام line های آن وجود دارد ور به ازای هر line یه بند از یک موضوع مجزا داریم. ابتدا موضوع آن جدا می شود و به عنوان کلید در یک dictionary قرار می گیرید و value برای این کلید خود dictionary دیگری است که به ازای هر کلمه و هر جفت کلمه ی به دنبال هم تعداد آن را خواهیم داشت. پس در پایان به ازای هر موضوع تعداد کل کلمه ها و جفت کلمه های به دنبال هم را داریم و جود پیاده سازی با استفاده از dictionary بوده است زمان دسترسی به این مقادیر در O(1) انجام می شود.

همچنین احتمال هر موضوع را هم با تقسیم تعداد تکرار آن موضوع در کل موضوعات فایل آموزشی بدست می آید که با محاسبه آن، این مقادیر را هم ذخیره می کنیم.

پس از بدست آوردن داده های مورد نیاز، از فایل آموزشی برای بررسی مدل زبانی unigram و بررسی داده های فایل تست برای این مدل، تابع calculate_class_in_unigram را فراخوانی می کنیم و برای بررسی داده ها در مدل bigram تابع calculate_class_in_bigram را فراخوانی می کنیم. در این توابع داده های بدست امده از مرحله قبل را به عنوان ورودی می دهیم و مشابه مرحله قبل داده ها را از فایل می خوانیم و موضوع را ذخیره می کنیم تا در انتها بتوانیم با مقدار پیش بینی شده مقایسه کنیم. سپس در مدل unigram و bigram مطابق تعریف آن احتمال هر زبان را با استفاده از داده های بدست امده از فایل آموزشی بدست می آوریم. به این شکل که به ازای هر کلمه در unigram و هر کلمه و جفت کلمه در bigram احتمال آن را در هر یک از موضوعات بدست می آوریم و به ازای آن موضوع ذخیره می کنیم و برای کلمه بعدی احتمال بدست آمده را در احتمال قبلی ضرب می کنیم و در انتهای آن بند احتمال های بدست آمده برای هر موضوع را بررسی می کنیم و موضوع با

بیشترین احتمال را به عنوان موضوع احتمالی آن بند در نظر می گیریم و با مقدار واقعی آن مقایسه می کنیم و اگر درست بود یک واحد به تعداد TP (که در ابتدا صفر بوده) برای آن موضوع اصلی اضافه می کنیم و اگر اشتباه بود یک واحد برای FP (که ابتدا صفر بوده) برای آن موضوع اصلی و آن موضوع اشتباه تشخیص داده شده اضافه می کنیم.

چون ضرب احتمالات بسیار کوچک میشود از جمع لگاریتم احتمالات استفاه شده است. همچنین با استفاده از قانون بیز احتمال برای هر موضوع بدست آمده است.

محاسبه احتمال ها به شکل زیر است:

$$unigam: P(c_{1:N}) = P(c1) \times P(c2) \times \dots \times P(cn) \rightarrow$$
$$\log\big(P(c_{1:N})\big) = \log\big(P(c1)\big) + \log\big(P(c2)\big) + \dots + \log(P(cn))$$

$$bigram: P(c_{1:N}) = P(c1| < s >) \times P(c2|c1) \times \dots \times P(c_n|c_{n-1}) \rightarrow$$
$$\log\big(P(c_{1:N})\big) = \log\big(P(c1| < s >)\big) + \log\big(P(c2|c1)\big) + \dots + \log(P(c_n|c_{n-1}))$$

همچنین در محاسبه احتمالات unigram اگر به مقدار صفر رسیدیم برای جلوگیری از صفر شدن کل حاصل احتمال(یا اگر از لگاریتم استفاده می کنیم برای امکان لگاریتم گرفتن از آن) برای احتمال صفر مقدار کوچک 0.000001 در نظر گرفته شده است و در روش bigram برای جلو گیری از این موضوع از روش Backoff استفاده شده که به این شکل است که برای هر یک از احتمالات bigram احتمال unigram را هم محاسبه می کنیم و ضریبی برای این دو در نظر می گیریم و حاصل را با یک دیگر جمع می کنیم. با استفاده از آزمون و خطا های مختلف می توانیم بهترین ضریب را برای بیشترین تشخیص درست بیابیم که با توجه به سه بررسی انجام شده در این مورد بهترین ضریب ها به شکل زیر هستند:

$$\lambda_1 P(ci|ci-1) + \lambda_2 P(ci) \rightarrow$$

$$\lambda_1 = 0.90, \lambda_2 = 0.20 : \begin{cases} unigram \rightarrow true = 785, false = 75 \\ bigram \rightarrow true = 802, false = 58 \end{cases}$$

$$\lambda_1 = 0.80, \lambda_2 = 0.20 : \begin{cases} unigram \rightarrow true = 785, false = 75 \\ bigram \rightarrow true = 803, false = 57 \end{cases}$$

$$\lambda_1 = 0.70, \lambda_2 = 0.30 : \begin{cases} unigram \rightarrow true = 785, false = 75 \\ bigram \rightarrow true = 803, false = 57 \end{cases}$$

$$\lambda_1 = 0.60, \lambda_2 = 0.40 : \begin{cases} unigram \rightarrow true = 785, false = 75 \\ bigram \rightarrow true = 804, false = 56 \end{cases}$$

$$\lambda_1 = 0.50, \lambda_2 = 0.50 : \begin{cases} unigram \rightarrow true = 785, false = 75 \\ bigram \rightarrow true = 803, false = 57 \end{cases}$$

پس بهترین مقدار $\lambda_1 = 0.60, \lambda_2 = 0.40$ می باشد.

بررسی خروجی برنامه:

پس از اجرای برنامه به ازای فایل های داده شده در مدل ***unigram*** تعداد تشخیص های درست برابر 785 تا و تعداد تشخیص های غلط برابر 75 تا است و مقدار ***precision*** و ***recall*** برای این مدل به شکل زیر است:

اقتصاد: precision-> 0.966824644549763

سیاسی: precision-> 0.905

ادب و هنر: precision-> 0.896551724137931

اجتماعی: precision-> 0.7484662576687117

ورزش: precision-> 0.9912280701754386

اقتصاد: recall-> 0.9026548672566371

سیاسی: recall-> 0.8341013824884793

ادب و هنر: recall-> 0.9454545454545454

اجتماعی: recall-> 0.9104477611940298

ورزش: recall-> 0.9912280701754386

F_measure:

اقتصاد : 0.9336384439359268

سیاسی : 0.86810551558753

ادب و هنر : 0.920353982300885

اجتماعی : 0.8215488215488216

ورزش : 0.9912280701754386

پس از اجرای برنامه به ازای فایل های داده شده در مدل ***bigram*** تعداد تشخیص های درست برابر ***804*** تا و تعداد تشخیص های غلط برابر ***59*** تا است و مقدار ***precision*** و ***recall*** برای این مدل به شکل زیر است:

اقتصاد: precision-> 0.9715639810426541

سیاسی: precision-> 0.915

ادب و هنر: precision-> 0.8620689655172413

اجتماعی: precision-> 0.8650306748466258

ورزش: precision-> 0.9868421052631579

اقتصاد: recall-> 0.9318181818181818

سیاسی: recall-> 0.8840579710144928

ادب و هنر: recall-> 1.0

اجتماعی: recall-> 0.9038461538461539

ورزش: recall-> 0.9911894273127754

F_measure:

اقتصاد : 0.951276102088167

سیاسی : 0.896551724137931

ادب و هنر : 0.9259259259259259

اجتماعی : 0.88125

ورزش : 0.989010989010989