بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

پردازش زبان طبیعی

نیم‌سال دوم ۰۱ـ۰۰

استاد: احسان‌الدین عسگری

تمرین سوم | **مدل‌های زبانی** | مهلت ارسال: ۸ خرداد

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در تمرین‌هایی که چند ترک دارند، فقط یک نفر از هر گروه در سامانه CW باید ترک مورد نظر گروه را انتخاب کند. امکان تغییر ترک تا قبل از زمان ددلاین انتخاب ترک وجود دارد. البته ذکر این نکته ضروری است که هر ترک محدودیتی برای تعداد افرادی که آن را انتخاب می‌کنند، دارد. بنابرین در اسرع وقت برای انتخاب ترک اقدام کنید.
- در طول ترم امکان ارسال با تاخیر تمرین‌ها بدون کسر نمره تا سقف ۱۲ روز وجود دارد. محل بارگزاری جواب تمرین‌ها بعد از ۳ روز بسته خواهد شد و پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند شد. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۵ درصد از نمره تمرین کسر خواهد شد.
- توجه داشته‌باشید که نوت‌بوکهای شما باید قابلیت بازاجرای ۱۰۰ درصد داشته باشند و در صورت نیاز به نصب یک کتابخانه یا دسترسی به یک فایل، مراحل نصب و دانلود (از یک محل عمومی) در نوت‌بوک وجود داشته باشد.
- تمامی فایل‌های مرتبط به پروژه که حجم کمی دارند باید به شکل فایل زیپ در سامانه CW اپلود شوند. اگر حجم یک فایل زیاد بود (مانند فایل ذخیره شده یک مدل در صورتیکه بیش از ۲۰۰ مگابایت باشد)، تنها همان فایل را در یک محل عمومی، مثل گوگل درایو آپلود بفرمایید و لینک دانلود را در نوت‌بوک و مستندات قرار دهید.
- در پروژه‌های گروهی کافی است که فقط یکی از اعضای گروه پروژه را آپلود کند. اما حتما در گزارش کار نام همه اعضای گروه همراه با شماره دانشجویی آن‌ها آورده شود.
- بخشی از نمره شما به گزارش کار شما اختصاص دارد. در گزارش کار لازم نیست خط به خط کاری را که کرده‌اید توضیح دهید. بلکه باید به شکل کلی ایده‌تان برای حل مساله را شرح دهید. لازم است چند نمونه از خروجی‌های مساله را در گزارش بیاورید و براساس آن رفتار برنامه‌تان را تحلیل کنید. همچنین اگر پارامتری در صورت مساله خواسته شده (مانند دقت، صحت و ...) که در گزارش آورده شود شما باید آن را حساب کنید و در گزارش خود بیاورید.
- در صورت وجود هرگونه ابهام یا مشکل، در کوئرای درس آن مشکل را بیان کنید و از پیغام دادن مستقیم به دستیاران آموزشی خودداری کنید.

توضیحات کلی

در این تمرین شما به حل مسائل پردازش زبان به کمک ابزار مدل زبانی و جاسازی کلمه [1] می‌پردازید. این تمرین دارای ۴ ترک است.

[1] Word Embedding

تشخیص و تصحیح غلط‌های املایی متن

همانطور که در بخش ابتدایی درس مشاهده کردید یکی از روش‌های تصحیح غلط‌های املایی استفاده از فاصله‌ی ویرایشی[۲] است، هرچند فاصله ویرایشی دارای محدودیت‌های جدی است و لزوما نمی‌تواند تمام غلط‌های متن را اصلاح کند. یکی از مهم‌ترین روش‌هایی که می‌تواند کنار فاصله ویرایشی برای اصلاح متن قرار بگبرد استفاده از مدل زبانی است. برای مثال اگر بخواهید برای اصلاح جمله "دیوار حانل مستحکم نیست"، تنها از فاصله ویرایشی استفاده کنید کلمه "حانل" احتمالا به "حامل" تغییر می‌یابد درحالی که کلمه موردنظر "حائل" است. اما با اضافه کردن مدل زبانی احتمال اینکه شما به کلمه "حائل" دست یابید بالا می‌رود. در این تمرین شما باید با استفاده از مدل‌زبانی و فاصله ویرایشی برنامه‌ای را طراحی کنید که بتواند غلط‌های املایی متن را تا حد امکان بدرستی اصلاح کند. بدین‌منظور ورودی برنامه شما باید یک متن و خروجی آن اصلاح شده متن موردنظر به همراه غلط‌های املایی و محل آن‌ها و تصحیح‌شده غلط‌های املایی است.
در این ترک شما می‌توانید از مدل‌های پیش‌آموزش‌دیده استفاده کنید. داده‌ای نیز در اختیار شما قرار می‌گیرد ولی استفاده از این داده ضروری نیست و شما می‌توانید فقط از مدل‌های پیش‌آموزش‌دیده استفاده کنید هرچند اگر نیاز داشتید مدل‌زبانی را آموزش دهید یا میزان‌سازی روی مدل‌های فعلی انجام دهید می‌توانید از این داده استفاده کنید. البته تمرکز اصلی این ترک باید بر روی تصحیح غلط‌های املایی باشد. دادگان را می‌توانید از این لینک دریافت کنید.
همچنین اگر بتوانید در مدل خود اصلاح علائم نگارشی را نیز انجام دهید بسته به مقدار تلاش شما مقداری **نمره‌ی اضافی** به شما تعلق می‌گیرد.

به نکات زیر توجه فرمایید:

- خیلی از مواقع ممکن است کلماتی از متن شما غلط باشند اما این غلط به‌نحوی باشد که کلمه جدید خودش معنا داشته باشد که در این صورت هم کد شما باید بتواند شناسایی و تصحیح لازم را انجام دهد. برای مثال اگر کد شما جمله "دیوار حال مستحکم نیست" را دریافت کند هرچند که کلمه "حال" یک کلمه معنادار است اما بوضوح منظور کلمه "حائل" بود و حرف "ئ" جا افتاده است. در این صورت نیز کد شما باید به درستی خطا را شناسایی و اصلاح کند.

- شما برای انجام این تمرین باید از **حداقل دو مدل زبانی** استفاده کنید که یکی از آن‌ها باید مدل زبانی تبدیلگر[۳] باشد. البته می‌توانید هر دو مدل زبانی را به‌صورت ترکیبی نیز استفاده کنید.

- شما در این ترک عملا باید از مدل‌زبانی برای تشخیص و تصحیح غلط‌های املایی استفاده کنید و فاصله ویرایشی صرفا یک هیوریستیک پیشنهادی کنار مدل‌زبانی است. خودتان نیز می‌توانید از هیوریستیک بهتری استفاده کنید و تا زمانی که روش شما منطقی باشد و دقت مدل‌تان پایین نیاید مجاز به انجام هر کاری هستید.

[2] Edit Distance
[3] Transformer

ورودی	خروجی

پس از سال‌ها تلاش رازی موفق به کسف الکل شد. این دانشمند تیرانی باعث افتخار در تاریخ کور است.	[
	{
	"raw": "کسف",
	"corrected": "کشف",
	"span": [31,34]
	},
	{
	"raw": "تیرانی",
	"corrected": "ایرانی",
	"span": [56,62]
	},
	{
	"raw": "کور",
	"corrected": "کشور",
	"span": [84,87]
	}
	]

بسیاری از مباحث علوم غیرطبیعی با استفاده از فیریک دنیای مادی ابل توجیح نیست و برای یادگیری باید به فلسفه‌های خاصی رجو کرد.	[
	{
	"raw": "فیریک",
	"corrected": "فیزیک",
	"span": [44, 49]
	},
	{
	"raw": "ابل",
	"corrected": "قابل",
	"span": [61, 64]
	},
	{
	"raw": "توجیح",
	"corrected": "توجیه",
	"span": [65, 70]
	},
	{
	"raw": "رجو",
	"corrected": "رجوع",
	"span": [115, 118]
	}
	]

فضای جاسازی دوزبانه

در درس مشاهده کردید که جاسازی یک کلمه حاوی معنای آن کلمه می‌باشد. در واقع با محاسبه بردار معنایی یک عبارت یا یک کلمه،ما آن‌ها را در فضای معنایی خواهیم داشت. با کمک بردار معنایی می‌توان جست‌وجوی معنایی انجام داد. برای مثال توقع می‌رود که با جست‌وجوی بردار سیب به بردار پرتقال به عنوان یک بردار شبیه برسیم. چون هر دو میوه هستند و در جملات به جای همدیگر می‌توانند به کار روند.
یکی از حالت‌های پیشرفته‌تر این جست و جو، میان چند زبان است. مشکلی که وجود دارد فضای معنایی جاسازی‌های آموزش دیده یکسان نیست و قطعا بین دو زبان متفاوت کاملا متفاوت خواهد بود. روش‌هایی وجود دارد که این فضای معنایی را مشترک می‌کند. در این ترک قرار است که شما در ابتدا جاسازی کلمات انگلیسی و فارسی را محاسبه کنید. سپس فضای معنایی دو زبان را یکی کنید و یک جست‌وگر معنایی دو زبانه بسازید.
اندازه بردارهای جاسازی را برابر با ۱۰۰ در نظر بگیرید. شما باید گام‌های زیر را برای انجام ترک انجام دهید:

۱. داده ورودی شما داده قرآنی است و شما حق انتخاب دارید که از زبان‌های فارسی، انگلیسی یا عربی دو تا از آن‌ها را انتخاب کرده و بر اساس آن پروژه را انجام دهید. برای جمع‌آوری داده زبان‌های انگلیسی و فارسی از این لینک استفاده کنید. می‌توانید برای آموزش بهتر بردار معنایی از تمام ترجمه‌های یک زبان به عنوان ورودی استفاده کنید.

۲. باید با استفاده از skip-gram بردارهای جاسازی کلمات دو زبان را محاسبه کنید. برای مشاهده نمونه کد می‌توانید به این لینک مراجعه کنید.

۳. در این پروژه برای ساده‌سازی، تبدیل فضای معنایی با استفاده از یک تبدیل خطی انجام می‌دهید. برای مطالعه سایر روش‌ها به این لینک مراجعه‌کنید.
با استفاده یک مدل شبکه عصبی تک لایه خطی که از فرمول زیر تبعیت می‌کند و با کمک لیستی از بردارهای معادل زبان، نزدیک‌ترین تابع تبدیل بین این دو فضای معنایی را پیدا کنید. (راهنمایی: برای مثال در داده‌های قرآنی می‌توانید بردار معادل هر آیه را حساب کنید و از آن‌ها استفاده کنید)

$$Wx + b$$

به نکات زیر در مورد این ترک توجه داشته باشید:

۱. اگر بر روی هر سه زبان این کار را انجام دهید (دو زبان دیگر را به فضای برداری زبان سوم ببرید) نمره امتیازی به شما تعلق می‌گیرد.

۲. باید کدهایی که برای آموزش جاسازی کلمه زده‌اید همراه با پروژه اپلود شوند. اما در فایل main پروژه که تست نهایی با آن انجام می‌شود باید بردارهایی هایی که قبلا آموزش داده‌اید را فقط بارگزاری کنید.

۳. در واقع شما در این ترک سه مدل را آموزش می‌دهید (بردارهای معنایی زبان اول، بردارهای معنایی زبان دوم و مدل تبدیل یکی به دیگری) که فایل وزن‌های مدل آموزش دیده هر کدام باید همراه پروژه‌تان آپلود شود.

۴. توابع محاسبه جاسازی کلمتان طبیعتا باید tokenization را قبل از محاسبه انجام داده باشد. بعد از محاسبه جاسازی هر توکن می‌توانید با یک میانگین گرفتن ساده بردار معنایی جمله را محاسبه کنید.

۵. پیشنهاد می‌شود که در هنگام آموزش دادن تابع تبدیل بین دو فضای معنایی، بردارهای ورودی و خروجی را نرمال کنید.

$$||\text{Embedding}||_2 = 1$$

۶. در گزارش خود مقدار شباهت کسینوسی (ضرب داخلی جبری) بردارهای آیات یکسان در دو زبان را، با آیات متفاوت بررسی کنید. (۵ مثال برای مقایسه کافی است)

برای گرفتن ایده کلی فایل اصلی پروژه می‌توانید به شبه کد زیر در پایتون دقت کنید:

```
en_emb_model = load_model(...)
fa_emb_model = load_model(...)
fa_to_en_model = load_model(...)

def encode_en(sentence: str):
    tokens = tokenize_en(sentence)
    embs = []
    for token in tokens:
        emb.append(en_emb_model.encode(token))
    return_value = np.average(embs, axis=0)
    return return_value / np.linalg.norm(return_value)

def encode_fa(sentence: str):
    tokens = tokenize_fa(sentence)
    embs = []
    for token in tokens:
        emb.append(fa_emb_model.encode(token))
    return_value = np.average(embs, axis=0)
    return return_value / np.linalg.norm(return_value)

def convert_emb_from_fa_to_en(fa_emb):
    return fa_to_en_model(fa_emb)

def encode_en_same_space(sentence: str):
    return encode_en(sentence)

def encode_fa_same_space(sentence: str):
    return convert_emb_from_fa_to_en(encode_fa(sentence))
```

پیدا کردن مشابه‌ترین آیات قرآن و عبارت‌های کتب مقدس دیگر از روی عبارت ورودی

این قسمت شامل ۳ بخش است که به روش های متفاوت نمایش کلمه می‌پردازد و با استفاده از آن‌ها، سعی در پیدا کردن شباهت عبارت ورودی با آیات قرآن/خطبه‌ها، نامه‌ها وحکمت‌های نهج البلاغه/دعاهای صحیفه سجادیه/عبارات کتب مقدس می‌کند و در نهایت مواردی که دارای نزدیک‌ترین جاسازی به جاسازی عبارت ورودی است را به عنوان مشابه‌ترین موارد برمی‌گرداند.
شما میتوانید از "داده‌های قرآنی، نهج‌البلاغه و صحیفه سجادیه" یا از "داده‌های قرآنی و کتاب‌های مقدس" استفاده کنید. در هر دو صورت می‌بایست از نسخه عربی داده‌های مذکور استفاده کنید. در صورت تمایل می‌توانید از مدل‌های چند زبانه برای پشتیبانی از هر دو زبان فارسی و عربی نیز استفاده کنید که در این صورت با توجه به کارایی مدل، نمره اضافی در نظر گرفته خواهد شد.

۱. در روش اول با محاسبه tfidf در سطح ۱)کلمه و ۲)کاراکتر، شبیه ترین موارد را محاسبه می‌کنید.

۲. در روش دوم با استفاده از بردارهای آماده FastText ، شبیه ترین موارد را محاسبه می‌کنید.

۳. در روش سوم با استفاده از بردارهای آماده و Fine-Tune کردن آن ها با استفاده از داده های معرفی شده، شبیه ترین موارد را محاسبه می‌کنید.

ـ جاسازی‌های ویکیپدیا عربی و فارسی مدل fasttext را می‌توانید از لینک‌های قرار داده شده دانلود کنید. همچنین برای دسترسی به تمامی جاسازی‌ها میتوانید به وب‌سایت fasttext مراجعه بفرمایید.

ـ برای استفاده و آموزش جاسازی‌های fasttext در پایتون، می‌توانید در این لینک راهنمای پکیج Gensim را مطالعه کنید.

ـ برای دسترسی به داده‌های قرآنی می‌توانید از سایت تنزیل استفاده کنید.

ـ همچنین می‌توانید داده‌های مورد نظر خود را در صفحه گیت‌هاب درس پیدا کنید. داده‌های کتب مقدس به گروه‌های انتخاب کننده این قسمت تحویل داده خواهد شد.

تکمیل مصراع دوم در بیت با رعایت وزن شعر

هدف این تمرین استفاده مناسب از مدل‌های زبانی در تولید مصراع دوم یک بیت می‌باشد بدین شکل که مدل نهایی شما باید قادر باشد با گرفتن یک تعداد کلمه در ورودی خود که نماینده یک مصراع هستند ، جمله ای تولید کند که بتوان به عنوان مصراع دوم پذیرفت؛ به عنوان مثال :
اگر جمله‌ی «گاه می گویند دریا خاک شد» به عنوان ورودی داده شود در خروجی مصراع‌هایی مانند زیر تولید شود:
* گاه اندر موج ما چالاک شد
* گاه در گرداب ما غمناک شد
* گاه گویند دریا چو گوهر پاک شد
(تمامی مصراع‌های مثال زده شده توسط ماشین تولید شده‌اند)

برای ایجاد چنین مدلی شما نیاز به یک دیتاست اشعار دارید. می‌توانید از دیتاست اشعار فارسی که از سایت گنجور جمع آوری شده است استفاده کنید. در انتخاب دادگان مجاز هستید از اشعار هر شاعر دلخواهی استفاده کنید دقت داشته باشید اگر از چند شاعر مختلف استفاده می‌کنید حتما به سبک شاعرهای انتخاب شده دقت کنید همچنین می‌توانید مدل خود را به گونه‌ای تنظیم نمایید که به جای تولید مصراع دوم مصراع جاری را تکمیل نماید.

نکات پیاده سازی:

۱. برای شروع و پایان مصراع‌ها از نمادهایی همچون __BOM__ و __EOM__ استفاده کنید. همچنین در صورتی که نیاز به یکسان سازی طول مصراع‌ها وجود دارد می‌توانید از نماد __PAD__ استفاده کنید.

۲. رعایت قافیه یا عدم رعایت قافیه را به عنوان یک پارامتر ورودی در نظر بگیرید.

۳. وارد کردن عنصر وزن شعر در خروجی نمره اضافی دارد.

تحویل دادنی و مقایسه مدل‌ها با یکدیگر:

۱. با استفاده از یک مدل N-Gram ادامه مصراع یا مصراع بعدی را تولید کنید.

۲. با استفاده از یک مدل encoder-decoder که encoder آن یک شبکه LSTM است ادامه مصراع یا مصراع بعدی را تولید کنید.

۳. با استفاده از یک مدلی که از مکانیزم توجه استفاده می‌کند اقدام به تکمیل مصراع یا مصراع بعدی کنید. به این منظور می‌توانید از مدل‌هایی نظیر Bert، GPT یا دیگر مدل‌هایی از این دست استفاده کنید.