به نام خدا

تمرین سری سوم درس پردازش زبان طبیعی

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف

پاییز ۱۴۰۰

تمرین سری سوم: مدل‌های زبانی

آدرس سایت درس

http://language.ml/courses/

درباره‌ی تمرین

هدف از این کار با مدل‌های زبانی[1] است. تمرین پیش‌رو شامل ۴ بخش است و تیم‌ها به صورت گروه‌های حداکثر ۳ نفره باید به انجام این تمرین بپردازند. در این تمرین هر تیم فقط یکی از چهار بخش را پیاده‌سازی می‌کند. مهلت انتخاب موضوع تمرین سری سوم حداکثر تا دوشنبه ساعت ۱۱:۵۹ - ۲۹ آذر ماه است.

دسته اول - تغییر الگوی گفتاری غیر رسمی و شعری به فرم رسمی نوشتاری

در این بخش هدف آن است که یک جمله‌ی بهم ریخته از لحاظ چینش کلمات به عنوان ورودی داده شود و شما باید بتوانید کلمات را به درستی در کنار یکدیگر بچینید و به عنوان خروجی نمایش دهید. نمونه‌هایی از ورودی و خروجی‌های مورد انتظار عبارت‌اند از:

- ورودی: کی زنگ می‌زنی به من؟
  - خروجی: کی به من زنگ می‌زنی؟
- ورودی: برگزار کردیم جلسه را.
  - خروجی: جلسه را برگزار کردیم.
- ورودی: حجره خورشید تویی خانه ناهید تویی / روضهٔ امید تویی راه ده ای یار مرا
  - خروجی: حجره خورشید تویی خانه ناهید تویی / روضهٔ امید تویی ای یار مرا راه ده
- ورودی: زند موجی بر آن کشتی که تخته تخته بشکافد / که هر تخته فروریزد ز گردش‌های گوناگون.
  - خروجی: موجی بر آن کشتی زند که تخته تخته بشکافد / که هر تخته از گردش‌های گوناگون فروریزد.
- ورودی: دارم حرف می‌زنم من.
  - خروجی: من دارم حرف می‌زنم.

راهنمایی: شما برای حل این مساله از هر روش خلاقانه‌ای با استفاده از مدلهای زبانی می‌توانید بهره ببرید. یکی از روشهای پیاده‌سازیزی این بخش یادگیری یک مدل زبانی (از مدلهای ساده مثل n-gram تا استفاده از مدلهای شبکه‌های عصبی) بر روی POS جملات متون رسمی است. برای این مورد دیتای اخبار فارسی در حوزه‌های مختلف برای شما فراهم شده است. شما می‌توانید بخش محدودی از این داده را به عنوان نمونه انتخاب کنید. سپس با استفاده از POS-tagger بر روی کلمات جمله ورودی (جایگزینی کلمات با محتمل‌ترین POS آنها) جایگشتی که بالاترین امتیاز را از نظر language model دارد در نظر بگیرید.

[1] Language models

دسته دوم - کامل کردن کلمه‌ی جاری در جمله برای زبان فارسی (و یا یک زبان ایرانی)

در این بخش، باید یک مدل زبانی برای فهم بافت جمله یاد گرفته شود. سپس با استفاده از مدل زبانی یاد گرفته شده، محتمل‌ترین کاراکترها برای تکمیل کلمه‌ی جاری مورد استفاده قرار بگیرد. مدل زبانی مورد استفاده می‌تواند مبتنی بر کاراکتر و یا کلمه باشد. همچنین می تواند ترکیبی از مدل زبانی مبتنی بر کاراکتر و یا کلمه را استفاده کنید. برای نمونه ممکن است شما بخواهید ابتدا با آموزش مدل زبانی مورد نظر، دامنه‌ی کلمات را کم کرده و سپس از میان کلمات برگزیده شده، محتمل‌ترین کلمه را با توجه به مدل زبانی مبتنی بر کاراکتر انتخاب کنید.

- ورودی: دیروز که داشتم به دانشگاه می‌رفتم متوجه شدم که کل.
  - خروجی: دیروز که داشتم به دانشگاه می‌رفتم متوجه شدم که کلاس
- ورودی: دیوان شمس از مو
  - خروجی: دیوان شمس از مولانا
- ورودی: به نام خ.
  - خروجی: به نام خدا

نمونه‌های بالا صرفا جهت درک بهتر است و دامنه‌ی اصلی مورد پوشش شما، متن‌های خبری خواهد بود که در اختیارتان قرار می‌گیرد.

دسته سوم –ساخت یک سامانه بازیابی اطلاعات بر اساس مدل زبانی

بدین منظور بردارهای تعبیه به کمک مدل زبانی برای سندها (و یا جمله‌ها) و همچنین پرس‌وجوها ساخته می‌شوند و با توجه به شباهت پرس‌وجوی ورودی کاربر با سندها، مرتبط‌ترین‌شان بازگردانده می‌شوند. معیار ارزیابی P@K خواهد بود. ورودی در این بخش پرس‌وجو کاربر است و خروجی باید شامل لیستی از سندهای مرتبط (به ترتیب مرتبط‌تر بودن) باشد. برای یافتن مرتبطترین سندها حداقل از ۳ روش بازنمایی سندها استفاده نمایید و برای داده ارزیابی جدول ارزیابی تولید نمایید.

۱) tf-idf (word-level/bpe) - (ngram=۱،۲ کلمات)

۲) tf-idf weighted summation of embeddings (different window sizes)

۳) برت فارسی

۴) روش خلاقانه

دسته چهارم – یادگیری و ارزیابی مدل زبانی فارسی (– یا یک زبان دلخواه) (در سطح: کاراکتر، bpe، کلمات) و مدلهای مختلف هموار شده n-gram و شبکه‌های عصبی (LSTM، یا ترنسفرمرها)

در این بخش، دستکم سه مدل زبانی مختلف بر روی یک مجموعه‌ی داده‌ی مشخص در نظر گرفته می‌شود و آموزش داده می‌شود و هر سه مدل با یکدیگر مقایسه می‌شود. این مقایسه می‌تواند از منظرهای مختلف صورت بگیرد ولی گزارش کردن مقدار perplexity برای هر روش ضروری است. همچنین یادگیری مدل زبانی برای هر سه روش، در سه حالت مختلف byte pair encoding، کاراکتر و کلمه باید انجام شود.

نکات پایانی

- در مورد میزان به کارگیری داده‌ها برای آموزش مدل در هر یک از چهار بخش، این مورد بر عهده‌ی گروه‌ها است. از آن‌جایی که در این تمرین فرض شده است که شما نهایتا با سیستمی با مشخصات گوگل گولب کار خواهید کرد، نیاز نیست که با همه‌ی داده‌ها کار کنید. بنابراین در آغاز کار باید از داده‌های موجود، یک نمونه‌گیری (به روش دلخواه خودتان) انجام شود و بر روی آن بخش از داده‌ها کار انجام بشود.
- داده‌ها را می‌توانید از طریق این نشانی[2] دریافت کنید
- در هر بخش، در صورتی که بیش از خواسته‌ی سوال کار انجام بگیرد، می‌تواند منجر به گرفتن نمره‌ی بیش از تمرین بشود (نمره‌ی اضافی). البته این مورد قبلا باید هماهنگ شده باشد و هر ویژگی اضافی لزوما نمره اضافه در بر نخواهد داشت.
- برای راحتی کار گروه‌ها در آزمایش مدل‌های ساخته شده، یک وب‌سرویس مبتنی بر فلسک[3] تا پایان روز دوشنبه در اختیار شما قرار می‌گیرد تا بتوانید کار خود را در صورت نیاز، با این وب‌سرویس آزمایش کنید.
- زبان مجاز برای پیاده‌سازی این تمرین، پایتون است.
- همه‌ی اعضای گروه باید به کدهای نوشته شده، خروجی‌ها و همچنین نحوه‌ی کارکرد کد آگاهی کافی داشته باشند.
- انجام دادن تمرین به صورت انفرادی هیچ مزیتی بر انجام تمرین به صورت گروهی ندارد.
- برای هر تمرین یک گزارش شامل مستندات کد و توضیحات و خروجی‌ها آماده کنید. همچنین کدها خوانا باشند و کامنت‌گذاری نیز انجام شود.

[2] https://github.com/language-ml/2-LM-embedding-projects
[3] Flask