« تمرین درسی »

درس	عنوان	مهلت	نمره	نوع
مباحث ویژه (یادگیری ماشین)	**تمرین ۱**	**۱۴ آبان ۱۴۰۱ - ۲۳:۵۹**	**۴**	**اختیاری**

I. **هدف آموزشی تمرین:** در این تمرین قصد داریم با نحوه پیش‌پردازش متون انگلیسی و فارسی آشنا شویم. بدین منظور چهار فایل متنی فراهم شده است که ضمیمه پیام است. ابتدا فایل‌های متنی را باز کرده، متغیرهای لازم را تعریف کنید و مراحل تمرین را گام به گام پیاده‌سازی کنید. برای هر گام علاوه بر پیاده‌سازی بر روی دیتاست ارائه شده، یک مثال دیگر نیز بزنید و در گزارش کار، گزارش کنید.

II. **نحوه ارسال تمرین:** پیاده‌سازی انجام شده را در قالب یک فایل Jupyter notebook یا پایتون (.py) به همراه یک گزارش کار در قالب Word و PDF. محتوای نهایی تمرین به شکل فشرده در بخش تکالیف درسی در سامانه سمیاد ارسال شود.

III. **شرح و محتوای تمرین:**

۱. **سازمان‌دهی فضاهای خالی**[1]**:** فضاهای خالی موجود در متن می‌تواند فاصله[2]، تب[3] و یا خط جدید[4] باشند. در این مرحله باید فضاهای خالی اضافی حذف شود و بین هر دو کلمه فقط یک فاصله باشد. برای این کار از متد WhitespaceTokenizer موجود در کتابخانه NLTK استفاده کنید. ابتدا با استفاده از این متد داده‌های متنی را تجزیه کنید. سپس ایراد متد را گزارش کرده و در انتها خروجی حاصل را به یک متن تبدیل کنید.

۲. **یکپارچه‌سازی حروف:** برای متون انگلیسی باید حروف تمامی کلمات موجود در متن کوچک باشد. متد لازم برای این مرحله را پیدا کرده و آن را روی متون انگلیسی استخراج شده از مرحله قبل اعمال کنید

۳. **استخراج جملات و توکن‌ها:** در این مرحله باید جملات و توکن‌های موجود در متون پردازش شده را استخراج کنید. ابتدا جملات را با استفاده از PunktSentenceTokenizer استخراج کنید. آیا این روش برای متن کوتاه انگلیسی و متن کوتاه فارسی جملات را به درستی تفکیک کرده است؟ برای رفع این مشکل چه متدی را پیشنهاد می‌کنید؟ سپس توکن‌ها را با استفاده از متد TreebankWordTokenizer استخراج کنید. تعداد کل جملات، تایپ‌ها و توکن‌های موجود در هر متن را نیز گزارش کنید.

۴. **حذف اعلام نگارشی:** در این مرحله قصد داریم کاراکترهای الفبایی و غیر الفبایی را جداسازی کنیم؛ به عبارت دیگر اعلام نگارشی، اعداد و کارکترهای خاص باید حذف شوند. این کار توسط متدهای زیادی انجام می‌شود. در این مرحله نحوه‌ی عملکرد متدهای RegexpTokenizer را بررسی می‌کنیم. با استفاده از این متد توکن‌های استخراجی از مرحله قبل را پردازش کرده و خروجی را گزارش کنید.

۵. **مفهوم StopWord:** ابتدا در مورد مفهوم StopWordها تحقیق کنید و سپس برای متون انگلیسی StopWord را از توکن‌های استخراجی حذف کنید و تعداد توکن‌ها برای هر متن گزارش کنید.

۶. **Stemming:** در این مرحله قصد داریم ریشه توکن‌های یافت شده را استخراج کنیم. این کار با استفاده از متدهای LancasterStemmer و PorterStemmer برای توکن‌هایی با اندیس ۳، ۱۱، ۶۰ و ۶۸ برای متن Beanstalk و توکن با اندیس ۲ برای متن ShortSampleEnglish انجام دهید.

---

[1] White Space
[2] Space
[3] Tab
[4] New line

۷. **Lemmatization:** با استفاده از WordNetLemmatizer کلمات زیر را به حالت نگارشی اولیه آنها بازگردانید.

went	better	was	eaten	butterflies	fishing	signaling

آیا استفاده از متد lemmatize با ورودی‌های پیش‌فرض، برای همه این کلمات پاسخ درست را برمی‌گرداند؟ اگر پاسخ خیر است، پیشنهاد شما برای اینکه با این روش بتوان برای همه این کلمات نتایج صحیح گرفت چیست؟ (راهنمایی: به ورودی pos از متد lemmatize دقت کنید و آن را متناسب با کلمات تغییر دهید.)

- موفق باشـــید...