به نام خدا

دانشکده مهندسی کامپیوتر

مبانی هوش مصنوعی ترم بهار ۰۰-۹۹

پاسخنامه تمرین پنجم : پردازش زبان طبیعی (فصل بیست و دوم)

سوال ۱

یک سیستم را در نظر بگیرید که برای تشخیص زبان استفاده می‌شود. این سیستم یک متن به عنوان ورودی گرفته و زبان آن را در خروجی نمایش می‌دهد. روش استفاده شده در این سیستم استفاده از مدل‌های احتمالاتی است.

الف) فرض کنید این سیستم فقط برای تشخیص زبان‌های فرانسه و عربی استفاده می‌شود. برای مدلسازی زبان-های این سیستم، کدام مدل بهتر است؟ مدل کلمه‌ای یا حرفی*؟ علت خود را توضیح دهید.

برای زمینه، دو جمله‌ی زیر به ترتیب فرانسوی و عربی می‌باشند:

FR: tout ce que tu avais à faire était de suivre le foutu train

AR: كل ما عليك فعله هو اتباع القطار اللعين

مدل حرفی؛ چون کاراکتر های دو زبان به طور واضح با هم تفاوت دارند و برای تشخیص این دو زبان نیازی به بررسی کلمات نیست.

ب) اگر این سیستم برای تشخیص زبان‌های عربی و فارسی استفاده شود، کدام مدل بهتر است؟ مدل کلمه‌ای یا حرفی؟ مدل unigram یا bigram؟ تحلیل کنید.

مدل کلمه‌ای: به دلیل تشابه الفبای فارسی و عربی، احتمال اینکه جمله‌های فارسی، کاراکترهایی را که در الفبای عربی نیستند داشته باشند، کم نیست به همین دلیل باید با کمک کلمات این دو زبان را تشخیص داد. برای مدل کلمه‌ای هر دو روش unigram و bigram می‌تواند روش مناسبی باشد، اما با توجه به اینکه کلمه‌های عربی و فارسی اشتراک زیادی ندارند، استفاده از unigram کارایی بهتری دارد و کافی است.

در صورتی که پاسخ شما با تحلیل مناسب مدل کلمه‌ای bigram باشد هم نمره کامل دریافت می‌کنید. توجه کنید که کلمات عربی که در فارسی استفاده می‌شود معمولا طرز نوشتار متفاوتی دارند (مثلا قطار و القطار)

ج) درباره مزایا و معایب مدل unigram و bigram و همین طور مدل کلمه‌ای و حرفی توضیح دهید.

مدل کاراکتری بدلیل دیکشنری محدود، سرعت بیشتری دارد اما همیشه نمی‌تواند یک مدل قوی برای تشخیص دو زبان/کلاس بسازد. مدل unigram نیز به دلیل دیکشنری محدودتر، از مدل bigram سریع تر است، اما برای تشخیص کلاس‌های نزدیک به هم (مثلا تشخیص طرز بیان دو سیاست‌مدار به زبان انگلیسی) از دقت کافی برخوردار نیست؛ چون کلمات را بدون context بررسی می‌کند.

* مدل حرفی (کاراکتری) بجای کلمات، با کاراکترها کار می‌کند. مثلا در مدل unigram کاراکتری:

$$P(\text{"the"}) = P(\text{"}t\text{"}) * P(\text{"}h\text{"}) * P(\text{"}e\text{"})$$

سوال ۲

فرض کنید پیام‌هایی که از زبان دو ربات ثبت شده به شکل زیر می‌باشد.

GLaDOS	101	1010	10101	110
CL4P-TP	1001	1000	0101	010

احتمال مخابره‌ی پیام از GLaDOS برابر ۰.۴ و CL4P-TP برابر ۰.۶ است.

الف) با استفاده از مدل unigram، کلمه 001 مربوط به کدام ربات است؟

$$P(0|GLaDOS) = \frac{6}{15} = 0.4$$

$$P(1|GLaDOS) = \frac{9}{15} = 0.6$$

$$P(001|GLaDOS) = P(0|GLaDOS) * P(0|GLaDOS) * P(1|GLaDOS)$$

$$= 0.4 * 0.4 * 0.6 = 0.96$$

$$P(GLaDOS|001) \propto P(001|GLaDOS) * P(GLaDOS) = 0.96 * 0.4 = \mathbf{0.384}$$

$$P(0|CL4PTP) = \frac{9}{15} = 0.6$$

$$P(1|CL4PTP) = \frac{6}{15} = 0.4$$

$$P(001|CL4PTP) = P(0|CL4PTP) * P(0|CL4PTP) * P(1|CL4PTP)$$

$$= 0.6 * 0.6 * 0.4 = 0.144$$

$$P(CL4P-TP|001) \propto P(001|CL4PTP) * P(CL4PTP) = 0.144 * 0.6 = \mathbf{0.864}$$

$$P(GLaDOS|001) < P(CL4PTP|001)$$

پس طبق مدل unigram مخابره‌ی ۰۰۱ از CL4P-TP بوده است.

ب) با استفاده از مدل bigram چطور؟ از هموارسازی لاپلاس استفاده کنید.

چون تعداد کاراکترهای یکتا ۴ تا است (۰ و ۱ و <s> و </s>) باید برای هموارسازی لاپلاس ۴ را به مخرج اضافه کنیم.

$$P(< s > 0|GLaDOS) = \frac{0+1}{4+4} = \frac{1}{8}$$

$$P(00|GLaDOS) = \frac{0+1}{6+4} = \frac{1}{10}$$

$$P(01|GLaDOS) = \frac{4+1}{6+4} = \frac{5}{10}$$

$$P(1 </s >) = \frac{2+1}{9+4} = \frac{3}{13}$$

$$P(001|GLaDOS) = \frac{1}{8} * \frac{1}{10} * \frac{5}{10} * \frac{3}{13} = 0.00144$$

$$P(GLaDOS|001) \propto P(001|GLaDOS) * P(GLaDOS) = 0.00144 * 0.4 = 0.000576$$

$$P(< s > 0|CL4PTP) = \frac{2+1}{4+4} = \frac{3}{8}$$

$$P(00|CL4PTP) = \frac{3+1}{9+4} = \frac{4}{13}$$

$$P(01|CL4PTP) = \frac{4+1}{9+4} = \frac{5}{13}$$

$$P(1 </s >) = \frac{2+1}{6+4} = \frac{3}{10}$$

$$P(001|CL4PTP) = \frac{3}{8} * \frac{4}{13} * \frac{5}{13} * \frac{3}{10} = 0.0133$$

$$P(CL4PTP|001) \propto P(001|CL4PTP) * P(CL4PTP) = 0.0133 * 0.6 = 0.0079$$

$$P(GLaDOS|001) < P(CL4PTP|001)$$

پس طبق مدل bigram هم مخابره‌ی ۰۰۱ از CL4P-TP بوده است.

سوال ۳

یک سیستم بازیابی اطلاعات ساده را در نظر بگیرید و به سوالات زیر پاسخ دهید.

الف) فرض کنید می‌خواهیم برای این موتور جستجو، مقادیر precision و recall را بدست آوریم. در این زمینه precision و recall را مقایسه کنید.

Precision: چه نسبتی از کل نتایج بازگردانده شده واقعا مربوط به کوئری داده شده هستند.

Recall: نتایج بازگردانده شده، چه نسبتی از کل نتایج مربوط به کوئری داده شده هستند.

ب) به نظر شما برای ارزیابی موتور جستجو، بهبود کدامیک از این دو معیار ها (precision/recall) اولویت بیشتری دارد؟ توضیح دهید.

توجه: پاسخ این سوال با توجه به تحلیل شما ممکن است متفاوت باشد.

در موتور جستجو تعداد کمتر False Negative اولویت بیشتری دارد؛ یعنی که موتور جستجوی ما می‌تواند مقدار زیادی از نتایج مربوط را بازگرداند. این یعنی recall با این توجیه اولویت بیشتری دارد.

اگر اولویت با False Positive کمتر باشد؛ یعنی موتور جستجو نتایج غلط کمتری برگرداند. یعنی اولویت با precision بهتر است. با این اولویت کاربران همه‌ی نتایج درست را نمی‌بینند ولی نتیجه‌ی غلط کمتری می‌بینند.

در واقعیت موتورهای جستجو بر اساس کارایی که دارند روی یکی از این دو اولویت بیشتری می‌گذارند ولی نمی-توان به طور قطع فقط یکی از این معیارها را انتخاب کرد.

ج) موتورهای جستجو برای برچسب زدن (false positive، true positive و ...) به برچسب‌ها دسترسی ندارند. یک روش برای استخراج این اطلاعات بر اساس رفتار کاربر ارائه دهید.

یک روش گرفتن زمانی است که کاربر در صفحه‌ی یک نتیجه باقی مانده است؛ هر چه این زمان کمتر باشد، یعنی نتیجه‌ی کمتر رضایت بخش بوده است. روش دیگر بررسی این است که کاربر پس از ورود به یک صفحه نتیجه، دیگر صفحه‌ی دیگری را باز نکرده است.

د) طبق جدول زیر مقادیر precision و recall را بدست آورید. آیا این نتایج برای ماژول جستجوی یک سایت خرید می‌تواند قابل قبول باشد؟ برای یک موتور جستجو چطور؟

Model prediction

Reality		Positive	Negative
	Positive	10	5
	Negative	3	12

$$Precision = \frac{10}{13}$$

$$Recall = \frac{10}{15}$$

برای ماژول سرچ سایت خرید، نشان دادن نتایج اشتباه ضرر زیادی ندارد و هدف نشان دادن تمام نتایج مربوط است (یعنی recall بهتر) در این مثال ۵ محصول مربوط بازگردانده نشده است و recall خوبی ندارد. اگر این ماژول صرفا برای پیشنهاد دهنده سرچبار بود که فقط چند نتیجه‌ی مربوط برتر را پیشنهاد می‌داد می‌توانست مورد قبول باشد، چون هدف پیشنهاددهنده سرچبار precision بهتر است.

برای موتور جستجو این نتایج می‌تواند قابل قبول باشد، چون از ۱۵ سند مربوط، ۱۰ تا بازگردانده شده و فقط ۳ نتیجه غلط داده شده است. با این حال با توجه به موضوع موتور جستجو این مقدار نیز می‌تواند مورد رضایت نباشد.

سوال ۴

می‌خواهیم یک دیتاست برای آموزش یک مدل طراحی کنیم که بصورت خودکار به نظرات یک وبسایت برچسب مثبت و منفی می‌دهد. برای استخراج اطلاعات این وبسایت از یک اسکریپر[1] استفاده کردیم ولی متاسفانه کامنت-های استخراج شده دارای ناخالصی می‌باشند. یکی از ابزارهای استخراج اطلاعات، regex ها هستند که با آن می-توان یک متن را فیلتر کرد یا قسمتی از آن را استخراج کرد.

یک مجموعه regex طراحی کنید که ناخالصی‌های متنی شبیه به متن زیر را از بین ببرد و کامنت خالص را برگرداند.

```
<p dir = "ltr">
      Wise man say: <q>Forgiveness is divine, but never pay full
      price for late pizza.</q>
      &#128544;
       
</p>
```

برای این کار می‌توانید از کتابخانه regex موجود در زبان دلخواه خود استفاده کنید.

توجه: کامنت داده شده باید بصورت جمله‌ی زیر برگردد (یعنی Wise man say هم جزو کامنت است):

```
Wise man say: Forgiveness is divine, but never pay full price for late pizza.
```

ولی چون به طور واضح ذکر نشده بود برگرداندن جمله‌ی زیر هم مورد قبول است:

```
Forgiveness is divine, but never pay full price for late pizza.
```

[1] Scraper

برای فیلتر تگ‌ها و علائم html می‌توان از regex زیر استفاده کرد:

```
<.*?>|&.*;
```

با این regex تمام <> ها و محتویات بین آنها انتخاب می‌شود و همینطور تمام &; ها با محتویات بینشان انتخاب می‌شود. (فرض کرده‌ایم که درمحتوای کامنت <> وجود ندارد)

سپس با یک زبان برنامه‌نویسی قسمت انتخاب شده را با "" replace می‌کنیم که آن‌ها را حذف کند و کامنت خالص را برگرداند.

امتیازی: سوال ۵

اکثر زبان‌های باستانی برای جداسازی کلمات از هم، از فاصله و نیم‌فاصله استفاده نمی‌کنند و جداسازی کلمات را به ذهن خواننده می‌سپارند. امروزه هم زبان‌هایی مثل تایلندی، از فاصله استفاده نمی‌کنند. تحقیق کنید که برای مسئله جداسازی کلمات یک جمله در یک بردار توسط کامپیوتر، می‌توان چه کارهایی انجام داد؟ (حداقل دو پاراگراف)

سوال تحقیقی