



پردازش زبان طبیعی

نیم سال دوم ۰۱-۰۲

مدرس: احسان الدین عسگری

تمرین سوم

مدل‌های زبانی

مهلت ارسال: ۲۹ اردیبهشت

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص شده است.
- در تمرین‌هایی که چند چالش دارند، فقط یک نفر از هر گروه در سامانه CW باید چالش مورد نظر گروه را انتخاب کند. امکان تغییر چالش تا قبل از زمان ددلاین انتخاب چالش وجود دارد. البته ذکر این نکته ضروری است که هر چالش محدودیتی برای تعداد افرادی که آن را انتخاب می‌کنند، دارد. بنابراین در اسرع وقت برای انتخاب چالش اقدام کنید.
- در طول ترم امکان ارسال با تاخیر تمرین‌ها بدون کسر نمره تا سقف ۱۲ روز وجود دارد. محل بارگزاری جواب تمرین‌ها مطابق زمان مشخص شده در تقویم، 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 داشته باشید 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 آپلود 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 مواردی از این دست) که در گزارش آورده شود شما باید آن را حساب کنید و در گزارش خود بیاورید.
- کد نهایی شما باید یک کلاس شامل تابع run داشته باشد که برای هر چالش طبق ورودی‌های مشخص شده، خروجی موردنظر را بدهد. شما می‌توانید به کمک روند معرفی شده در فایل CONTRIBUTION کدهای خود را به کتابخانه‌ی parsi.io اضافه کنید. در صورتی که pull-request شما پذیرفته شود نمره امتیازی به شما تعلق خواهد گرفت.
- دقت داشته باشید، موارد امتیازی که در این تمرین آمده است، صرفاً بر روی امتیاز همین تمرین اثر دارد و بر روی نمرات تمارین و یا بخش‌های دیگر درس، تأثیر ندارد.
- در صورت وجود هرگونه ابهام یا مشکل، در کوئرای درس آن مشکل را بیان کنید و از پیغام دادن مستقیم به تیم تدریس خودداری کنید.

توضیحات کلی

در این تمرین شما به حل مسائل پردازش زبان به کمک ابزار مدل زبانی و جاسازی کلمه^۱ می‌پردازید. این تمرین دارای ۹ ترک می‌باشد.

نکته: نمره‌ای که برای انجام موارد "امتیازی" اشاره شده در ترک‌ها در نظر گرفته می‌شود، صرفاً برای جبران نمره‌های کسر شده احتمالی در انجام این تمرین می‌باشد و نمره نهایی اخذ شده نمی‌تواند بیشتر از نمره تمرین باشد.

¹ Word Embedding

کامل کردن کلمه جاری در یک دامنه مشخص

در این تمرین شما قرار است متناسب با کلمات قبلی یک جمله، کلمه بعدی یا کلمه فعلی را تکمیل نمایید. بدین منظور از این [دیتاست](#) که از سایت‌های «نمناک» و «های‌دکتر» جمع‌آوری شده است، می‌توانید استفاده نمایید.

ورودی	خروجی
احتمال افزایش تورم در سال	آینده
احتمال افزایش تورم در سال آ	ینده
بهترین روش برای غلبه بر استرس	پیاده‌روی
بهترین روش برای غلبه بر استرس ن	وشیدن
بهترین روش برای غلبه بر استرس نوشیدن	قهوه

برای این انجام این بخش تمامی موارد زیر بایستی تکمیل شود:

- از مدل زبانی ^۲ n-gram به عنوان یک مدل پایه استفاده نمایید، توجه داشته باشید که مدل شما باید در سطح (۱) کاراکتر و (۲) کلمه قادر به کامل کردن جمله باشد.
 - از مدل‌ها^۳ به عنوان یک مدل عمیق استفاده نمایید، توجه داشته باشید که مدل شما باید در سطح (۱) کاراکتر و (۲) کلمه قادر به کامل کردن جمله باشد. برای تنظیم دقیق^۴ کردن مدل عمیق، روش مناسبی را با تحقیق در مورد فرآیندها^۵ ارائه دهید. به عنوان نمونه فریز کردن برخی لایه‌ها و همچنین از استفاده از نرخ یادگیری متفاوت برای لایه‌های متمایز شبکه عمیق از جمله روش‌های تنظیم دقیق باشد. بدین منظور می‌توانید از این [مقاله](#) کمک بگیرید.
 - مدل عمیقی را طوری آموزش دهید که بتواند تا دو کلمه پیش‌رو را، پیش‌بینی و تکمیل نماید. به منظور آشنایی بیشتر با روش‌های رمزگشایی
 - برای داشتن اطلاعات بیشتر و آشنایی بیشتر شما با این ترک، یک نمونه گزارش به همراه کد از دانشجویان ترم پیش در این [لینک](#) در دسترس شما قرار گرفته است. می‌توانید از گزارش و کد مورد نظر، برای بهبود کار خودتان نسبت به کار پیشین انجام شده استفاده کنید. صرفاً کپی کردن کدها بدون داشتن ایده‌ی جدید باعث از دست دادن نمره می‌شود.^۶ برای تولید متن، می‌توانید به این [لینک](#) مراجعه کنید.
- در صورت عملکرد مطلوب مدل شما، نمره کامل به شما تعلق خواهد گرفت. در صورتی که مدل بتواند تعداد قابل قبولی از کلمات را در جملات مفهومی و پیچیده‌تر کامل نماید، نمره امتیازی به شما تعلق خواهد گرفت.
- برای راحتی آزمایش مدل توسعه داده شده توسط شما، یک واسط مبتنی بر فلسفک توسعه داده شده که در این نشانی [گیت‌هاب](#) موجود است و می‌توانید از این واسط استفاده کنید (استفاده از این واسط اجباری نیست).

^۲Language Model

^۳Transformers

^۴Fine-tune

^۵Hyper-parameter

^۶Decoding

با افزایش تعداد داروها به یاد آوردن نام داروها مشکل شده است. گاهی اوقات با دانستن قسمتی از نام دارو به دنبال دارو هستیم. در اینجا مایل هستیم ابزاری آماده کنیم تا بتوانیم با جست‌وجو نام یا فضای مربوطه‌ی آن، داروی مد نظر را پیدا کنیم.

یکی از حالت‌های پیشرفته‌تر این جست‌وجو، میان چند زبان است. 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 جست‌وجوگر دو زبانه در حوزه دارو بسازید.

شما باید گام‌های زیر را برای انجام ترک انجام دهید:

۱. داده ورودی شما از جنس دارو است و شما باید لیستی از این داده‌ها به زبان انگلیسی و فارسی داشته باشید. در [این لینک](#) یک داده نمونه وجود دارد.

۲. 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](#) مراجعه کنید.

۴. **امتیازی:** با جست‌وجو کاربرد دارو مانند بهبود سردرد بتوانیم داروهای مورد نظر را بازیابی کنیم.

به نکات زیر در مورد این ترک توجه داشته باشید:

۱. باید کدهایی که برای آموزش جاسازی کلمه زده‌اید همراه با پروژه آپلود شوند. اما در فایل main پروژه که تست نهایی با آن انجام می‌شود باید بردارهایی که قبلاً آموزش داده‌اید را فقط بارگذاری کنید.

۲. 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.

۳. توابع محاسبه جاسازی کلمات طبعاً 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.

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

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

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

۶. برای داشتن اطلاعات بیشتر و آشنایی بیشتر شما با این ترک، یک نمونه گزارش به همراه کد از دانشجویان ترم پیش در [این لینک](#) در دسترس شما قرار گرفته است. می‌توانید از گزارش و کد مورد نظر، برای بهبود کار خودتان نسبت به کار پیشین انجام شده استفاده کنید. صرفاً کپی کردن کدها بدون داشتن ایده‌ی جدید باعث از دست دادن نمره می‌شود.

پیدا کردن واحدهای مشابه قرآن، نهج البلاغه و صحیفه سجاده در کتاب مقدس و برعکس

در این ترک، شما با استفاده از مدل‌زبانی و بردارهای جاسازی، به یافتن عبارات مشابه موجود در قرآن کریم، نهج البلاغه و صحیفه سجاده در **کتاب مقدس** (مجموعه عهد قدیم و جدید) و برعکس می‌پردازید. این کار به سه روش زیر می‌بایست انجام شود:

۱. استفاده از مدل‌زبانی از پیش آموزش داده شده fasttext

۲. استفاده از بردارهای کلمات هم‌تراز شده ^۷ fasttext

۳. استفاده از مدل‌زبانی **غیروابسته به زبان** ^۸ LaBSE

بدین منظور گام‌های زیر را انجام دهید (برخی از گام‌ها صرفاً در بعضی از روش‌های بالا نیاز به طی شدن دارند که در ابتدای هر مورد ذکر می‌شود)

۱. (صرفاً در روش سوم) مدل زبانی از پیش آموزش دیده شده LaBSE را روی متن عربی **قرآن**، **نهج البلاغه**، **صحیفه‌سجاده** و **ترجمه انگلیسی کتاب مقدس** تنظیم دقیق ^۹ کنید.

۲. (صرفاً در روش اول) یک تبدیل فضای معنایی از امبدینگ fasttext عربی به امبدینگ fasttext انگلیسی آموزش دهید. در این جا این تبدیل فضای معنایی با استفاده از یک تبدیل خطی انجام می‌شود (برای مطالعه سایر روش‌ها، می‌توانید به **این لینک** مراجعه کنید). برای آموزش این تبدیل، با استفاده از یک مدل شبکه عصبی تک لایه خطی که از فرمول زیر تبعیت می‌کند و با کمک لیستی از بردارهای معادل در دو زبان، نزدیک‌ترین تابع تبدیل بین این دو فضای معنایی را پیدا کنید. این تبدیل را می‌توانید با داده‌های موازی عربی - انگلیسی آموزش دهید؛ به طور مثال می‌توانید با استفاده از ترجمه انگلیسی **قرآن**، بردار معادل هر آیه را حساب کرده و از آن‌ها استفاده کنید.

$$Wx + b$$

۳. برای تمامی واحدهای موجود در قرآن، نهج البلاغه و صحیفه سجاده با استفاده از مدل‌زبانی عربی و تبدیل خطی یادگرفته شده و همچنین تمامی واحدهای کتاب مقدس با استفاده از مدل‌زبانی انگلیسی، بردار جاسازی بدست آورید. یک روش برای این کار در روش‌های اول و دوم، میانگین‌گیری از بردار جاسازی‌های تمامی کلمات موجود در یک واحد است. اما برای اینکه در این دو روش نتیجه بهتری بدست بیاورید، می‌بایست با محاسبه ^{۱۰} IDF هر کلمه (در داده‌هایی که قبلاً معرفی شد) و در نظر گرفتن آن عدد به عنوان ضریب کلمه، **میانگین وزن‌دار** محاسبه کنید.

۴. تابعی بنویسید که یک واحد از قرآن، نهج البلاغه یا صحیفه سجاده را به عنوان ورودی دریافت کند. سپس با استفاده از مدل‌زبانی عربی و تبدیل خطی یادگرفته شده، بردار جاسازی مربوط به عبارت ورودی را در فضای برداری انگلیسی تولید کند. سپس با مقایسه آن بردار با بردار واحدهای موجود در کتاب مقدس، ۱۰ تا از نزدیک‌ترین واحدهای موجود در کتاب مقدس را به عنوان خروجی برگرداند. برای یافتن نزدیک‌ترین بردار، می‌توانید از شباهت کسینوسی استفاده کنید.

۵. مشابه مورد قبل، تابع دیگری بنویسید که این بار یک واحد از کتاب مقدس را دریافت و پس از تولید بردار جاسازی آن در فضای برداری انگلیسی، ۱۰ تا از نزدیک‌ترین واحدهای موجود در هر یک از سه کتاب قرآن، نهج البلاغه و صحیفه سجاده (مجموعاً ۳۰ واحد) به عنوان خروجی برگرداند.

⁷ Aligned Word Vector

⁸ Language-agnostic

⁹ Fine tune

¹⁰ Inverse Document Frequency

به نکات زیر توجه کنید:

- منظور از "واحد" در هر یک از کتاب‌ها عبارت است از:
قرآن: یک آیه
نهج‌البلاغه: یک خطبه، یک نامه یا یک حکمت
صحیفه سجادیه: یک خط که با x##y شروع می‌شود
کتاب مقدس: یک verse
- برخی از واحدهای موجود در نهج‌البلاغه طولانی هستند. برای اینکه نتایج بهتری بدست بیاورید، می‌توانید آن‌ها را به واحدهای کوچکتری تقسیم کنید. انجام مناسب این مورد، نمره امتیازی دارد.
- کتاب مقدس از دو مجموعه تشکیل می‌شود، "عهد قدیم"^{۱۱} و "عهد جدید"^{۱۲} که نسخه اصلی آن‌ها به ترتیب به زبان **عبری** و **یونانی** است. در این ترک برای ساده‌سازی، از شما خواسته شده است که از ترجمه انگلیسی هر دو مجموعه استفاده کنید. اما همان‌طور که می‌دانید، معمولاً ترجمه یک کتاب به طور کامل نمی‌تواند شامل تمامی مفاهیم موجود در نسخه اصلی باشد. از این رو بهتر است مراحل ذکر شده در این ترک را برای سه زبان عربی، عبری و یونانی انجام داد. انجام این مورد پیشنهاد شده و نمره امتیازی دارد. (علاوه بر لینک‌های ذکر شده برای متن کتاب مقدس، برای دسترسی به سایر نسخه‌ها و زبان‌های این کتاب، می‌توانید به **این لینک** مراجعه کنید.)
- شما می‌بایست در هنگام آموزش دادن تابع تبدیل بین دو فضای معنایی در روش اول و همچنین محاسبه بردارهای جاسازی برای هریک از واحدها در روش‌های اول و دوم، تمامی بردارها را نرمال کنید:

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

- در گزارش خود، خروجی توابعی که در گام‌های ۴ و ۵ نوشته اید را برای چند نمونه متنوع از هر دو داده، محاسبه و بررسی کنید.
- در گزارش خود، مقدار شباهت کسینوسی بردارهای واحدهای معادل عربی و انگلیسی که در یادگیری تبدیل خطی استفاده کرده بودید را با واحدهای متفاوت بررسی کنید (صرفاً در روش اول؛ ۵ مثال برای مقایسه کافی است).
- تمامی کدهایی که برای آموزش تبدیل خطی، بازیابی واحدهای مشابه و ... نوشته‌اید را به همراه گزارش آپلود کنید.

¹¹Old Testament

¹²New Testament

فضای جاسازی مشترک بین متن و تصویر

در مدل‌های زبانی، معمولاً سعی می‌شود امبدینگ‌ها به صورتی تعریف شوند که جملات با معنی مشابه در این فضا در نزدیکی یکدیگر قرار بگیرند به طوری که برای جمله‌ها یا کلمه‌های مشابه خروجی امبدینگ به یکدیگر نزدیک باشد. هدف این تمرین ایجاد مدل امبدینگ برای متن و تصویر یا متن و صوت است.

نمونه مدل‌ها

کارهای زیادی مانند مدل‌های جستجو چند رسانه‌ای^{۱۳}، ایجاد تصویر از متن و ایجاد عنوان برای تصویر نیازمند فضای امبدینگ مشترک برای تصویر و متن هستند به صورتی که تصاویر مشابه با متن نزدیک یکدیگر باشند. نمونه‌های چنین مدلی برای زبان‌های مختلف وجود دارد^{۱۴}. شما می‌توانید از مدل‌های موجود در زبان‌های دیگر برای طراحی و آموزش مدل خود در زبان فارسی کمک بگیرید.

مجموعه دادگان

برای آموزش مدل امبدینگ تصویر و متن یا صوت و متن نیازمند داده‌های موازی در این دو فضا هستیم. شما در این تمرین مجاز به استفاده از مجموعه داده‌های موجود زبان فارسی هستید. با این وجود، ایجاد داده نیز تشویق می‌شود و در صورت نیاز فضای لازم برای این کار نیز توسط تیم درس برایتان آماده می‌شود. برای مدل تصویر و متن می‌توانید از داده‌های [این لینک](#) و برای مدل صوت و متن از داده‌های [این لینک](#) می‌توانید استفاده کنید.

خروجی‌ها

مدل شما باید موارد زیر را پوشش دهد.

- برای مدل متن و تصویر باید ورودی مدل متن یا تصویر باشد. (به طور مشابه برای مدل متن و صوت باید ورودی مدل متن یا صوت باشد)
- خروجی مدل یک آرایه‌ی امبدینگ باشد.
- خروجی مدل برای تصاویر و متن‌های مشابه باید فاصله‌ی اقلیدسی کمی داشته باشد.
- خروجی مدل برای تصاویر و متن‌های غیر مشابه نباید فاصله‌ی اقلیدسی نزدیک باشند.
- **امتیازی** در صورتی که مدل شما تصویر و متن را به طور همزمان ورودی گرفت باید بتواند در تصویر قسمتی که توسط متن توصیف شده است را پیدا کند. همچنین اگر مدل شما برای فضای صوت و متن ایجاد شده است باید بتواند با ورودی گرفتن صوت و متن، بازه‌ی صوتی که بیشترین شباهت به متن داده شده را دارد را به عنوان خروجی بدهد.
- **امتیازی** مدلی برای نزدیک کردن امبدینگ متن به امبدینگ نقشه در یک کاربرد آموزش دهید. به عنوان مثال، در نقشه‌های هواشناسی توضیحاتی مرتبط با وضعیت بارش و سرعت باد و ... از روی داده‌ای که از نقشه استخراج می‌شود، وجود دارد. شما می‌توانید با یافتن دیتاستی شامل نقشه‌های هواشناسی و توضیح متنی مرتبط با آن‌ها، امبدینگ‌های نزدیک به هم برای نقشه و متن تولید کنید. این مورد قابل تعمیم به کاربردهای پیشنهادی دیگر از سوی شما مانند یافتن امبدینگ نزدیک به هم برای نقشه و پوشش گیاهی موجود در آن و ... می‌باشد.

¹³Multi Modal

¹⁴[Open AI CLIP](#)

- برای داشتن اطلاعات بیشتر و آشنایی بیشتر شما با این ترک، یک نمونه گزارش به همراه کد از دانشجویان ترم پیش در [این لینک](#) در دسترس شما قرار گرفته است. می‌توانید از گزارش و کد مورد نظر، برای بهبود کار خودتان نسبت به کار پیشین انجام شده استفاده کنید. صرفاً کپی کردن کدها بدون داشتن ایده‌ی جدید باعث از دست دادن نمره می‌شود.

تشخیص و تصحیح غلط‌های املائی متن برای زبانهای ایرانی

همانطور که در بخش ابتدایی درس مشاهده کردید یکی از روش‌های تصحیح غلط‌های املائی استفاده از فاصله‌ی ویرایشی^{۱۵} است، هرچند فاصله‌ی ویرایشی دارای محدودیت‌های جدی است و لزوماً نمی‌تواند تمام غلط‌های متن را اصلاح کند. یکی از مهم‌ترین روش‌هایی که می‌تواند کنار فاصله‌ی ویرایشی برای اصلاح متن قرار بگیرد استفاده از مدل زبانی است. برای مثال اگر بخواهید برای اصلاح جمله “دیوار حائل مستحکم نیست”، تنها از فاصله‌ی ویرایشی استفاده کنید کلمه “حائل” احتمالاً 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 مدل زبانی و فاصله‌ی ویرایشی برنامه‌ای را طراحی کنید که بتواند غلط‌های املائی متن را تا حد امکان بدرستی اصلاح کند. بدین منظور ورودی برنامه شما باید یک متن و خروجی آن اصلاح شده متن موردنظر به همراه غلط‌های املائی و محل آن‌ها و تصحیح‌شده غلط‌های املائی 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 مدل زبانی را آموزش دهید یا تنظیم دقیق^{۱۶} روی مدل‌های فعلی انجام دهید می‌توانید از این داده استفاده کنید. البته تمرکز اصلی این ترک باید بر روی تصحیح غلط‌های املائی باشد. دادگان را می‌توانید از این [لینک](#) دریافت کنید.

به نکات زیر توجه فرمایید:

- خیلی از مواقع ممکن است کلماتی از متن شما غلط باشند اما این غلط به نحوی 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^{۱۷} باشد. البته می‌توانید هر دو مدل زبانی را به صورت ترکیبی نیز استفاده کنید.

- علاوه بر حداقل دو مدل زبانی یک مدل پایه^{۱۸} هم باید تست شود. مدل پایه به این صورت است که شما با استفاده از یک مجموعه از کلمات فارسی و فاصله‌ی ویرایشی سعی در تصحیح متن موجود می‌کنید. سپس عملکرد دو مدل قبلی با این مدل پایه مقایسه خواهد شد.

- کارایی کد شما از لحاظ دقت با معیار f۱ سنجیده می‌شود. بدین منظور شما باید یک مجموعه داده تست درست کنید و دقت مدل‌های خود را روی آن بررسی کنید. مجموعه تست شما باید حداقل دارای ۱۵ جمله باشد. این ۱۵ جمله را به همراه متن تصحیح‌شده نیز در فایل خود به منظور ارزیابی انسانی نیز قابل مشاهده باشد (تصحیح‌شده هر جمله زیر خود جمله اصلی).

- شما در این ترک عملاً باید از مدل زبانی برای تشخیص و تصحیح غلط‌های املائی استفاده کنید و فاصله‌ی ویرایشی صرفاً یک مکاشفه^{۱۹} پیشنهادی کنار مدل زبانی است. خودتان نیز می‌توانید از مکاشفه بهتری استفاده کنید و تا زمانی که روش شما منطقی باشد و دقت مدل‌تان پایین نیاید مجاز به انجام هر کاری هستید.

¹⁵Edit Distance

¹⁶Fine tune

¹⁷Transformer

¹⁸Baseline

¹⁹Heuristic

خروجی	ورودی
<pre>[{ "raw": "كسف", "corrected": "كشف", "span": [31,34] }, { "raw": "تیرانی", "corrected": "ایرانی", "span": [56,62] }, { "raw": "کور", "corrected": "کشور", "span": [84,87] }]</pre>	<p>پس از سال‌ها تلاش رازی موفق به کشف الكل شد. این دانشمند تیرانی باعث افتخار در تاریخ کوراست.</p>
<pre>[{ "raw": "فیریک", "corrected": "فیزیک", "span": [44, 49] }, { "raw": "ا بل", "corrected": "ق ا بل", "span": [61, 64] }, { "raw": "توجیح", "corrected": "توجیه", "span": [65, 70] }, { "raw": "رجو", "corrected": "رجوع", "span": [115, 118] }]</pre>	<p>بسیاری از مباحث علوم غیرطبیعی با استفاده از فیریک دنیای مادی ابل توجیح نیست و برای یادگیری باید به فلسفه‌های خاصی رجو کرد.</p>

همچنین در اینجا چند ایده آورده شده که می‌توانید از آن‌ها استفاده کنید.

- با توجه به خلاقیت خودتان می‌توانید دو روش خود را با هم ترکیب کنید و در دقت یا سرعت کار بهبود ایجاد کنید.
- همانطور که در مثال‌ها آورده شده در یک جمله ممکن است چند غلط املائی وجود داشته باشد. از این رو می‌توانید چه حالت ترتیبی یعنی تصحیح یک یک کلمات و حالت ترکیبی یعنی اصلاح یکباره تمام کلمات را امتحان کنید. واضحاً حالت ترکیبی سریعتر است اما امکان دارد حالت ترتیبی دقیقتر عمل کند

نمره امتیازی: مدل خود را به گونه ای آموزش دهید ک بتواند علائم نگارشی را نیز تا حدی تصحیح کند.

تبدیل زبان محاوره‌ای به رسمی

تبدیل زبان محاوره‌ای به رسمی و برعکس، جزو مسائل کاربردی در حوزه‌ی تولید زبان طبیعی^{۲۰} است. یکی از مراحل که می‌تواند به تحقق این هدف کمک کند، تغییر ترتیب کلمات متناسب با الگوی زبان (برای مثال تغییر جمله‌ی «رفتم دانشگاه.» در زبان محاوره‌ای به «دانشگاه رفتم.» در زبان رسمی و برعکس) و همچنین اصلاح فرآیندهای واجی (برای مثال تغییر کلمه‌ی «خونه» به «خانه» در تبدیل عامیانه به رسمی و برعکس) است. در این تمرین قصد داریم با کمک آنچه تا کنون در درس آموخته‌ایم، گامی در راستای حل این مسئله برداریم.

در بخش اول این تمرین، تغییر ترتیب کلمات زبان عامیانه به زبان رسمی مورد بحث بوده و حالت برعکس (رسمی به عامیانه) دارای نمره‌ی امتیازی است. لذا ورودی این مسئله کلمات یک جمله‌اند که لزوماً ترتیب استاندارد ندارند و خروجی آن پیشنهاد بهترین حالت جایگشت این کلمات در زبان رسمی است. برای این کار نیاز است دو مدل زبانی n-gram و مبتنی بر مبدل را بررسی و مقایسه کنید. در انتخاب شیوه‌ی حل مسئله مختارید اما یک پیشنهاد می‌تواند کمک گرفتن از اجزای سخن^{۲۱} و آموزش یک مدل زبانی بر روی داده‌ی برچسب‌خورده باشد. در خصوص دادگان زبان رسمی، می‌توانید از دیتاست اخبار و یا هر داده‌ی واجد شرایط دیگری استفاده کنید و در صورتی که به داده‌ی زبان محاوره‌ای نیاز داشتید، از دیتاست LSCP بهره بگیرید. ملاک ارزیابی مدل را معیاری مشابه با امتیاز BLEU در نظر بگیرید که در آن نسبت تعداد n-gramهایی که در خروجی به صورت صحیح آمده‌اند به تعداد کل n-gramهای موجود محاسبه می‌شود.

در بخش دوم نیاز است یک ارزیابی ذاتی^{۲۲} انجام داده و به کمک بردارهای جاسازی کلمات، ارتباط میان فرم کلمه در زبان رسمی و زبان عامیانه را بررسی کنید؛ در واقع هدف آنست دریابیم آیا با داشتن فرم کلمه در زبان عامیانه (برای مثال «خونه») می‌توان به فرم کلمه در زبان رسمی (که معادل «خانه» است) رسید؟ برای این کار می‌توانید از بردارهای جاسازی کلمات موجود نظیر `fasttext` استفاده کنید.

نهایتاً لازم است دو بخش یاد شده را به شیوه‌ای مناسب ترکیب کنید؛ به گونه‌ای که هم ترتیب کلمات به صورت خواسته‌شده درآید و هم تا حد ممکن فرآیندهای واجی اصلاح شده باشد.

²⁰Natural Language Generation (NLG)

²¹Part-of-speech (POS)

²²Intrinsic Evaluation

بررسی و اصلاح بایاس‌های مختلف در مدل‌های زبانی

بایاس در یادگیری ماشین به معنای جهت‌دار بودن تشخیص مدل یادگرفته شده است به عنوان مثال یک مدل برای تشخیص رای دادگاه براساس شواهد آموزش می‌دهیم و این مدل ممکن است براساس توزیع خاص دادگان آموزش نسبت به یک نژاد دارای قضاوت‌های خاصی باشد در این صورت مدل دارای بایاس هست.

این تمرین دارای دو بخش زیر است:

۱ - بررسی بایاس در یک مدل زبانی

۲ - اصلاح بایاس در یک مدل زبانی

همچنین بایاس‌ها می‌توانند انواع مختلفی داشته باشند که در این تمرین دو نوع بایاس براساس نژاد و بایاس براساس جنسیت مد نظر هست.

برای این تمرین باید یک مدل زبانی را برای بررسی انتخاب کنید که می‌توانید از یک مدل زبانی دلخواه مانند برت استفاده کنید که هم در زبان فارسی و هم در زبان انگلیسی موجود هست.

همچنین می‌توانید یک زبان دلخواه برای بررسی‌های خود انتخاب کنید و برای راحتی کار خود از زبان‌هایی استفاده کنید که دارای مدل زبانی از پیش آموزش داده شده باشند و بهتر است سراغ زبان‌های ایرانی یا زبان‌هایی که بایاس در آن‌ها کمتر بررسی شده است بروید یا حتی می‌توانید از مدل‌های چند زبانه استفاده کنید و بایاس را همزمان روی چند زبان بررسی کنید. (بررسی بایاس در مدل‌های چند زبانه دارای نمره امتیاز می‌باشد)

در بخش اول باید در مدل انتخاب شده دو نوع بایاس گفته شده را بررسی کنید و ببینید که از هر نوع بایاس در این مدل چه میزان وجود دارد و امتیازی برای شدت بایاس موجود طراحی کنید به عنوان مثال می‌توانید جملاتی در خصوص شغل افراد طراحی کنید و ببینید که مدل شغل‌های مختلف را با چه امتیازی به جنسیت یا نژادهای مختلف نسبت می‌دهد و مشابه شغل برای موضوعات مختلف این کار را انجام دهید و به شکل میانگین یک امتیاز برای هرکدام از این دو نوع بایاس در یک مدل زبانی طراحی کنید.

در بخش دوم باید یک روش ارائه دهید که شدت این بایاس‌ها را کم کند. روش شما می‌تواند در رابطه با اصلاح توزیع خود دادگان آموزش یا اصلاح خود مدل برای مقاومت در برابر بایاس یا به شکل یک مرحله بعد از پیش‌بینی مدل باشد و سپس امتیاز طراحی شده خود را دوباره اندازه‌گیری کنید.

تشخیص عبارات و کلمات هم معنا و متضاد

در این تمرین می خواهیم کلمات مترادف و متضاد را تشخیص دهیم. منظور از مترادف در این تمرین، کلماتی است که به جای هم به کار برده می شوند و استفاده از آن ها جای هم هیچ تفاوتی در معنا به وجود نمی آورد؛ کلماتی که از نظر ظاهری و یا جاسازی^{۲۳} مشابه هم هستند. همچنین کلماتی که از نظر روابط معنایی با هم در یک دسته قرار می گیرند را می توانید به عنوان مترادف در نظر بگیرید؛ مثل خودرو و وسیله نقلیه. ارتباط واژگانی مختلف را می توانید برای کلمات مترادف در نظر بگیرید. همچنین ممکن است کلمات به جای هم به کار نروند و کنار هم بیایند؛ مثل ایالات متحده آمریکا که ایالات متحده را می توان مترادف آمریکا در نظر گرفت. مجموعه داده مورد استفاده در این تمرین مجموعه داده ویکی پدیا فارسی است که در این [این لینک](#) قرار دارد. همچنین دادگانی از کلمات مترادف در [این لینک](#) موجود است که برای آموزش مدل تان می توانید از آن استفاده نمایید. این تمرین شامل مراحل زیر است:

- برای این کار ابتدا صفت ها، اسم ها و فعل ها را بیابید و تعدادشان را در مجموعه داده گزارش کنید و یک جاسازی با مدل های زبانی به کار ببرید که بتواند کلمات مشابه را بیابد. برای ارزیابی مدل تان از معیارهای *Precision, Recall, F1* برای فعل ها و اسم ها و صفت های مجموعه داده استفاده نمایید. برای مقایسه روش پیشنهادی تان از روش های پایه دیگر مانند *Glove* و *Word2vec* استفاده نمایید و میزان بهبودتان را در گزارش ذکر نمایید.
- در این قسمت با استفاده از دادگان مترادفی که در تمرین داده شده و روابط معنایی موجود در فارسی نت یک مجموعه داده ای از لغات مترادف مورد نظر تمرین بسازید.
- کلمات مترادف را بر روی کلمات کلیدی استخراج شده از مجموعه داده ویکی پدیا را بیابید.
- **امتیازی** روش پیشنهادی تان را برای پیدا کردن کلمات متضاد نیز آزمایش کنید. اگر بتوانید با تغییراتی در مدل پیشنهادی تان هم کلمات مترادف و هم متضاد تشخیص داده شود.
- **امتیازی** همچنین برای کلمات دخیل در فارسی مدل تان را آزمایش کنید. یعنی برای کلمات انگلیسی که در فارسی استفاده می شوند کلماتی با معنی مشابه را پیدا کنید. مثلا برای کلمه هلیکوپتر: بالگرد، اسانس: عطرمایه ، اتوماتیک: خودکار و انیمیشن: پویانمایی، مدل واژه مناسب فارسی را بتواند بیابد.

²³Embedding

یکی از انواع روابط واژگانی که در درس به آن پرداخته شد، چند معنایی بودن^{۲۴} است. به این معنا که یک لغت می‌تواند معانی گوناگونی داشته باشد. به عنوان مثال برای زبان فارسی، واژه شیر می‌تواند معنای شیر جنگل، شیر آب و شیر لبنیات را داشته باشد. واژه Bank در انگلیسی نیز از این دست واژگان است که می‌تواند به معنای موسسه بانکی و یا حاشیه شیب‌دار رودخانه باشد. همان‌طور که انتظار می‌رود، این لغات می‌توانند باعث بروز ابهام شوند و به کار رفتن هر کدام از وجوه معنایی این واژگان^{۲۵} در سیاق درست حائز اهمیت است. هدف از این تسک، ارائه‌ی یک روش ارزیابی جدید برای سنجش عملکرد مدل‌های زبانی از پیش آموزش داده شده در یادگیری وجوه معنایی مختلف واژگان است تا ببینیم که مدل‌های زبانی تا چه حد می‌توانند یک واژه را بر اساس وجوه معنایی گوناگونی که دارد، در سیاق‌های مختلف پوشش داده و درست جایگذاری کنند. برای درک بهتر مساله فرض کنید دو مدل زبانی از پیش آموزش داده داریم. می‌خواهیم عملکرد آن‌ها را در درک وجوه معنایی واژه Bank بسنجیم. دو جمله زیر را در نظر بگیرید.

- bank1 : ...a **bank** can hold the investments in a custodial account ...
- bank2 : ...as agriculture burgeons on the east **bank**, the river ...

برای این منظور اگر دو جمله بالا را به عنوان ورودی داده و واژه bank را mask کنیم، مدل زبانی‌ای عملکرد بهتری در درک وجه معنایی بانک داشته که این واژه را در رتبه‌ی بهتری نسبت به واژگان دیگر برای هر دو جمله در سیاق‌های مختلف پیشنهاد دهد. برای سنجش می‌توانید از معیار Mean Reciprocal Rank استفاده کنید. بنابراین برای این تسک چند مدل زبانی از پیش آموزش داده شده روی زبان‌های فارسی و انگلیسی را از hugging face به دلخواه و بر اساس اهمیت‌شان انتخاب کرده و مجموعه داده‌ای جهت ارزیابی تهیه فرمایید. به عنوان نمونه از مدل‌های زبانی ROBERTa ، ALBERT و BERT برای زبان انگلیسی و از مدل‌های زبانی ParsBERT و ParsBigBird برای زبان فارسی، همچنین برای ایجاد مجموعه داده از Wordnet و Farsnet می‌توانید استفاده نمایید. نمونه‌ای از مجموعه داده‌گان موجود برای زبان انگلیسی، SemCor و برای زبان فارسی SBU-WSD-Corpus هستند. نهایتاً، باید روشی برای ارزیابی ارائه کرده و عملکرد مدل‌های زبانی موجود را هم بر روی داده‌های زبان فارسی و هم زبان انگلیسی سنجیده و گزارش کنید.

²⁴Polysemy

²⁵Word Sense