بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

پردازش زبان طبیعی

نیم‌سال دوم ۰۲-۰۱

مدرس: احسان‌الدین عسگری

تمرین اول | کاوش متن | مهلت ارسال: ۱۷ اسفند

- 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 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 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.

کاوش متن

در این تمرین انفرادی قرار است مراحل مختلف پیش‌پردازش را روی متن انتخابی خود اعمال کرده و در پایان یک تسک دلخواه مانند استخراج عبارات کلیدی روی متن‌تان تعریف کرده و آن را حل کنید. هدف این تمرین آن است که با چالش‌های اولیه پردازش یک متن آشنا شوید.
ترجیح شدید در این تمرین این است که متن مورد نظرتان را خودتان به دست آورید. با این کار هم با روش‌های به دست آوردن متن آشنا می‌شوید. هم متن مورد نظر با احتمال بالایی دارای مسائل پیش پردازشی خواهد بود.
اگر مایل به این کار نیستید می‌توانید از متون از پیش آماده شده استفاده کنید. اما در این صورت تسکی که روی متن انجام می‌دهید باید فراتر از استخراج عبارات کلیدی و نمونه تسک‌های مطرح در کلاس باشد وگرنه نمره تمرین را از دست می‌دهید!
نمره شما در این تمرین بر اساس میزان زحمتی که کشیده‌اید داده می‌شود. در واقع صرف اجرا کردن نوت‌بوک نمونه روی یک متن از پیش آماده شده برای نمره‌ی این تمرین کافی نیست و باید مساله شما یا چالش استخراج متن و تمیزسازی

و یا چالش تسک داشته باشد.
برای مشاهده نمونه متن و نمونه نوت‌بوک‌های انجام شده به این لینک مراجعه کنید.
در این تمرین به نکات زیر دقت کنید:

۱. تلاش کنید که تمام گزارش خود را با رعایت راست‌چین و چپ‌چین بودن متن در داخل نوتبوک قرار دهید. این کار را می‌توانید با استفاده از HTML در سلول‌های نوتبوک انجام دهید. اگر این کار برایتان مقدور نیست می‌توانید گزارش را به شکل یک pdf جداگانه در کنار پروژه آپلود کنید.

۲. در گزارش کارتان (یا همان نوت‌بوک پروژه) منبعی که از آن داده را به دست آورده‌اید ذکر کنید و در مورد داده‌ی خود توضیحات اولیه‌ای بدهید.

۳. اگر داده را با استفاده از Crawl کردن به دست آورده‌اید در کنار پروژه اسکریپتی که با آن این کار را انجام داده‌اید را قرار دهید.

۴. کد اصلی شما که در واقع فایل نوت‌بوک است باید بدون مشکل بتواند در کامپیوتری دیگر اجرا شود. برای همین در صورت استفاده از یک پکیج خاص، در قالب زیر دستور نصب آن در اول نوتبوک قرار دهید.

```
!pip install <package_name>
```

۵. پیش‌پردازشی که قرار است روی متن اعمال کنید بایستی برای متن و تسک مورد نظر شخصی‌سازی شود. برای مثال ممکن است بخواهید stopword ، pos-tagger ، lemmatizer و دیگر مواردی از این دست را به نحوی استفاده یا پیاده‌سازی کنید که متناسب با متن و تسک شما باشند.
از کاربرد انجام این کار می‌توان به پردازش متون کهن نام برد. چون با این که ابزارهای Hazm و DadmaTools برای زبان فارسی وجود دارند اما برای متون کهن عملکرد مناسبی ندارند.
برای خلاقیت و ایده‌های تازه در پیش‌پردازش متن، نمره‌ی امتیازی در نظر گرفته شده است.

۶. حداقل یکی از انواع پیش پردازش‌هایی که استفاده کرده‌اید را تست کنید. برای مثال یک بار بدون آن پیش‌پردازش و یک بار همراه با آن، تسک را انجام دهید و نتایج را مقایسه کنید. این مقایسه می‌تواند به صورت کمی یا کیفی باشد. در نهایت هدف این است که برای تصمیم‌هایی که در پیش‌پردازش می‌گیرید، دلیل مستدل و داده‌محوری ارائه کنید.

۷. فایل متنی ورودیِ پردازش نشده‌ی شما باید همراه با پروژه آپلود شود. حجم فایل حدودا باید بین ۱ تا ۵ مگابایت باشد. در صورت کم حجم بودن فایل ورودی ممکن است بخشی از نمره را از دست بدهید!

۸. بایستی (مشابه با سبک نوت‌بوک‌هایی که در اختیارتان قرار داده شده است) ابتدا مشخصات کلی متن (شامل آمار تعداد جملات و کلمات و ...) را ذکر کنید و در هر بخش توضیحات کار را در نوت‌بوک بنویسید.

۹. برای تسکی که انتخاب می‌کنید، می‌توانید از مثال‌های زیر ایده بگیرید:

- محاسبه گسترده و با جزییات شاخص‌هایی در مورد متون شعرای مختلف با هدف مقایسه و مشاهده تفاوت‌ها و شباهت‌ها
- مقایسه متون نوشته شده در قرن‌های مختلف با هدف استخراج عبارات نو و یا محاسبه دیگر شاخص‌ها
- مقایسه سخن‌رانی‌های چند رئیس جمهور و استخراج عبارات کلیدی
- تولید پایپ‌لاین‌های پیش‌پردازشی برای زبان‌ها یا گویش‌های ایرانی
- استخراج عبارات (چند کلمه‌ای) کلیدی متن یا پیدا کردن عنوان، برای متن

۱۰. سعی کنید در گزارشی که می‌نویسید موارد زیر وجود داشته باشد:

- توضیح درباره‌ی کلیت کد در حدی که بدون خواندن جزئیات کد، بتوان آن را متوجه شد.

- خلاصه‌ای از نتایج به دست آمده و تحلیل آن. اگر چیزی را امتحان کردید و به نتیجه‌ی خوبی نرسید هم آن را در نوتبوک و گزارش بیاورید تا روند رسیدن به خروجی نهایی مشاهده شود.
- دلیل تصمیم‌هایی که گرفتید. برای مثال اگر تصمیم گرفتید از نوع خاصی از stemming استفاده کنید، چرایی‌اش را در گزارش بیاورید.