بسمه تعالی

سیستم‌های توصیه‌گر

Recommender Systems

تکلیف ۱

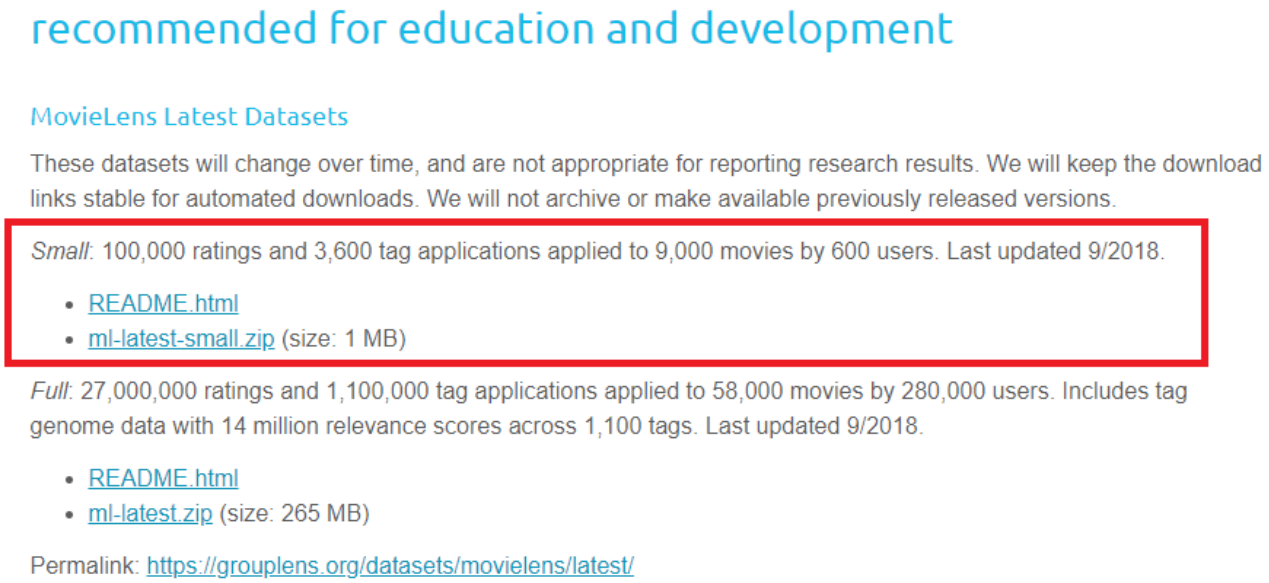
توجه

۱. **برای پیاده سازی این تکلیف از زبان برنامه نویسی Python استفاده نمایید.**
   - **از کتابخانه های pandas، numpy و math برای لود کردن مجموعه داده و محاسبات استفاده شود.**
   - **کد ها ترجیحا به صورت شیء گرا و با استفاده از کلاس پیاده سازی شوند.**
   - **برای ترسیم از کتابخانه matplotlib استفاده شود و نمودار تولید شده را به گزارش اضافه نمایید.**
۲. **خروجی انجام تکلیف، مشتمل بر موارد زیر است:**
   - **کدهای پیاده سازی به زبان پایتون**
   - **یک فایل Document.docx که مستندات نحوه انجام پروژه است.**
     - **شامل نحوه پیاده سازی، نتایج، تحلیل نتایج و سایر موارد خواسته شده است.**
     - **به عنوان نمونه، به فایل ضمیمه تکلیف مراجعه فرمایید.**
۳. **کلیه کد ها را در یک پوشه با نام Code قرار داده و به همراه فایل مستندات فشرده نمایید**
   - **(پسوند rar یا zip)**
۴. **در صورتی که فایل ارسالی مشکل داشته باشد، تبعات آن بر عهده دانشجو است**
   - **مثلا فایل فشرده Extract نشود**
۵. **ارسال تکلیف را به روز و ساعات آخر موکول نکنید شاید اینترنت یا سامانه مشکل داشته باشد**
   - **ارسال تکلیف فقط و فقط از طریق سامانه مشخص شده پذیرفتنی است و ارسال آن به هر شکل دیگری در نظر گرفته نخواهد شد.**

تکلیف ۱

مجموعه داده

۱- نسخه ای از مجموعه داده MovieLens را که در شکل زیر مشخص شده، از آدرس (https://grouplens.org/datasets/movielens) دانلود نمایید.


recommended for education and development

MovieLens Latest Datasets

These datasets will change over time, and are not appropriate for reporting research results. We will keep the download links stable for automated downloads. We will not archive or make available previously released versions.

Small: 100,000 ratings and 3,600 tag applications applied to 9,000 movies by 600 users. Last updated 9/2018.

- README.html
- ml-latest-small.zip (size: 1 MB)

Full: 27,000,000 ratings and 1,100,000 tag applications applied to 58,000 movies by 280,000 users. Includes tag genome data with 14 million relevance scores across 1,100 tags. Last updated 9/2018.

- README.html
- ml-latest.zip (size: 265 MB)

Permalink: https://grouplens.org/datasets/movielens/latest/

۱- مشخصات مجموعه داده MovieLens 100K را مستند و گزارش نمایید.

۲- برای این مجموعه داده هیستوگرام توزیع رای‌های آن را ترسیم نماید

a) محور x ها معرف رای و محور y معرف میزان درصد از کل رای ها است.

۳- رای ها فایل ratings.csv را بر اساس timestamp به صورت صعودی مرتب کرده و برای هر کاربر، رای های او را به ۵ بخش افراز نمایید.

a) با اجماع رای های بخش های معادل همه کاربران، نهایتا مجموعه داده کلیه رای ها، به ۵ بخش Fold-1 تا Fold-5 تقسیم می شود.

b) در ادامه از این ۵ بخش برای انجام 5-fold cross-validation استفاده نمایید.

پیاده سازی الگوریتم ها

۴- الگوریتم User-Based CF را که در اسلاید های ۲۵ تا ۲۷ درس توضیح داده شده است را پیاده سازی نمایید.

۵- الگوریتم Per User Average را با فرمول پیش بینی زیر پیاده سازی نمایید:

$$p_{u,i} = \bar{r}_{u,*}$$

۶- الگوریتم Per Item Average را با فرمول پیش بینی زیر پیاده سازی نمایید:

$$p_{u,i} = \bar{r}_{*,i}$$

۷- الگوریتم Global Average با فرمول پیش بینی زیر پیاده سازی نمایید:

$$p_{u,i} = \bar{r}_{*,*}$$

که در آن $\bar{r}_{*,*}$ میانگین کل رای های داده شده در مجموعه داده است

ارزیابی الگوریتم های پیاده سازی شده

۸- کارایی سیستم‌های توصیه گر Global Average، Per User Average ، Per Item Average و User-Based را با استفاده از مجموعه داده MovieLens و طبق روالی که در ادامه توضیح داده می شود، مقایسه و گزارش نماید.

a) برای پیاده سازی الگوریتم User-Based از پارامترهای $\theta = 0,\ k = \infty$ استفاده نمایید.

b) طبق توضیحی که در بند ۳ داده شد و اعمال تکنیک 5-fold cross-validation، هر بار از یک بخش برای تست و از ۴ بخش دیگر برای آموزش استفاده شود. در نهایتا از نتایج متوسط گیری و گزارش شود.

c) برای ارزیابی الگوریتم ها از دو معیار Mean Absolute Error (MAE) و Prediction Coverage استفاده نمایید.

$$\text{MAE} = \frac{\sum_{u \in U} \sum_{i \in Test_u} |p_{u,i} - r_{u,i}|}{\sum_{u \in U} |Test_u|}$$

$$\text{Coverage} = \frac{\sum_{u \in U} \sum_{i \in Test_u} \rho_{u,i}}{\sum_{u \in U} |Test_u|}$$

$$\rho_{u,i} = \begin{cases} 1, & p_{u,i} \neq \bullet \\ 0, & p_{u,i} = \bullet \end{cases}$$

d) برای هر یک از دو معیار یک نمودار میله ای ترسیم کنید که محور x الگوریتم ها و محور y معیار مد نظر است

e) نتایج حاصل را مورد تحلیل، ارزیابی و بحث قرار دهید.

ارزیابی نقش پارامترهای الگوریتم کاربر محور

۹- کارایی سیستم توصیه گر User-Based را با استفاده از مجموعه داده MovieLens و پارامترهای زیر، مقایسه و گزارش نماید.

a) استفاده از k-نزدیکترین همسایه برای مقادیر $k = 1,2,3, \dots, 25, \infty$

- برای هر یک از دو معیار MAE و Coverage یک نمودار میله ای ترسیم کنید که محور x مقادیر k و محور y معیار مد نظر است.
- نتایج حاصل را مورد تحلیل، ارزیابی و بحث قرار دهید.

b) استفاده از حد آستانه θ برایم مقادیر $\theta = -1, -0.9, -0.8, -0.7, \dots, 0, 0.1, 0.2, \dots, 0.9, 1$

- برای هر یک از دو معیار MAE و Coverage یک نمودار میله ای ترسیم کنید که محور x مقادیر θ و محور y معیار مد نظر است.
- نتایج حاصل را مورد تحلیل، ارزیابی و بحث قرار دهید.

۱۰- کد برنامه به همراه فایل مستندات (نسخه docx و نسخه pdf) را مطابق توضیحات ارائه شده در ابتدای تکلیف ارسال فرمایید.

a) لازم به ذکر است که گزارش ضمیمه، صرفا یک قالب پیشنهادی است که باید تکمیل شود. دانشجویان عزیز می توانند به هر نحوی بر اساس تکلیف درخواست شده آن را ویرایش نمایند

با آرزوی توفیق الهی.