به نام خدا

دانشکده مهندسی کامپیوتر

مبانی هوش مصنوعی ترم بهار ۰۰-۹۹

پروژه چهارم : پردازش زبان طبیعی (NLP)

مهلت تحویل ۲۸ تیر ۱۴۰۰

صورت مسئله

بسیاری از مردم فکر می‌کنند که هوش‌مصنوعی روزی جای انسان‌ها را خواهد گرفت و هر ساله شغل‌های -بدون خلاقیت- زیادی را از بین می‌برد. در این میان گروه‌های افراطی دست به اعتراضات زده‌اند و شرکت شما که اخیرا برای خود نامی دست و پا کرده، یکی از هدف‌های این گروه‌ها است. به دلیل شیوع کووید، اعتراضات نمی‌تواند بصورت حضوری برگزار شود و این گروه‌ها تصمیم می‌گیرند به سایت شرکت شما حمله کنند و کامنت‌های مخرب بگذارند.

شما که به سختی شرکت خود را به جایی رسانده‌اید می‌خواهید کاری کنید که این کامنت‌ها، قبل از اینکه در سایت ثبت شوند، فیلتر شوند تا اعتبار شرکت شما از بین نرود.

برای این کار ساده‌ترین کار استخدام چندین کارمند است که این کامنت‌ها را شناسایی و حذف کنند. با این کار باید هزینه‌ی زیادی بپردازید ولی احتمالا اگر تعداد بالایی از کارمندها استخدام کنید، این گروه‌های افراطی به هدف خود میرسند؛ پس شما تصمیم می‌گیرید بجای اینکه کارهای آسان را به آدم‌ها بسپارید، یک سیستم تشخیص احساسات توسعه دهید که بصورت خودکار این کار را انجام دهد.

دیتاست آموزش:

دیتاست ضمیمه شده شامل دو فایل rt-polarity.neg و rt-polarity.pos که به ترتیب شامل نظرات منفی و مثبت یک سایت می‌باشد.

در هر خط هر کدام از این فایل‌ها یک نظر قرار دارد که باید با کمک آن‌ها مدل زبانی خود را آموزش دهید.

جزئیات پیاده‌سازی

برای تشخیص اینکه یک نظر مثبت است یا منفی، باید در ابتدا مدل‌های زبانی مرتبط ساخته شود. برای این پروژه از مدل زبانی **bigram** استفاده می‌کنیم. برای ساخت مدل زبانی:

۰. پیش‌پردازش دیتاست

قبل از هر کاری، باید مطمئن شوید حروف بی‌کاربرد از جملات حذف شده باشند؛ برای مثال اینکه کاراکتر * در نظری باشد، نباید تاثیری داشته باشد.

۱. ساخت دیکشنری لغات

برای این کار باید فایل دیتاست را خوانده و با توجه به برچسب هر جمله، آن را در یکی از دو دیکشنری قرار دهید.

۲. شمارش کلمات هر دیکشنری

بعد از جمع‌آوری کلمات، باید تعداد تکرار آنها را بشمارید و آن را ذخیره کنید (مثلا استفاده از ساختار داده "دیکشنری"). از آنجایی که کلمات کم‌تکرار ارزش زیادی ندارند؛ برای سرعت بخشیدن مرحله‌ی بعدی، می‌توانید کلمات با تعداد کمتر از یک تعداد معین را از دیکشنری حذف کنید (مثلا کمتر از ۲). علاوه بر این کلمات با تکرار خیلی زیاد، مثلا the نباید ارزش زیادی داشته باشد، پس می‌توانید تعدادی از کلمات پرتکرار را نیز حذف کنید (مثلا ۱۰ کلمه با بالاترین فرکانس)

برای مدل زبانی بایگرم علاوه بر شمارش کلمات، نیاز به شمارش جفت کلمات نیز می‌باشد.

۳. محاسبه احتمالات

در نهایت با محاسبه‌ی احتمال هر کلمه و جفت کلمه، مدل بایگرم کامل می‌شود.

احتمال هر کلمه در یک زبان، برابر تعداد تکرار آن کلمه در آن زبان، تقسیم بر مجموع تکرار همه‌ی کلمات در آن زبان است.

$$P(w_i) = \frac{count(w_i)}{M}$$

احتمال وقوع هر جفت کلمه برابر تعداد تکرار تعداد آن جفت کلمه در آن زبان، تقسیم بر تعداد تکرار کلمه‌ی اول در آن زبان است.

$$P(w_i|w_{i-1}) = \frac{count(w_{i-1}w_i)}{count(w_{i-1})}$$

پس از آموزش دادن مدل مربوط به زبان منفی و یا مثبت، باید بتوانیم احتمال تعلق هر جمله‌ی دلخواه به هرکدامیک از این زبان‌ها را حساب کنیم. در نهایت زبانی انتخاب می‌شود که جمله‌ی داده شده با احتمال بیشتری به آن تعلق دارد.

$$P(l_i|w_1w_2 \dots w_{n-1}w_n) \propto P(w_1w_2 \dots w_{n-1}w_n|l_i) \times P(l_i)$$

فرض کنید احتمال هر کدام از زبان‌ها برابر ۰.۵ باشد.

برای محاسبه‌ی احتمال یک جمله $(w_1w_2 \dots w_{n-1}w_n)$ در یک زبان:

$$P(w_1w_2 \dots w_{n-1}w_n) = P(w_1) * \prod P(w_i|w_{i-1})$$

برای محاسبه‌ی $P(w_i|w_{i-1})$ از فرمول زیر استفاده کنید:

$$P(w_i|w_{i-1}) = \lambda_3 P(w_i|w_{i-1}) + \lambda_2 P(w_i) + \lambda_1 \epsilon$$

$$\lambda_3 + \lambda_2 + \lambda_1 = 1$$

$$0 < \epsilon < 1$$

امتیازی: پیاده‌سازی با مدل unigram

نکته: برای ارزیابی مدل می‌توانید از بخشی از دیتاست آموزشی را آموزش ندهید (مثلا ۱۰ درصد از هر کدام از فایل-ها را آموزش ندهید و آنها را در شافل کنید) و به عنوان داده تست استفاده کنید.

فرمت ورودی و خروجی

پس از آموزش یا ذخیره مدل، برنامه باید بتواند تا وقتی که دستور خروج **(!q)** داده نشده است، یک رشته از ورودی بخواند و عبارت filter this (برای جملاتی با زبان منفی) یا not filter this (برای جملات خنثی و مثبت) را در خروجی چاپ کند؛ برای مثال

```
> python comment-filter.py
> why did you make me do this?
filter this
> you’re fighting so you can watch everyone around you die
filter this
> think mark
not filter this
> you’ll outlast every fragile insignificant being on this planet
filter this
> !q
```

گزارش

گزارشی شامل موارد زیر تهیه کرده و در فایل تحویلی اضافه کنید:

- تاثیر حذف کلمات پرتکرار و کم تکرار در دقت بدست آمده
- تاثیر مقدار λ و ϵ دقت بدست آمده
- بهترین دقت دست‌یافته و تحلیل تاثیر پارامترها در آن

توضیحات تکمیلی

- این پروژه را بصورت انفرادی یا در گروه دو نفره انجام دهید.
- در صورت گروهی انجام دادن پروژه باید از گیت استفاده کنید.
- در صورت انجام پروژه به صورت گروهی، هر دو عضو گروه باید بصورت جداگانه فایل خود را در سامانه آپلود کنند.
- در صورت مشاهده تقلب، نمره دریافت شده، بین افراد خاطی تقسیم می‌شود.
- پروژه تحویل مجازی دارد و بخشی از نمره به تسلط اعضای گروه به کد اختصاص دارد.
- زبان انجام پروژه آزاد است.
- گزارش کد شامل موارد گفته شده را در یک فایل pdf در فایل زیپ اضافه کنید.
- فایل‌های کد و گزارش را در قالب نام‌گذاری AI_P4_9931099.zip در سامانه کورسز آپلود کنید. (نیازی به آپلود فایل‌های گیت نیست)
- در صورت هرگونه سوال یا مشکل با ایمیل ce.ai.spring00@gmail.com یا آی‌دی تلگرام @uramirbin در تماس باشید.
- ددلاین این پروژه **۲۸ تیر ۱۴۰۰ ساعت ۲۳:۵۵** است. با توجه به اینکه تحویل پروژه ها از ۲۹ تیر شروع می‌شود، **فقط تا ۸ ساعت تاخیر پذیرفته می‌شود.**