
 <p>دانشگاه صنعتی امیرکبیر</p>	<p>دوره آموزشی پردازش زبان طبیعی (NLP)</p> <p>تمرین سوم</p> <p>مهلت تحویل:</p> <p>۳۰ مردادماه ۱۴۰۰</p>	 <p>آکادمی همراه</p>
---	--	---

برای این تمرین دو فایل با عنوان train.csv به عنوان مجموعه داده آموزش و یک فایل با عنوان test.csv به عنوان مجموعه داده آزمون در اختیار شما قرار گرفته شده است. این مجموعه داده دارای دو ستون است. ستون article متن سند و ستون id شناسه سند است.

نکته ۱: برای دسترسی به مجموعه داده آموزش و تست به ترتیب از دستورات زیر در سرویس ابری گوگل کولب استفاده کنید.

```
!gdown --id 1rovazK48q7pHcEM271aX70Dr594NYQ77
!gdown --id 1ZCHuj6JtyOkb5ismn3qF3Rp2DRGJ1tRk
```

سوال ۱) با استفاده از مجموعه داده آموزش، مدل های بازنمایی زیر را آموزش دهید. (بردار هر متن ۳۰۰ بعد در نظر گرفته شود).

الف) آموزش بردار کلمات روی مجموعه داده با استفاده از word2vec مدل skip-gram.

ب) آموزش بردار اسناد روی مجموعه داده با استفاده از doc2vec مدل.

نکته ۲: برای پیاده سازی هر دو مدل word2vec و doc2vec می توانید از کتابخانه gensim استفاده کنید. بردار کلمات را برای کلمات بالای ۴ بار تکرار و با window ۱۰ به دست آورید.

نکته ۳: برای محاسبه بازنمایی گزینه (ب) کل سند را به عنوان ورودی به مدل doc2vec بدهید.

سوال ۲) برای اسناد موجود در پیکره آزمون، بردار بازنمایی اسناد را با استفاده از روش های زیر به دست آورید.

الف) استفاده از بردار کلمات آموزش دیده توسط word2vec مدل skip-gram روی مجموعه آموزش و محاسبه بردار جملات پیکره آزمون با استفاده از میانگین وزن دار بازنمایی کلمات سند با استفاده از TF-IDF هر یک از کلمات.

ب) محاسبه بردار اسناد براساس مدل doc2vec آموزش داده شده.

سوال ۳) برای هر یک از اسناد زیر (شناسه این اسناد در ادامه نوشته شده است) شبیه ترین متن به آن ها را از مجموعه آموزش بیابید. برای محاسبه شباهت از معیار شباهت کسینوسی استفاده کنید و برای محاسبه بردار هر یک از دو روش ذکر شده در سوال ۲ را استفاده نمایید. (برای محاسبه معیار شباهت کسینوسی پیشنهاد می شود از کتابخانه sklearn استفاده کنید).

۱. Doc443

۲. Doc428

۳. Doc635

نکته ۴: مثالی از خروجی این سوال در ادامه آورده شده است. فرض کنید شناسه سند اصلی Doc951 باشد. حرف A نشان دهنده این است که این شبیه‌ترین سند و میزان شباهت آن با استفاده از بازنمایی حالت (الف) سوال ۲ و B با استفاده از بازنمایی حالت (ب) سوال ۲ به دست آمده است.

Doc951: (Doc101, A), (0, 6529), (Doc854, B), (0, 5501)

سوال ۴) برای هر یک از کلمات زیر سه شبیه‌ترین کلمه را از میان تمام کلمات که در بازنمایی word2vec از داده آموزش استخراج کردید بیابید. برای این کار از تابع most_similar مدل word2vec ایجاد شده با کتابخانه gensim استفاده کنید.

۱. بهداشت

۲. استقلال

۳. رودخانه

نکته ۵: مثالی از خروجی این سوال در ادامه آورده شده است. فرض کنید کلمه اصلی کلمه 'دفاع' باشد.

دفاع: ('مقدس', 0, 6529), ('پشتیبانی', 0, 5501), ('ضد موشکی', 0, 5276)

سوال ۵) بردار هر کلمه سوال ۴ را به همراه ده شبیه‌ترین کلمات به هر یک از آن‌ها را در یک plot نمایش دهید. تحلیل خود را از این plot بنویسید. برای این کار ابتدا باید بردار هر کلمه را از فضای ۳۰۰ بعدی به فضای ۲ بعدی ببرید. در این مرحله از لینک موجود در پاورقی^۱ کمک بگیرید. برای استفاده از کتابخانه معرفی شده در پاورقی پارامتر init را برابر با مقدار 'pca' قرار دهید تا برای کاهش بعد از الگوریتم pca استفاده شود. سپس برای نمایش بردارهای دو بعدی کلمات مشابه از کتابخانه matplotlib استفاده کنید.

نکته ۶: پیشنهاد می‌شود مقادیر پارامترها برای کتابخانه پیشنهادی در لینک موجود در پاورقی به صورت زیر تنظیم شود.

```
TSNE(perplexity=10, n_components=2, init='pca', n_iter=3500, random_state=32)
```

نکته ۷: خروجی این سوال نمودار ذکر شده در سوال و تحلیل شما از نمودار خواهد بود.

¹ <https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html>