به نام خدا

دانشکده مهندسی کامپیوتر

مبانی هوش مصنوعی ترم بهار ۰۰-۹۹

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

مهلت تحویل ۱۱ تیر ۱۴۰۰

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

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.

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

* 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؟

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

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.

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

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.

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

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.

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)

[1] Scraper

توضیحات تکمیلی

- پاسخ به تمرین‌ها باید بصورت فردی انجام شود. در صورت مشاهده تقلب، نمره بین دو طرف تقسیم می‌شود.
- پاسخ خود را در یک فایل pdf بصورت خوانا در سامانه کورسز آپلود کنید.
- اگر سوال ۴ را با کد پیاده‌سازی کردید، از کد و نتیجه اسکرین شات بگیرید و در pdf اضافه کنید.
- فرمت نام‌گذاری تمرین باید مانند AI_HW5_6931420.pdf باشد.
- در صورت هرگونه سوال یا مشکل با ایمیل ce.ai.spring00@gmail.com یا آی‌دی @uramirbin در تماس باشید.
- ددلاین این تمرین **۱۱ تیر ۱۴۰۰ ساعت ۲۳:۵۵** است. هر روز تاخیر باعث کاهش **۲۵%** نمره‌ی دریافت شده می‌شود.