



بازیابی پیشرفته اطلاعات (فاز دوم)

نیم سال 4042

مدرس: دکتر اسکندری

دانشکده‌ی مهندسی کامپیوتر – آزاد شیراز

زمان تحویل: 6 خرداد

طراحان: حمید نامجو، امیرحسین همتی و

علی مجاهد

- پروژه انفرادی و یا نهایتاً 2 نفره است.
- فقط در موارد ذکرشده مجاز به استفاده از کتابخانه‌های آماده هستید.
- هر گروه باید کد پروژه خود را در مهلت اعلام شده ارسال نماید.
- علاوه بر گزارش پروژه، در انتهای ترم یک ارائه و دفاع نیز از این پروژه گرفته خواهد شد. مهلت ارسال پاسخ تا ساعت 23:59 روز مشخص شده است.
- همکاری و هم فکری شما در انجام پروژه مانع ندارد اما پاسخ‌های ارسالی هر گروه حتماً باید توسط خود او نوشته شده باشد.
- در صورت هم فکری و یا استفاده از هر منابع خارج درسی، نام هم فکران و آدرس منابع مورد استفاده برای حل سوال مورد نظر را ذکر کنید.

1 شرح کلی تمرین

در این فاز از پروژه‌ی عملی، قرار است یک زنجیره‌ی کامل IR/ML برای مجموعه‌ای کوچک از فیلم‌ها پیاده‌سازی کنید. نقطه‌ی شروع شما داده‌های crawl شده از IMDB و یک inverted index آماده است و در انتها باید:

- ژانر اصل هر فیلم را استخراج و یک برچسب واحد برای هر سند تعریف کنید.
- نمایش برداری اسناد را به صورت ماتریس تنک TF یا TF-IDF بسازید.
- روی ویژگی‌ها پیش پردازش، هرس و انتخاب ویژگی انجام دهید.
- چند طبقه‌بند (Linear SVM، k-NN، Rocchio) را آموزش و ارزیابی کنید.
- روی همان فضای برداری، خوشه‌بندی انجام دهید و کیفیت خوشه‌ها را نسبت به برچسب‌های ژانر بسنجید.
- و در نهایت اثر بُعد ویژگی را روی کارایی دسته‌بندی و خوشه‌بندی بررسی کنید.

تمام کدنویس و تحلیل ها داخل نوت بوک IR_Spring_2026_Phase_2_Question.ipynb انجام می شود. ساختار نوتبوک با همین شماره بندی بخش ها طراح شده و تمام قسمت هایی که باید پیاده سازی کنید با #TODO علامت گذاری شده اند.

موفق و موید):