درس داده‌کاوی

تمرین سری پنجم

نکات مهم

- به ازای هر سوال، یک پوشه با نام question_i بسازید و فایل‌ها/جواب‌های مرتبط با آن سوال را آنجا قرار دهید. چه در سوال‌های تئوری، چه در سوال‌های عملی.

- جواب سوال‌های تئوری را می‌توانید تایپ و یا اسکن کنید و یا حتی در محیط jupyter بنویسید. در هر حالت، فایل آن را در همان پوشه مربوطه قرار دهید.

- در انتها تمامی پوشه‌های سوالات (question_i) را زیپ کنید و نام آن را HW5_YourName_YourStudentNumber بگذارید و در کوئرا آپلود کنید.

- متن سوالات را در همان ابتدا مطالعه کنید تا بتوانید مشکلات و ابهامات را هرچه زودتر در کوئرا بپرسید.

تمرین اول

آنالیز اسپاتیفای

تقریبا همه‌ی ما با اسپاتیفای آشنا هستیم. یک پلتفورم برای گوش کردن به موسیقی‌ها. در این تمرین می‌خواهیم چندین سوال را در رابطه با این پلتفورم بررسی کنیم.

1. فرض کنید که در ورودی شرکت اسپاتیفای، یک مانیتور خیلی بزرگ وجود دارد و روی آن تنها یک عدد نوشته شده است و هر روز آپدیت می‌شود. قصد آن‌ها این بوده که تنها با نشان دادن یک عدد، وضعیت پیشرفت کلی شرکت نمایان شود. به نظر شما این عدد چه چیزی بوده است؟

2. فرض می‌کنیم عددی که در نظر گرفتید، معیار x را محاسبه می‌کرده است (به طور مثال، تعداد نصب‌های برنامه در گوگل پلی). برای بیشتر کردن این عدد، چه کارهایی می‌توان انجام داد؟ یک ریکامندر سیستم ساده طراحی کنید که به افزایش معیار x کمک کند.

3. در صورت امکان برنامه اسپاتیفای را باز کنید و یا از تصاویر آن در سایت‌ها و گوگل استفاده کنید. حداقل سه مورد از ریکامندر سیستم‌های استفاده شده در اسپاتیفای را پیدا کنید، مدل و پارامترهای هر کدام را به صورت شهودی توضیح دهید.

4. در اسپاتیفای یک قابلیت به نام discover weekly وجود دارد. در این قسمت، هر هفته به شما آهنگ‌های جدیدی که مورد علاقه شما خواهند بود نشان داده می‌شود. با جست و جو در اینترنت سعی کنید الگوریتم این ریکامندر سیستم را بفهمید و آن را توضیح دهید.

تمرین دوم

انواع و اقسام

این تمرین جهت بالا بردن اطلاعات شما از ریکامندر سیستم‌ها است. در اینجا تقریبا تمامی روش‌های مختلف نام برده شده است. این روش‌ها را مطالعه کنید و به صورت خلاصه الگوریتم هر کدام را توضیح دهید.

1. Content based recommendation
2. Collaborative filtering
3. Nearest neighbours
4. Latent factor methods
5. Matrix factorization
6. Deep learning embedding
7. Hybrid approaches
8. در سال ۲۰۰۹، نتفلیکس یک مسابقه راه انداخت و به بهترین الگوریتم ریکامندر سیستم در آن مسابقه، ۱ میلیون دلار جایزه داد! این الگوریتم را بخوانید و آن را توضیح دهید. چرا این الگوریتم روی دیتاست نتفلیکس انقدر کارا بود؟ (امتیازی)

تمرین سوم

مشکلات سر راه

همه چیز آنقدرها هم خوب و عالی نیست. ریکامندر سیستم‌ها هم مشکلات خودشان را دارند. در اینجا ۴ مورد از اصلی‌ترین مشکلات ریکامندر سیستم‌ها نوشته شده است، هر کدام را توضیح دهید و برای هر کدام راه حل‌های مناسبی پیشنهاد دهید.

1. Cold start
2. Exploition
3. Interpretability
4. Scalibility

تمرین چهارم

وقت عمل است

تا اینجا دیگر به صورت کامل به تئوری‌های ریکامندر سیستم‌ها مسلط شده‌ایم. حال وقت کد زدن فرا رسیده. دیتاست MovieLens، یک دیتاست از فیلم‌ها و ریتینگ کاربران به آن‌ها است. فایل **ml-latest-small.zip** را دانلود کنید. در این فایل همه‌ی اطلاعات لازم برای راه اندازی یک ریکامندر سیستم وجود دارد.

با روش دلخواه خودتان یک مدل ریکامندر سیستم بسازید که در ورودی آی‌دی یک کاربر دلخواه را بگیرد و در خروجی نام ۳ فیلم را برای آن کاربر پیشنهاد دهد.

تمرین پنجم

گراف دوبخشی

یک گراف دوبخشی با ۳ راس در مجموعه X و ۳ راس در مجموعه Y در نظر بگیرید. فرض می‌کنیم تمامی راس‌های مجموعه X به مجموعه Y لینک دارند. همچنین هر راس به خودش نیز لینک دارد. یک مثال از همچین گرافی در دنیای واقعی می‌تواند سایت‌های provider محصولات (مانند آمازون) و سایت‌های affiliation، که این محصولات را پروموت می‌کنند، باشد.

1. ماتریس مجاورت گراف ذکر شده را محاسبه کنید.

2. عملیات پیج‌رنک را به صورت دستی تا نقطه تعادل بنویسید. به صورت شهودی چگونه می‌توان این نتیجه را توضیح داد؟

3. با اضافه کردن مالیات ۵۰ درصدی به این سیستم، پیج‌رنک راس‌ها چگونه تغییر می‌کند؟

4. حالت کلی این سوال با 2n راس را در نظر بگیرید. چه تغییراتی در این گراف ایجاد کنیم تا پیج‌رنک همه راس‌ها در نهایت با یک‌دیگر برابر شود؟ (امتیازی)

تمرین ششم

رابطه‌های دوستی

در این سوال قصد داریم تا **داده‌های دوستی** را به کمک کتابخانه NetworkX بررسی کنیم. بدین منظور به سوال‌های زیر پاسخ دهید.

1. ابتدا گراف جهت‌داری از داده‌ها بسازید. تعداد راس‌ها و یال‌های این گراف چقدر است؟

2. میانگین درجه‌ها در این گراف چند است؟ هیستوگرامی از درجه‌های راس‌های این گراف نشان دهید و آن را تحلیل کنید.

3. متریک پیج‌رنک هر راس را محاسبه کنید و شماره ۱۰ راس اول با بیشترین پیج‌رنک را چاپ کنید.

4. طولانی‌ترین مسیر دوستی برای راس ۸۹۴۲ را پیدا کنید و آن را چاپ کنید.

5. الگوریتم پیج‌رنک را به صورت دستی پیاده‌سازی کنید. در این الگوریتم، راس‌هایی که درجه آن‌ها از ۲۰ بیشتر است را در teleport set قرار می‌دهیم و میزان مالیات را ۱۰٪ در نظر می‌گیریم. پیج‌رنک راس با آی‌دی ۱۰۰۰۰ چند می‌شود؟ (امتیازی)

تمرین هفتم

گوگلِ گوگولی

می‌دانیم الگوریتم پیج‌رنک توسط گوگل سرچ استفاده می‌شود. در این سوال قصد داریم که ببینیم گوگل از ۰ تا ۱۰۰ چگونه یک سرچ کوئری را محاسبه می‌کند. **کل نمره این سوال امتیازی است**. در این سوال به جواب‌ها به صورت نسبی نمره داده می‌شود. یعنی به بهترین جواب (با فرض اینکه از یک مینیممی بهتر باشد) نمره ۱۰۰ داده می‌شود و به بقیه به همان نسبت کمتر.

موضوعاتی که می‌توانید راجع به آن‌ها تحقیق کنید این موارد هستند:

1. گوگل در ورودی تنها یک متن می‌گیرد. این متن را چگونه تحلیل می‌کند؟

2. بعد از تحلیل و فهمیدن منظور سوال کاربر، چگونه پیج‌های مرتبط را به این سوال پیدا می‌کند؟

3. بعد از پیدا کردن این پیج‌ها، چگونه آن‌هارا بر اساس کیفیتشان مرتب می‌کند؟

4. گوگل چگونه اطلاعات کاربران را در سرچ تاثیر می‌دهد؟ (مثلا کسی که در آمریکا کلمه soccer را سرچ می‌کند دنبال نتیجه متفاوتی است از کسی که در انگلیس این کلمه را سرچ می‌کند. یا اگر کسی عاشق فیلم دیدن باشد، با سرچ کلمه football، گوگل احتمالا برای او فیلمی را می‌آورد که اسم آن football باشد. در واقع گوگل، اطلاعات کاربران مانند موقعیت مکانی و سلایق آن‌ها را در نتایج سرچ تاثیر می‌دهد)

از آنجا که کل این سوال امتیازی است، نمره دادن **سختگیرانه‌تر** می‌باشد.