به نام خدا

تمرین سری پنجم درس پردازش زبان‌های طبیعی
هدف از تمرینات این سری آشنایی شما عزیزان با مباحث برچسب زدن توالی(Sequence labelling) و ترجمه ماشینی(Machine Translation) می باشد. شما باید از بین دسته های زیر یکی را به دلخواه انتخاب کرده و تا تاریخ 23 دی ماه در این فرم وارد کنید.
برای مشخص کردن ترک انتخابی کافیست تنها یکی از اعضای گروه فرم را تکمیل کند و در صورتی که چند نفر از اعضای یک گروه فرم را تکمیل کنند جریمه خواهند شد و همچنین در پایان برای بارگذاری فایل ها و گزارش تمری خود نیز فردی که فرم را تکمیل کرده است، برگذاری کند.
دسته اول را می‌توانید با استفاده از هر کدام از روش های labeling یا translation حل کنید.
برای انجام این تمرین شما تا روز 8 بهمن ماه فرصت دارید. در دسته اول و سوم به شما یک فایل برای آموزش مدل ها داده شده که می توانید از آن برای آموزش و ارزیابی(validation) استفاده کنید و یک فایل برای آزمون(test) که بدون برچسب(جمله) است و شما باید برچسب ها(جملات) را پیش بینی کنید.
برای پیاده سازی مدل ها استفاده از پکیج ها و کتابخانه ها و مدل های از پیش آموزش دیده و همینطور استفاده از یک مجموعه داده مستقل از مجموعه دادگان معرفی شده برای بهسازی (fine-tune) کردن مدل ها مجاز است.

دسته اول

نویسه گردانی(transliteration)

برای این قسمت باید بتوانید نوشتارهای(خط) مختلف زبان ها را به یکدیگر تبدیل کنید. مثلا تبدیل فینگلیش(finglish) به فارسی و برعکس و یا تبدیل نوشتار تاجیکی به فارسی، نوشتاری ترکی استانبولی به ترکی آذری و یا تبدیل بین نوشتارهای زبان کوردی. از دیگر زبان ها مانند اشکال مختلف زبان لاتین نیز میتوانید استفاده کنید. برای این دسته به جز تبدیل فینگلیش به فارسی مجموعه داده ای در دسترس نمی باشد، لذا به گروه هایی که این دسته را انتخاب کرده و مجموعه داده جمع آوری می کنند نمره اضافه تعلق می گیرد.
برای ارزیابی می توانید از معیارهای ارزیابی مانند BLEU و NIST و Accuracy استفاده کنید.
برای تبدیل فینگلیش(finglish) به فارسی و برعکس مجموعه دادگان مورد نیاز آن تهیه شده است. برای این تمرین شما باید مدلی را طراحی کنید که بتواند متون فینگلیش را به فارسی تبدیل کند و برعکس. مجموعه داده های مورد نیاز در لینک وجود دارد و شما می توانید از آن ها استفاده کنید.
در ادامه می توانید مثال هایی را از این تبدیل مشاهده کنید:

az ham joda shodan kheili sakhteh
فارسی: از هم جدا شدن خیلی سخته

baleh motmaen bashid hastam
فارسی: بله ، مطمئن باشید هستم

ma mard hayeh awli naboudim
فارسی: ما مردهای عالی نبودیم

دسته دوم

اعراب گذاری متون عربی

در این تمرین شما باید بتوانید متون عربی بدون اعراب را اعراب گذاری کنید. برای آموزش نیز مجموعه داده های قرآن مجید، نهج البلاغه و صحیفه سجادیه در اختیار شما عزیزان قرار می گیرد که لینک آنها در ادامه آمده است. مانند دسته قبل برای ارزیابی می توانید از معیارهای ارزیابی مانند BLEU و NIST و Accuracy استفاده کنید.

در زیر چند نمونه را می توانید مشاهده کنید:

بدون اعراب: بسم الله الرحمن الرحیم

با اعراب: بِسْمِ اللَّـهِ الرَّحْمَـٰنِ الرَّحِيمِ

بدون اعراب: هل یستوی الذین یعلمون و الذین لا یعلمون

با اعراب: هَلْ يَسْتَوِي الَّذِينَ يَعْلَمُونَ وَالَّذِينَ لَا يَعْلَمُونَ

مجموعه داده قرآن مجید | مجموعه داده نهج البلاغه | مجموعه داده صحیفه سجادیه در لینک وجود دارد و شما می توانید از آن ها استفاده کنید.

دسته سوم

مدل های چند زبانه ی تشخیص موجودیت های نامدار(Multilingual NER)

در این دسته شما باید بتوانید اجزای متن که در زبان فارسی و انگلیسی هستند را بر اساس کلاس های شش گانه زیر دسته بندی کنید. برای این تمرین دو مجموعه داده یکی به زبان فارسی و دیگری به زبان انگلیسی در اختیار شما قرار می گیرد و شما باید مدلی طراحی کنید که برای هر دو دسته پاسخگو باشد.

راهنمایی: می توانید از مدل های چند زبانه مثل Roberta استفاده کنید. همچنین می توانید از مجموعه داده هایی مثل آرمان برای بهینه سازی مدل استفاده کنید. برای ارزیابی می توانید از معیارهای ارزیابی دسته بندی استفاده کنید(Precision, Recall, Accuracy, F1-score)

1. PER : Person
2. LOC : Location
3. GRP : Group
4. CORP : Corporation
5. PROD : Product
6. CW: Creative Work

لینک مجموعه داده مورد نیاز در ادامه آورده شده است. مثال هایی از این مورد:

His playlist includes sonny sharrock, gza, country teasers and the notorious b.i.g

His	O
playlist	O
includes	O
sonny	B-PER
sharrock	I-PER
,	O
gza	B-PER
,	O
country	B-GRP
teasers	I-GRP
and	O
The	B-PER
notorious	I-PER
b.i.g	I-PER

دسته چهارم

ترجمه ماشینی

در این دسته شما باید به دلخواه یک زبان غیر فارسی را به یکی از زبان‌های ایرانی ترجمه کنید. از این لینک می توانید به عنوان مجموعه داده استفاده کنید. شما باید مجموعه داده ای که استفاده می کنید را مانند سایر دسته ها به سه دسته آموزش، ارزیابی و آزمون تقسیم کنید(دسته آزمون هنگام آموزش نباید توسط مدل دیده شود). برای ارزیابی می توانید از معیارهای ارزیابی مانند BLEU و NIST و Accuracy استفاده کنید.