تمرین سری اول درس کاوش دادگان انبوه

سؤال اول: (کتبی)

به سؤالات مشخص شده از کتاب مرجع درس پاسخ دهید.

3.2.3	3.3.7	3.5.9	3.7.1	3.8.2
3.3.3	3.4.1	3.6.1	3.7.2	3.9.1
3.3.5	3.5.1	3.6.2	3.8.1	3.9.2

مهلت ارسال سوال اول: ۲۸ آبان

توجه داشته باشید که این سوالات را میتوانید تایپ کرده و ارسال کنید و یا روی کاغذ نوشته و فایل اسکن آن را به ایمیل اعلام شده ارسال نمایید.

کتاب مرجع:

Mining of massive datasets

Written by: Jure Leskovec, Jeffrey Ullman, Anand Rajaraman

کتاب مرجع ضمیمه شده است.

سؤال دوم: (پیاده سازی)

فایل شینگلی از ۱۲۳۲ فایل خبری در اختیار داریم. ساختار این فایل در ادامه مشخص شده است.

- شباهت فایل های خبری را بر اساس min hash signature به دست آورید. تعداد جایگشت های تصادفی را ۵۰ و ۱۰۰ در نظر بگیرید.
- کدهای پیاده سازی خود را به همراه گزارش ارسال نمایید. کد ارسالی باید پس از اجرا آدرس فایل شینگل و تعداد جایگشت های تصادفی را از کاربر دریافت کند و لیست شباهت خبرها را در یک فایل خروجی ذخیره نماید.
- برای پیاده سازی از کدهای آماده استفاده نکنید.
- فایل گزارش خود را مطابق نمونه ارسالی کامل کنید.
- **مهلت ارسال سوال دوم: ۵ آذرماه**
- بعد از ارسال تمرین ها تا تاریخ مشخص شده، زمان ارائه ی حضوری تمرین اعلام می شود.
- پاسخ های خود را به ایمیل های n.hazrati.7@gmail.com و mdaneshvar69@gmail.com ارسال نمایید. حتماً در عنوان ایمیل خود عبارت MDS-HW02 را قرار دهید.

ساختار فایل Shingle

$$\begin{bmatrix} K_{1,1} & \dots & K_{1,j} \\ \vdots & & \vdots \\ K_{i,1} & \dots & K_{i,j} \end{bmatrix}$$

در این فایل هر ستون مربوط به یک کلمه‌ی خاص و هر سطر مربوط به یک فایل خبری خاص است. بنابراین مقدار موجود در درایه‌ی $K_{i,j}$ ، مشخص کننده‌ی حضور یا عدم حضور کلمه‌ی j اُم در فایل i اُم است. مقادیر موجود در این فایل با کاما از هم جدا شده اند.