
**پروژه سوم درس داده کاوی**

**عنوان :**

خوشه بندی مقالات علمی با استفاده از مجموعه داده covid-19

**استاد:**

دکتر حسین رحمانی

پاییز ۱۴۰۲

# راهنمای پروژه

- مهلت ارسال پروژه تا **ساعت ۲۳:۵۹ تاریخ ۲۰ / ۰۹ / ۱۴۰۲** است و قابل تمدید نخواهد بود.
- به ازای هر روز تاخیر **۲۵ درصد** از نمره پروژه کسر خواهد شد.
- پاسخ به سوالات این پروژه باید در قالب یک گزارش با **فرمت PDF** یا به همراه **توضیحات فایل نوتبوک** (**Markdown**) ارائه شود.
- در صورت ارائه گزارش در قالب **توضیحات فایل نوتبوک**، توضیحات باید کامل، جامع و شفاف باشد.
- در صورت ارائه گزارش با **فرمت PDF**، فایل **کدهای اجراشده** نیز پیوست شود.
- تمامی فایل‌های این پروژه (گزارش و کدها) در قالب یک **فایل فشرده rar یا zip** با نام‌گذاری زیر ارسال شود .

StudentNumber_FirstName_LastName_Prj03.zip

- فایل تمرین را **حتما** در سامانه LMS آپلود نمایید. بدیهی است که تحویل از طریق ایمیل و یا سایر راه‌های ارتباطی قابل پذیرش نخواهد بود.
- **رعایت نکات نگارشی** در نوشتن گزارش نمره مثبت خواهد داشت.
- برای پاسخ به سوالات این پروژه **حتما** باید از زبان برنامه‌نویسی **پایتون** استفاده شود.

**مباحث تحت پوشش:** Clustering, Topic modeling

**هدف پروژه:** به دلیل حجم بالا و روزافزون داده‌های متنی آنلاین و محدودیت توانایی خواندن انسان، تقاضا برای سیستم‌های topic modeling و خوشه بندی افزایش یافته است. در این پروژه قصد داریم که دانشجویان به خوشه بندی مجموعه داده covid-19 بپردازند و روابط معنی داری بین مقالات این مجموعه داده استخراج و آن‌ها را بر اساس topic خوشه بندی کنند. این پروژه بر اساس مقاله (An integrated clustering and BERT framework for improved topic modeling) در این زمینه طراحی شده است و برای انجام مراحل کار می‌توانید آن را مطالعه کنید.

**مراحل انجام پروژه به ترتیب در ادامه آمده است:**

**۱- فایل ورودی**

CORD-19 مجموعه ای از مقالات علمی در مورد COVID-19 است که توسط تیم Semantic Scholar در موسسه Allen برای پشتیبانی از متن کاوی و تحقیقات پردازش زبان طبیعی نگهداری می‌شود. می‌توانید مقاله (CORD-19: The COVID-19 Open Research Dataset) که این مجموعه داده را معرفی کرده است، برای توضیح بیشتر مطالعه کنید. در این پروژه از اولین نسخه منتشر شده مجموعه داده استفاده خواهید کرد که به این فایل پیوست شده است. پارامترهای در نظر گرفته شده برای این پروژه زمان انتشار، عنوان و چکیده مقالات می‌باشند.

**۲- پیش‌پردازش‌داده**

یکی از مهم ترین مراحل قبل از شروع تحلیل، پیش پردازش داده‌های ورودی است. با توجه به نوع و ساختار داده های ورودی عملیات های متفاوتی می توان انجام داد شامل: حذف ستون های بدون استفاده، حذف داده‌های تکراری و مقادیر null، پیش پردازش‌های مورد نیاز برای داده متنی(stop word , stemming , lemmatization , … )

**۳- استخراج ویژگی**

در اینجا استخراج ویژگی برای تبدیل داده‌های متنی به بردارهای عددی استفاده می‌شود. از روش های متفاوتی برای استخراج ویژگی در داده های متنی می توان استفاده کرد مانند BOW, GloVe, TF-IDF, Word2Vec … . **در این بخش حداقل از دو روش استفاده شود.**

**۴- Topic Modeling - LDA**

یک روش طبقه‌بندی بدون نظارت متن‌ها و یا اسناد است که شبیه به خوشه‌بندی داده‌های عددی می‌باشد. کارکرد آن پیدا کردن یک سری موضوع (topic)-است حتی وقتی که اطمینان ندارید که دنبال چه چیزی

می‌گردید. یکی از پر استفاده ترین روش های (Latent Dirichlet Allocation) LDA است که این کار را براساس کلمات انجام می دهد.

### ۵- کاهش ابعاد (Dimensionality Reduction)

پس از انجام مراحل بالا، نیاز خواهید داشت که از روش های کاهش ابعاد استفاده کنید تا بتوانید که با حذف ویژگی های نامرتبط، تحلیل سریع تر و دقیق تری داشته باشید. از روش های زیر می توانید استفاده بکنید:

Principal Component Analysis (PCA)

Linear Discriminant Analysis (LDA)

Generalized Discriminant Analysis (GDA)

### ۶- انتخاب مدل خوشه بندی

با استفاده از الگوریتم‌ها و روش‌های خوشه بندی مناسب، داده‌های خود را به صورت بدون نظارت دسته بندی کنید. از روش Elbow برای پیدا کردن k (تعداد خوشه‌ها) مناسب استفاده کنید. داده‌های پیش‌پردازش شده را به مدل های مختلف خوشه‌بندی تزریق کرده و مدل را برای خوشه‌بندی داده‌ها آموزش دهید. **حداقل از دو روش استفاده شود.**

### ۷- نتیجه‌گیری

با روش‌های ارزیابی (همانند معیار ارزیابی **Silhouette**) و بصری سازی مدل‌های مختلف نتایج خود را تحلیل و به سوالات زیر پاسخ دهید.

- کدام یک از روش استخراج ویژگی و مدل خوشه بندی عملکرد بهتری دارند؟
- دلیل خود برای این انتخاب را ذکر کنید.
- با بررسی خوشه‌ها تحلیل کنید که هر خوشه نماینده‌ی کدام نوع از موضوعات می‌باشند.
- داده‌ها را به صورت دو بعدی نمایش دهید و با استفاده از رنگ‌های مختلف خوشه‌ی هر داده را مشخص کنید.

**نکته:** هرگونه ابتکار عمل، خلاقیت، تحلیل‌های دیگر و استفاده از روش‌های بیشتر نمره‌ی اضافه خواهد داشت.

موفق باشید.