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)
ایمان برازنده (barazandeh.iman@ut.ac.ir)

مجموعه تمرین سوم : کاوش جریان

مهلت تحویل این تمرین جمعه ۵ دی ماه ساعت ۲۳:۵۹ می‌باشد.

۱-(طراحی و پیاده سازی الگوریتم)

دانشگاه تصمیم گرفته برای حمایت از دانشجویان در روزهای همه‌گیری ویروس، با همکاری چند اُپراتور تلفن همراه به تعداد محدودی از آنها ۱۰۰ گیگابایت اینترنت هدیه اختصاص بدهد. دانشجویان باید در روز معین و در بازه دو ساعتی تعیین شده به سامانه ثبت نام مراجعه و با درج کد ملی و شماره تلفن خود ثبت نام کنند.

سامانه باید قادر باشد تک‌تک کد ملی‌های ورودی را در پایگاه داده خود جستجو کرده و نتیجه را اعلام کند. برای پیاده سازی فرایند جستجو از **فیلتر بلوم**[1] استفاده شده است. دیتاست uniquemelli.csv شامل کدهای ملی دانشجویان در اختیار شما قرار گرفته است.

خبر رسیده است که تعدادی از دانشجویان دانشگاه‌های دیگر با اطلاع از اینکه سامانه از فیلتر بلوم استفاده می‌کند تصمیم گرفته‌اند که شانس خود را برای ثبت نام امتحان کنند.

الگوریتم جستجو و ثبت نام را با در نظر گرفتن نکات زیر پیاده سازی کنید:

الف) پیاده سازی به صورت کلاینت-سرور باشد. کلاینت جریان را به سرور برای پردازش ارسال می‌کند.

ب) بطور تصادفی ۱۰۰۰ کد ملی از دیتاست و ۱۰۰۰ کد ملی خارج از دیتاست در نظر بگیرید.

پ) الگوریتم را برای تعداد متفاوت تابع هَش، اندازه‌های متفاوت آرایه بیتی و بیشتر پیاده سازی، اجرا و مقایسه کنید. کد و نتایج را با جزئیات گزارش کنید. برای هرکدام حتما مشخص کنید که چه تعداد دانشجو خارج از دانشگاه موفق به ثبت نام شده‌اند.

ت) آیا همه دانشجویان دانشگاه تهران موفق به ثبت نام می‌شوند؟ پاسخ خود را با دلیل بیان کنید.

ث) الگوریتم را به شکلی طراحی کنید که حداقل 0.01 از دانشجویان دانشگاه‌های دیگر موفق به ثبت نام شوند.

[1] Bloom Filter

۲-(طراحی و پیاده سازی الگوریتم)

دیتاست این تمرین (digikala.xlsx) شامل اطلاعات سفارشات فروشگاه دیجی‌کالا در یکی از روزهای سال ۲۰۱۸ میلادی است. با استفاده از الگوریتم **Flajolet-Martin** تعداد استانهای منحصر به فرد در این دیتاست را تخمین بزنید.

همچنین دیتاست دیگری نیز (urls.csv) در اختیار شما قرار گرفته است. این دیتاست شامل urlهایی است که کاربران در یک موتور جستجو کلیک کرده‌اند. با استفاده از الگوریتم **Flajolet-Martin** تعداد urlهای منحصر به فرد در این دیتاست را تخمین بزنید.

برای هر دو دیتاست، الگوریتم را با در نظر گرفتن نکات زیر پیاده سازی کنید:

الف) پیاده سازی به صورت کلاینت-سرور باشد. کلاینت جریان را به سرور برای پردازش ارسال می‌کند.

ب) برای تعداد متفاوت تابع هَش نتایج را گزارش و تحلیل کنید.

پ) میانگین و میانه را به شکل‌های متفاوتی محاسبه کرده و مشاهدات خود را 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_HW# مثلا MMDS_830498001_ImanBarazandeh_HW1 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.