کاوش داده‌گان انبوه

دکتر سامان هراتی زاده
دانشگاه تهران - دانشکده علوم و فنون
ترم اول سال تحصیلی – پاییز ۱۳۹۹

دستیاران آموزشی
مهدی صادقی (mahdisadeghi@ut.ac.ir)
شیوا پارسا راد (shiva.parsarad@ut.ac.ir)
علیرضا جعفری (Alirezajafari@ut.ac.ir)
ایمان برازنده (barazandeh.iman@ut.ac.ir)

پروژه نهایی : این ترم هم به آخر رسید

مهلت تحویل این تمرین سه شنبه ۵ اسفند ساعت ۲۳:۵۹ می‌باشد.

این ترم هم بالاخره تمام شد. به عنوان پروژه نهایی می توانید **از بین دو تمرین زیر یکی را انتخاب کرده** و پیاده سازی کنید.

(تمرین اول): موتور جستجوی تمرین چهارم را (روی همان دیتا ست) با استفاده از Map-Reduce پیاده سازی و روی پلتفرم هادوپ اجرا کنید. برای ارتباط با هادوپ از پایتون، می توانید از کتابخانه MRJob استفاده کنید.

(تمرین دوم: تا ۲۵٪ نمره اضافی): می‌خواهیم یک **سیستم توصیه‌گر مبتنی بر گراف** بنویسیم. دیتاست این تمرین نیز Movielenz100k است. (تعداد کاربران را محدود به کارایی سخت افزار کامپیوتر خود کنید.) کتابخانه Networkx پایتون کار شما را برای پیاده سازی راحت می‌کند.

برای پیاده‌سازی این سیستم بصورت زیر عمل کنید:

- یک گراف دو بخشی[1] کاربر-آیتم بسازید بطوریکه یک بخش آن را کابران و یک بخش آن را آیتم‌ها تشکیل می‌دهند. هر کاربر به شرطی به یک آیتم لینک می‌شود که امتیازش به آن بیشتر یا مساوی ۳ (با مقادیر متفاوت تست کنید و بهترین را انتخاب کنید) باشد.
- گراف را Fold کرده و آن را تبدیل به یک گراف کاربر-کاربر کنید. کاربرانی به هم متصل می‌شوند که حداقل به ۳ محصول بطور مشترک امتیاز داده باشند. (با مقادیر متفاوت تست کنید و بهترین را انتخاب کنید)
- با استفاده از یکی از الگوریتم‌های کلاسترینگ (اجتماع یابی[2]) مثل Louvain کاربران مشابه را پیدا کنید.
- آیتم‌هایی که کاربران در هر کلاستر به آنها امتیاز داده‌اند در یک لیست جمع کنید.

[1] Bipartite
[2] Community Detection

- برای هر کاربر آیتم‌هایی که مشاهده کرده را از آن حذف کنید و آیتم‌های باقیمانده را بر اساس معیار خاصی مرتب کنید (هر معیاری! مهم نیست، مثلا براساس مجموع امتیازهایی که این آیتم‌ها دارند یا تعداد امتیازها. هر چه معنی‌دار تر بهتر.).
- به هر کاربر Top-5 آیتم از این لیست را پیشنهاد دهید.

گراف‌ها در هر مرحله و همچنین خوشه‌ها بصورت گرافیکی نیز نمایش داده شوند. لیست پیشنهادها را نیز برای چند کاربر گزارش کنید.

نکات تکمیلی

- از زبان برنامه سازی پایتون کنید.
- کدها با نامگذاری مناسب در پوشه SOURCES قرار گیرد.
- گزارش کامل تمرین را با ساختار مناسب، بدون ابهام و به زبان ساده نوشته و با فرمت PDF در پوشه دیگری قرار دهید. برای نگارش گزارش‌ها نیز امتیاز مثبت و منفی در نظر گرفته شده است، پس برای نگارش گزارش خود وقت مناسبی اختصاص دهید.
- پاسخ تمرین خود را به ایمیل barazandeh.iman@ut.ac.ir ارسال نمایید. هر دو پوشه را در یک فایل ZIP با نام MMDS_STID_FullName_FP مثلا MMDS_830123451_ImanBarazandeh_FP قرار داده و در ایمیلی با همین عنوان نیز ارسال کنید.
- در صورت مشخص شدن هر نوع تقلب انفرادی یا مشترک نمره تمرین صفر در نظر گرفته می‌شود.
- به ازای هر روز تاخیر ۲۵٪ از امتیاز تمرین از دست خواهید داد.
- هر گونه ابهام را با دستیاران آموزشی مطرح کنید.