دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

تمرین سري سوم درس داده کاوي (امتیازی)

شهر Joo_Boo

استاد درس: دکتر شاکري

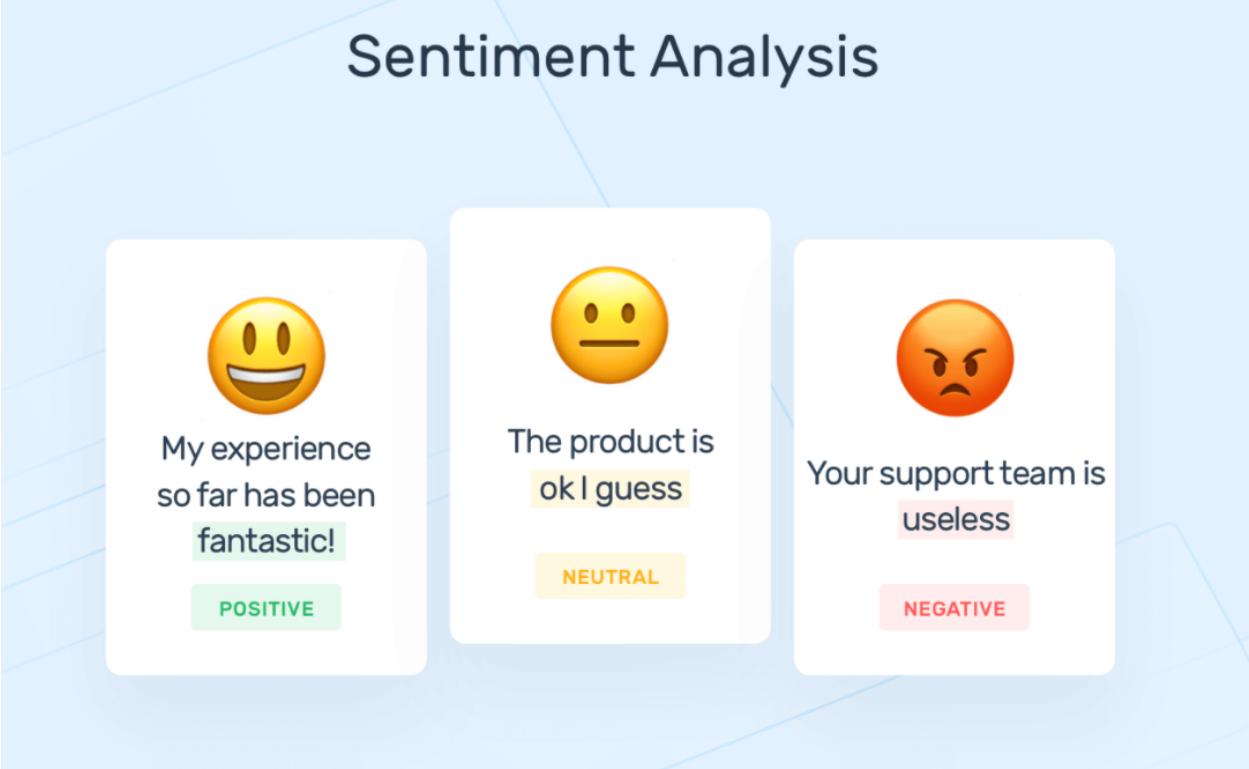
طراح: امید سقط چیان

تاریخ انتشار: چهارشنبه ۳ خرداد ماه ۱۴۰۲
مهلت تحویل : جمعه ۱۲ خرداد ماه ۱۴۰۲

شهر Joo-Boo

توضیحات اولیه:

میخواهیم با یکی از کاربردهای یادگیری ماشین در پردازش زبان های طبیعی آشنا شویم. (NLP)



یکی از تسک هایی که در این زمینه مطرح میشود به نام Sentiment Analysis هست. بدین معنی که ماشین وظیفه این را دارد که تعیین کند که یک داده جنبه مثبت، منفی یا خنثی دارد. این کار در حالت کلی جزییات قابل توجهی دارد که ما با آن کاری نداریم. (ما با دو حالت Happy و Sad کار خواهیم کرد، همچنین ما روی داده های متنی تمرکز خواهیم کرد.)

برای این کار یکی از رویکرد های بسیار ساده ای که وجود دارد این است که در نظر بگیریم که چه کلماتی به چه تعدادی در یک جمله استفاده شده اند و به هر کدام از کلمات نمره ای را نسبت دهیم (در اصل نشان دهیم هر کدام از کلمات چقدر موثرند روی حس جمله) این نمرات در بازه منفی یک تا یک خواهند بود. (مثبت یک یعنی بیشترین تاثیر در Happy شدن و بالعکس) در نهایت حاصلضرب وزن تمامی کلمات را در تعداد آن حساب کنیم و در نهایت با b جمع میکنیم، یعنی:

$$\sum_{1}^{n} w_n x_n + b$$

پارامتر b مقدار bias ما هست که در جلوتر توضیح داده شده است. حال داریم:

- اگر حاصل منفی شد جمله ناراحت است
- اگر حاصل مثبت شد جمله خوشحال است
- اگر حاصل صفر شد جمله خنثی است (ما از این حالت در تمرین صرف نظر میکنیم)

مثال:

جمله زیر را در نظر بگیرید:

I am sad

میتوان به "I" و "am" به ترتیب نمرات 0.01 و 0.04 را نسبت داد زیرا به نظر میرسد این دو کلمه تاثیر چندانی روی خوشحال بودن یا ناراحت بودن جمله ندارند. (توجه کنید که در اصل این وزن ها توسط ماشین یاد گرفته میشوند)

اما به کلمه sad نمره 0.9- را میتوان نسبت داد. حال تمامی این نمرات را ضرب در تعداد هرکدام از کلمات میکنیم و جمع میزنیم: (در اینجا bias را صفر در نظر گرفته ایم)

$$score_I * rep_I + score_{am} * rep_{am} + score_{sad} * rep_{sad} + b$$
$$0.01 * 1 + 0.04 * 1 + (-0.9 * 1) + 0$$
$$= -0.85$$

و چون این عدد منفی شد میگوییم این جمله ناراحت است.

دیتاست A:

حال فرض کنیم که شما به منطقه ای از شهری به نام Joo-Boo وارد شده اید که آدم های آن جا تنها از دو کلمه "Joo" و "Boo" استفاده میکنند (زبان آن ها تنها از این دو کلمه تشکیل شده است). شما صحبت های آدم های مختلف را میشنوید و متوجه میشوید که آن ها ناراحت هستند یا خوشحال. دیتاست زیر داده های جمع آوری شده توسط شما را نشان میدهد. (دو ویژگی x1 و x2 برای راحتی کار شما از ستون Sentence استخراج شده اند)

Sentence	**Sad/Happy (Y)**	**Number of Joo (x1)**	**Number of Boo (x2)**
Joo	Sad	1	0
Joo Joo	Sad	2	0
Boo Joo	Sad	1	1
Boo Joo Joo	Sad	2	1
Boo Boo Joo	Happy	1	2
Boo Boo	Happy	0	2
Boo	Happy	0	1

حال فرض کنید که ما میخواهیم مدلمان را اینگونه در نظر بگیریم: (هدف پیشبینی Y است)

$If\ (w1 * x1 + w2 * x2 + b > 0),\ then\ the\ sentence\ is\ Happy$

$If\ (w1 * x1 + w2 * x2 + b < 0),\ then\ the\ sentence\ is\ Sad$

همانطور که گفته شد به "b" همان bias میگوییم. شهودی که میتوان برای آن در نظر گرفت این است که اگر شما وارد این شهر شده اید و با یک شخص مواجه شده اید که به شما هیچ چیزی نمیگوید (x1 = x2 = 0) حال شما حال آن شخص را Happy یا Sad ارزیابی میکنید؟ (به نظر میرسد کسی که با شما صحبتی نکند ناراحت است، پس b باید سعی کند عبارت بالا را منفی کند که با منفی بودن b این اتفاق می افتد.) مثال دیگری میتوان در نظر گرفت: فرض کنید همین مدل را روی نظر های مردم روی یک فروشگاه آنلاین پیاده سازی میکنیم. فرض کنیم کسانی که از محصولات ناراضی هستند توضیحات کاملی میدهند که چه مشکلی وجود داشته و کسانی که راضی بوده اند صرفا نمره را میدهند و نظری هم نمیدهند. اینجا به وضوح b باید مثبت باشد.

سوالات دیتاست A

- سه پارامتر را چگونه تعیین کنیم تا روی دیتاست بالا دقت ۱۰۰ درصد داشته باشیم. (w1, w2, b)
- چه شهودی روی این دیتاست بدست می آورید؟ (راهنمایی: کدام کلمه باعث منفی شدن حس جمله و کدام کلمه باعث میشود جمله بار مثبت پیدا کند؟)
- به جای یک گزاره شرطی (if ...) میتوانیم از activation function استفاده کنیم و در اصل از یک perceptron استفاده کنیم. مشخص کنید از کدام activation function استفاده کنیم؟ (شکل perceptron را رسم کنید)
- اگر میخواستیم بجای اینکه به صورت گسسته بگوییم جمله ناراحت است یا خوشحال، از چه activation function ای باید استفاده میکردیم تا ورودی را به احتمال شاد بودن جمله تبدیل میکردیم؟
- چه شهود هندسی میتوان برای این دیتاست متصور شد ؟
  - نقاط را در دو بعد رسم کنید (در اصل هر جمله را میتوان با ۲ متغیر نشان داد) (انگار در این مدل ساده تنها تعداد ۲ کلمه برای ما اهمیت دارند، نه ترتیب و ...)
  - آیا به صورت خطی جدایی پذیر هستند؟
  - فضای دو بعدی را به دو قسمت تقسیم کنید و بگویید در کدام ناحیه جملات Happy و در کدام ناحیه جملات Sad هستند.

به ادامه سوال در صفحه بعدی توجه کنید

دیتاست B:

حال در منطقه دیگری از همان شهر ما با این دیتاست روبرو شده ایم:

Sentence	Sad/Happy (Y)	Number of Joo (x1)	Number of Boo (x2)
Joo	Sad	1	0
Joo Joo	Sad	2	0
Boo Joo	Sad	1	1
Boo Joo Joo	Happy	2	1
Boo Boo Joo	Happy	1	2
Boo Boo Boo Joo	Happy	1	3
Boo Joo Joo Boo	Happy	2	2
Boo Joo Boo Boo	Happy	1	3

مشابه دیتاست A مدلمان را میخواهیم در نظر بگیریم. (این دو سری سوال را مستقل از هم حل کنید.)

سوالات دیتاست B

- سه پارامتر را چگونه تعیین کنیم تا روی دیتاست بالا دقت ۱۰۰ درصد داشته باشیم. (w1, w2, b)
- چه شهودی روی این دیتاست بدست می آورید؟ (راهنمایی: آدم هایی که بیشتر صحبت میکنند خوشحال ترند یا کمتر؟ کدام کلمه بیشتر موثر است روی منفی شدن جمله)
- به جای یک گزاره شرطی (if ...) میتوانیم از activation function استفاده کنیم و در اصل از یک perceptron استفاده کنیم. مشخص کنید از کدام activation function استفاده کنیم؟ (شکل perceptron را رسم کنید)
- اگر میخواستیم بجای اینکه به صورت گسسته بگوییم جمله ناراحت است یا خوشحال، از چه activation function ای باید استفاده میکردیم تا ورودی را به احتمال شاد بودن جمله تبدیل میکردیم؟
- چه شهود هندسی میتوان برای این دیتاست متصور شد ؟
  - نقاط را در دو بعد رسم کنید. (در اصل هر جمله را میتوان با ۲ متغیر نشان داد) (انگار در این مدل ساده تنها تعداد ۲ کلمه برای ما اهمیت دارند، نه ترتیب و ...)
  - آیا به صورت خطی جدایی پذیر هستند؟
  - فضای دو بعدی را به دو قسمت تقسیم کنید و بگویید در کدام ناحیه جملات Happy و در کدام ناحیه جملات Sad هستند.

بیشتر بدانید

پردازش زبان های طبیعی جزییات قابل توجهی دارد. در صورت علاقه مندی میتوانید درباره شبکه های عصبی بازگشتی (RNN) ها ، LSTM, GRU و در نهایت درباره مفهوم Attention و مدل های Transformer base تحقیق کنید. در اصل خود ChatGPT هم روی همین معماری Transformer سوار است. نکاتی در پردازش زبان های طبیعی دارای اهمیت است. در پردازش زبان شما به رابطه بین کلمات باید توجه کنید. هر کلمه در Context خودش معنای متفاوتی را میتواند داشته باشد و مدل شما باید بتواند این موضوع را یاد بگیرد. موضوع مهم دیگری که وجود دارد بحث توجه (Attention) مدل است. به طور مثال در تسک های ترجمه ماشینی یا پرسش و پاسخ مدل باید بتواند در هر مرحله به بخش های خاصی از متن ورودی توجه ویژه ای داشته باشد. به طور مثال برای ترجمه یک متن طولانی در ابتدا مدل لازم نیست به جملات آخر متن توجه کند. این موضوع در سال ۲۰۱۷ در مقاله ای به نام "Attention Is All You Need" بیان شد.

در صورت داشتن هرگونه سوال مربوط به این سری به آیدی @omidiuuدر تلگرام پیام دهید.