درس کاوش دادگان انبوه

مدرس: دکتر سامان هراتی‌زاده

نیم‌سال اول ۹۹-۱۳۹۸

تمرین شماره‌ی دو

مهلت ارسال تمرین :۱۳۹۸/۰۴/۱۱ تا ساعت ۲۳

تحویل حضوری: اعلام خواهد شد

در این تمرین به شما تعدادی سند داده شده است و هدف این است که سندهای شبیه به هم پیدا شود. برای این کار از مراحلی که در کتاب گفته شده باید پیروی کنید:

۱- تبدیل هر سند به مجموعه‌های شینگلی

۲- ساختن ماتریسی که مشخص کند هر سند دارای چه شینگل‌هاییست. (مقادیر hash شینگل‌ها را باید استفاده کنید.)

۳- ساختن ماتریس minhash signature با استفاده از ماتریس قبلی (در این مرحله برای ترتیب‌های مختلف باید از تابع استفاده کنید و نباید از جابه‌جایی کامل سطرهای ماتریس استفاده شود. برای درک بهتر بخش ۳،۳،۵ را از کتاب مطالعه فرمایید.)

۴- یافتن موارد کاندید شباهت با استفاده از توابع LSH

۵- چک کردن کاندیدها و بررسی مقدار شباهتشان

نکات:

الف) نوشتن گزارش برای تمرین الزامیست ولی سعی کنید در کد نیز از کامنت‌گذاری مناسب استفاده کنید.

ب) سعی کنید روش استفاده شده در تمرین (min hash signature) را به اختصار توضیح دهید.

ج) روش انجام تمرین (مثل اینکه از چه تعداد جایگشت برای ساختن signature استفاده کرده‌اید و چرا؟) به طور کامل توضیح دهید. برای هر متغیری که انتخاب کرده‌اید باید توضیح کامل داده شود.

د) تعداد دقیق false-positive و تعداد تقریبی false-negativeها را محاسبه کنید.

ه) توصیه می‌شود که دور موارد بیان شده در کلاس مرور شود.

و) مهلت تمرین تمدید نخواهد شد!!

گزارش کار:

گزارش کارهای خود را در قالب یک فایل PDF بنویسید. درصورتیکه از مرجعی برای نوشتن تمرین‌ها استفاده کرده‌اید، حتما در فایل گزارش به آن ارجاع دهید. فایل گزارش باید شامل نام، شماره دانشجویی و متن گزارش به تفکیک هر سؤال باشد. گزارش کار بخشی از نمره نهایی این تمرین خواهد بود. لذا در نگارش آن نهایت دقت را داشته باشید.

نحوه ارائه:

حتما به تمام بخش‌های تمرین خود مسلط باشید. در زمان ارائه ممکن است از شما توضیحاتی تکمیلی در مورد تمرین و یا تغییراتی در کد خواسته شود. نمره‌ی اختصاص داده شده برای افرادی که مسلط به کار خود باشند قابل تغییر است. در زمان مشخص شده جهت ارائه حتما حضور داشته باشید. غیبت در روز ارائه به منزله عدم ارسال تمرین است و نمره‌ای به آن تعلق نخواهد گرفت.

نکات تکمیلی:

در صورت تأخیر تا ۲۴ ساعت، ۲۵٪ و تا ۴۸ ساعت ۵۰٪ از نمره تمرین کسر خواهد شد. به تمرین با تأخیر بیشتر از ۲ روز نمره‌ای تعلق نخواهد گرفت.

آدرس ایمیل جهت ارسال تمرین: adel.hessari@ut.ac.ir

فایل ارسالی یک فایل فشرده شده با قالب نامگذاری زیر باشد. (بجای NAME نام خود و به جای STDID شماره دانشجویی خود را قرار دهید.)

Hw2_NAME_STDID.zip

فایل فشرده شده باید شامل موارد زیر باشد:

۱- فایل‌های حاوی کدهای اجرایی – فایل متنی حاولی خروجی برنامه مربوط به تمرین اول و دوم.

۲- گزارش کار در قالب یک فایل PDF

موفق باشید. ☺