به نام خدا

# سامانه‌های یادگیری ماشین توزیع‌شده (پاییز ۱۴۰۴)

## تمرین کامپیوتری ۳

## مهلت ارسال: ۱۳ / ۱۰ / ۱۴۰۴

---


استاد درس: دکتر دوستی

دستیار طراح: علی خرم‌فر


---

## قوانین و ملاحظات

### نحوه ارسال تمرین:

- تمامی فایل‌ها باید در یک فایل فشرده با نام DMLS-CA3-StudentID ارسال شوند.
- کدهای مربوط به هر بخش را بر اساس جدول انتهای همین فایل ذخیره کرده و همراه گزارش ارسال کنید.
- تمامی کدهای ارسال‌شده باید امکان اجرای مجدد داشته باشند. اگر تنظیمات خاصی برای اجرا نیاز است، آن را ذکر کنید.
- کدهای ارسال‌شده باید توسط خودتان اجرا شده باشند و نتایج اجرا در فایل ارسالی مشخص باشد.

### رعایت اصول آکادمیک و صداقت علمی:

- این تمرین باید به‌صورت **فردی** انجام شود. هرگونه همکاری یا نوشتن تمرین به‌صورت گروهی ممنوع است.
- در صورت مشاهده تشابه در پاسخ، تمامی افراد درگیر نمره صفر دریافت خواهند کرد و موضوع به استاد گزارش خواهد شد.

### استفاده از ابزارهای هوش مصنوعی:

استفاده از ابزارهایی مانند ChatGPT، Gemini، Copilot و موارد مشابه مجاز است، اما تحت شرایط زیر:

- نحوه استفاده از این ابزارها را در گزارش خود توضیح دهید (ابزارهای استفاده‌شده، کاربردهای مشخص و موارد مرتبط).
- تمامی پرامپت‌ها و لینک‌های استفاده‌شده را در انتهای گزارش قرار دهید.
- عدم ارائه این اطلاعات به منزلهٔ سرقت علمی محسوب شده و منجر به نمره صفر خواهد شد.

### مهلت ارسال و جریمه تأخیر:

- امکان ارسال تمرین با تأخیر تا ۲ روز وجود دارد. برای هر روز تأخیر ۱۰ درصد جریمه اعمال می‌شود.
- تأخیر به‌صورت ساعتی محاسبه شده و پس از دو روز تأخیر، تمرین پذیرفته نخواهد شد.

### ارزیابی حضوری:

- ارزیابی تمرین به‌صورت حضوری انجام خواهد شد.
- محل ارزیابی: آزمایشگاه NLP، طبقه منفی یک، دانشکده مهندسی برق و کامپیوتر ساختمان شماره ۲.

## توضیحات تمرین

لطفاً پیش از شروع کار بر روی تمرین، به نکات زیر توجه فرمایید.

- برای دسترسی به UI ماشین Spark Master، به آدرس http://raspberrypi-dml0:8080 رفته و از نام کاربری admin و گذرواژه dmlsAdmin استفاده کنید.
- برای دسترسی به UI مربوط به HDFS به آدرس http://raspberrypi-dml0:9870 بروید.
- لطفاً فایل‌های دانشجویان دیگر را تغییر ندهید.
- برای دسترسی به HDFS در کد خود، می‌توانید از آدرس http:// raspberrypi-dml0:9000 استفاده کنید.
- برای سؤال‌های ۱ و ۲ می‌توانید از Colab یا کامپیوتر شخصیتان استفاده نمایید، ولی سؤال سوم باید روی کلاستر درس (بردهای رزبری پای) انجام شود.
- برای پیاده‌سازی از زبان پایتون و کتابخانه‌ی PySpark استفاده نمایید. برای استفاده از مدل‌ها یا توابع یادگیری ماشین از کتابخانه‌ی ml به جای mllib استفاده نمایید.
- قبل از شروع تمرین بهتر است ویدیوی آپلود شده در سامانه درس را مشاهده نمایید.
- سؤالات خود را در گروه تلگرام درس مطرح نمایید. به هیچ وجه کد یا پاسخ سؤالات را در گروه به اشتراک نگذارید.
- برای تمامی سؤالات در فایل گزارش، کدها را نیز توضیح دهید.
- در صورت بروز مشکل با ایمیل زیر در ارتباط باشید:

khoramfar@ut.ac.ir

## سؤال ۱. تحلیل داده‌های فروش بازی‌های ویدیویی با RDD

۳۰ نمره

در این سؤال فایل video_game_sales.csv شامل اطلاعات فروش بازی‌های ویدیویی در اختیار شما قرار داده شده است. هدف این سؤال، آشنایی با پردازش داده‌ها با استفاده از انتزاع RDD در چارچوب Spark است.

**نکات مهم:**

- در تمامی بخش‌های این سؤال الزاماً باید از Spark RDD استفاده نمایید.
- مدیریت مقادیر تهی (Missing Values) در داده‌ها بر عهده‌ی شماست.
- ترتیب مرتب‌سازی خروجی‌ها باید دقیقاً مطابق صورت سؤال باشد.

### ۱.۱ ۱۰ بازی پرفروش

فایل داده را خوانده و (header)، برای هر عنوان بازی (title) مجموع فروش جهانی (total_sales) را محاسبه کنید. سپس ۱۰ بازی پرفروش را بر اساس فروش کل، به ترتیب نزولی نمایش دهید.

### ۲.۱ تحلیل فروش بر اساس ژانر و منطقه

برای هر ژانر (genre)، مجموع فروش را در هر یک از مناطق زیر محاسبه کنید:

- آمریکای شمالی (na_sales)

- ژاپن (jp_sales)
- اروپا و سایر مناطق (pal_sales)

خروجی نهایی را به‌صورت نمودار گزارش دهید.

### ۳.۱ توسعه‌دهندگان با بالاترین امتیاز منتقدان (۱۰ نمره امتیازی)

در این بخش فقط رکوردهایی را که دارای مقدار امتیاز منتقدان (critic_score) هستند در نظر بگیرید. برای هر توسعه‌دهنده (developer) موارد زیر را محاسبه کنید:

- میانگین critic_score
- مجموع total_sales

سپس توسعه‌دهندگانی را که مجموع فروش آن‌ها حداقل ۱۰ میلیون است نگه داشته و ۱۰ توسعه‌دهنده برتر را به ترتیب نزولی میانگین امتیاز منتقدان نمایش دهید.

## سؤال ۲. مدل‌سازی معنایی کلمات با Word2Vec

۴۵ نمره

روش Word2Vec تکنیکی برای تبدیل کلمات به بردارهای معنایی است. این مدل که توسط توماس میکولوف در سال ۲۰۱۳ معرفی شد، به هر کلمه یک بردار عددی (Embedding) اختصاص می‌دهد؛ به‌گونه‌ای که کلمات با معنی مشابه، بردارهای نزدیک‌تری به یکدیگر در فضای برداری داشته باشند. علاوه بر این، بردارهای به‌دست‌آمده به‌گونه‌ای هستند که می‌توان برخی روابط معنایی بین کلمات را به‌صورت روابط جبری مشاهده کرد؛ به‌طور مثال رابطه‌ی زیر معمولاً برقرار است:

$$e(\text{king}) - e(\text{man}) + e(\text{woman}) \approx e(\text{queen})$$

در این سؤال هدف، آشنایی با کارکرد Word2Vec و بررسی روابط معنایی بین کلمات با استفاده از داده‌های متنی برگرفته از ویکی‌پدیا (Wikipedia) است.

داده های این سوال در سرور در آدرس زیر قرار دارند:

/home/shared_files/CA۲/wiki_corpus

### ۱.۲ آموزش مدل Word2Vec

با استفاده از داده‌های متنی که در اختیار شما قرار داده شده است، یک مدل Word2Vec را در محیط Spark آموزش دهید.

- برای پیاده‌سازی از زبان پایتون و کتابخانه‌ی PySpark استفاده نمایید.
- انتخاب تنظیمات مدل (مانند ابعاد بردار و اندازه‌ی پنجره) بر عهده‌ی شماست.

### ۲.۲ استخراج بردارهای کلمات

پس از آموزش مدل، بردارهای تعبیه‌شده (Embedding Vectors) مربوط به کلمات زیر را استخراج و گزارش نمایید:

```
iran, tehran, learning, science
```

### ۳.۲ محاسبه شباهت کلمات

برای هر یک از کلمات قسمت قبل (بخش ۲.۲)، ۵ کلمه‌ی نزدیک را به همراه میزان شباهت کسینوسی آن‌ها با استفاده از مدل آموزش‌داده‌شده محاسبه و گزارش کنید.

### ۴.۲ بررسی رابطه‌ی معنایی بردارها

رابطه‌ی زیر را برای بردارهای به‌دست‌آمده بررسی کرده و مشاهده نمایید تا چه حد سمت چپ عبارت با سمت راست آن برابر است:

$$e(\text{king}) - e(\text{man}) + e(\text{woman}) \approx e(\text{queen})$$

نتیجه‌ی به‌دست‌آمده را به‌صورت کوتاه تحلیل کنید.

**منابع مطالعاتی**

جهت مطالعه بیشتر می‌توانید به مقالات اصلی و منابع زیر مراجعه کنید :

- Efficient Estimation of Word Representations in Vector Space (2013)
- Distributed Representations of Words and Phrases and their Compositionality (2013)
- Word2Vec Explained - Blog Post

## سؤال ۳. تحلیل داده‌های تایتانیک و تعامل با HDFS

۳۰ نمره

این سوال باید الزاماً بر روی کلاستر درس انجام شود. هدف از این سوال کار با فایل سیستم توزیع‌شده (HDFS) و پیاده‌سازی خط لوله یادگیری ماشین روی داده‌های ذخیره شده در کلاستر است.

### ۱.۳ آپلود داده‌ها در HDFS

در HDFS پوشه‌ای با نام **شماره دانشجویی** خود ایجاد کرده و فایل Titanic-Dataset.csv را در آن آپلود کنید. دستورات استفاده شده برای این کار را در گزارش ذکر نمایید.

### ۲.۳ تحلیل اکتشافی داده‌ها

فایل را از HDFS خوانده و موارد زیر را محاسبه و گزارش کنید:

- نرخ بقای مسافران به تفکیک جنسیت (Sex).
- نرخ بقای مسافران به تفکیک کلاس بلیت (Pclass).
- میانگین سن (Age) برای مسافران زنده‌مانده و زنده‌نمانده به صورت مجزا.

### ۳.۳ آموزش مدل طبقه‌بند

پس از مدیریت مقادیر تهی به‌صورت مناسب، با استفاده از ویژگی‌های زیر یک مدل رگرسیون لجستیک (Logistic Regression) برای پیش‌بینی ستون Survived آموزش دهید:

```
Pclass, Sex, Age, Fare, Embarked
```

داده‌ها را به نسبت ٪۸۰ آموزش و ٪۲۰ تست به صورت تصادفی تقسیم کنید.

### ۴.۳ ارزیابی و ذخیره نتایج

معیار دقت (Accuracy) مدل را روی داده‌های تست محاسبه و گزارش کنید. نتیجه نهایی را در یک فایل متنی با نام accuracy.txt ذخیره کرده و این فایل را در پوشه‌ی خود (که در بخش ۱.۳ ایجاد کردید) در HDFS قرار دهید. برای تحلیل بهتر نتایج می‌توانید نمودار رسم کنید.

## نحوه تحویل تمرین کامپیوتری ۳

۱. گزارش پروژه را در قالب فایل PDF تهیه کنید.

۲. تمام فایل‌های مربوط به تمرین را طبق ساختار جدول زیر نام‌گذاری کرده و در انتها همه را در یک فایل zip با نام شماره دانشجویی خود در سامانه بارگذاری کنید.

| سؤال | نام فایل |
|---|---|
| گزارش | `report.pdf` |
| ۱ | `VideoGameSales.ipynb` |
| ۲ | `Word2Vec.ipynb` |
| ۳ | `TitanicHDFS.py` |