**دانشکده مهندسی برق**

# تمرین سری چهارم آنالیز داده های حجیم

**سال تحصیلی ۱۴۰۱-۱۴۰۰**

**سوال اول)**

در کتاب mining massive datasets در بخش ۳.۳.۵ الگوریتمی LSH به نوعی ذکر شده که یک permutation رندوم از سطرها انتخاب می‌شود. فرض کنید ستون n تایی با m تا مقدار ۱ داریم. فرض کنید فقط k مورد از n سطر را به صورت رندوم انتخاب می‌کنیم.

الف) در چه صورت نتیجه‌ی minhashing برابر با don't know می‌شود؟ برای ستون فوق الذکر ثابت کنید احتمال وقوع مقدار don't know برابر $(n-k/n)^m$ است.

ب) الگوریتم انتخابی در چه صورت ناموفق عمل می‌کند؟ فرض کنید میخواهیم احتمال وقوع don't know حداکثر برابر $e^{-۱۰}$ باشد. اعداد n و m را نیز اعدادی بزرگ در نظر بگیرید. به صورت تقریبی کوچکترین مقدار k را به نوعی بیابید که این واقعه رخ دهد.(راهنمایی. برای مقادیر بزرگ x داریم : $(1-\frac{1}{x})^x \approx e^{-1}$ )

**سوال دوم)**

در فرآیند پیدا کردن اثر انگشت به جای نگاه‌کردن به تصویر اثر انگشت به دنبال جزییاتی خاص (minutiae) در اثر انگشت میگردند. مثلا با تکه کردن تصویر اثر انگشت می‌توان تصویر را با وجود یا عدم وجود minutiae در هر تکه توصیف کرد. فرض کنید احتمال پیدا کردن اثر انگشت در یک تکه از شبکه‌ی عکس(فرض کنید مربع مربع تصویر شبکه شده‌است) ۲۰ درصد است. همچنین اگر در یک اثر انگشت در یک مربع یک minutiae یافت شود، در تصویر دیگر از این اثر انگشت به احتمال ۸۰ درصد این minutiae در این اثر انگشت یافت میشود.(یعنی دو تصویر از یک شی به احتمال ۸۰ درصد واحدهای مربعیشان نیز یکسان هست.). فرض کنید هر تصویر به ۱۰۰۰ مربع تقسیم شده است. همچنین فرض کنید یک خانواده از توابع بر حسب صفر یا یک بودن ۳ پیکسل خاص عکس(پیکسل‌ها برای هر تابع منحصر به فرد و ثابت است) آن را قبول یا رد میکنند.

الف) فرض کنید ۲۰۴۸ عضو از این خانواده به صورت رندوم به صورت OR کلی روی خروجی این توابع استفاده شود. احتمال false positive و false negative را بیابید. اگر این توابع را به دو گروه برابر تقسیم کنند و در هر گروه ابتدا and و سپس OR انجام دهیم احتمال‌ها به چه صورت می‌شوند.

ب) فرض کنید به جای ۲۰۴۸ تابع از n تابع استفاده کنید. به ازای چه مقدار n جمع مقادیر false positive و false negative کمینه می‌شود؟

**سوال سوم)**

هدف از این سوال شناسایی خودروهایی است که مسیر مشابه ای پیموده اند. . در سایت cw لینک درایو مجازی با عنوان **اطلاعات پروژه و تمرین درس** موجود است. توضیحات و نمونه کد برای پردازش اطلاعات در کولب نیز در این درایو وجود دارد. در این سوال می توانید از دیتاست Sample_Data.zip استفاده کنید. برای قسمت های ب و پ می توانید از rdd قسمت الف با نرخ ۰.۱ نمونه بگیرید.

الف) مسیر عبوری هر خودرو را به تفکیک روز در یک rdd مشخص کنید (همانند تمرین قبل). به عنوان مثال key= (Plate,Date) و value=[Device Code list].

ب) یک مسیر فرضی به صورت [Device Code list] فرض کنید و با محاسبه شباهت کسینوسی بین این مسیر و مسیر خودرو ها در rdd قسمت الف ، ۵ مسیر و خودرو با بیشترین شباهت را گزارش کنید.

پ) در این قسمت با استفاده از LSH قسمت ب را حل کنید. چند hyperplane فرض کنید و محاسبات مربوط به نحوه استفاده از آن ها را انجام دهید (and or). سپس مسیر های مشابه با مسیر فرضی را گزارش کنید. در صورت افزایش تعداد hyperplane ها دقت به چه صورت افزایش می یابد؟

**سوال چهارم)**

الگوریتم کلاسترینگ غیر اقلیدسی GRGPF را از نظر کاربرد و مراحل آن بررسی کنید.