به نام خدا

دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

درس پردازش داده‌های حجیم

استاد حقیرچهرقانی

تمرین دوم

علیرضا مازوچی

۴۰۰۱۳۱۰۷۵

بخش اول: سوالات تشریحی

سوال ۱

در جدول زیر فاصله دوبه‌دوی هر جفت کلمه آورده شده است:

	hen	she	he	then	when	مجموع
hen	۰	۲	۱	۱	۱	۵
she	۲	۰	۱	۳	۳	۹
he	۱	۱	۰	۲	۲	۶
then	۱	۳	۲	۰	۲	۸
when	۱	۳	۲	۲	۰	۸

الف) متناسب با مجموع فواصل کلمه hen مرکز این خوشه خواهد بود.

ب) کلمه she با فاصله ۲ تا کلمه hen دارای بیشترین فاصله تا خوشه است.

ج) بیشترین فاصله‌ای که بین جفت کلمات در جدول فواصل وجود دارد فاصله ۳ است؛ پس انسجام خوشه برابر با ۳ خواهد بود.

سوال ۲

۱. غلط؛ اولین بلاک شامل دو عدد یک است. اولین بلاک قطعا یک عدد یک دارد.
۲. غلط؛ چهارمین بلاک با صفر شروع شده است. بلاک‌ها با یک شروع می‌شوند.

۳. غلط؛ دومین بلاک با صفر شروع شده است. بلاک‌ها با یک شروع می‌شوند.
۴. صحیح

سوال ۳

الف) ۲۵۶؛ در قسمت ج مثالی ارائه شده است که در یک پنجره ۱۰۰۰ تایی بلاک‌های ۲۵۶ تایی ظاهر شده است. پس امکان پذیر است که بلاک‌های ۲۵۶ را مشاهده کنیم.

من ادعا می‌کنم امکان ندارد بلاک‌های بزرگتر از ۲۵۶ ظاهر شود و برای اثبات از برهان خلف استفاده می‌کنم. فرض کنید حداقل یک بلاک بزرگ‌تر از ۲۵۶ وجود داشته باشد. با توجه به آنکه از هر نوع بلاک ۱-تایی تا بزرگترین بلاک باید یک یا دو تای آن را داشته باشیم و از آنجایی که وجود دارد بلاکی که بزرگتر از ۲۵۶ باشد، پس حداقل یک بلاک ۱-تایی، حداقل یک بلاک ۲-تایی تا حداقل یک بلاک ۵۱۲-تایی خواهیم داشت. با این حساب حداقل $۱۰۲۳ = ۱ - ۲^{۱۰}$ عدد یک خواهیم داشت که طبیعتا در پنجره هزارتایی جا نمی‌شود. به تناقض می‌خوریم و حکم ثابت می‌شود.

با مثال و اثبات ارائه‌شده می‌توان نتیجه گرفت که بزرگترین بلاکی که امکان ظاهر شدن دارد ۲۵۶ است.

ب) ۲۵۶؛ امکان ندارد بزرگ‌ترین بلاک حداکثر ۱۲۸-تایی باشد. برای اثبات از برهان خلف کمک می‌گیرم. فرض کنید امکان‌پذیر باشد. پس در این شرایط حداکثر دو بلاک ۱-تایی، حداکثر دو بلاک ۲-تایی،... و حداکثر ۲ بلاک ۱۲۸-تایی خواهیم داشت که تعداد یک‌های موجود در بلاک‌ها در این حالت حداکثر برابر با $۵۱۰ = (۱ - ۲^{۸}) \times ۲$ خواهد بود. پس به تناقض می‌خوریم و حکم اثبات می‌شود. در نتیجه بزرگترین بلاک حداقل ۲۵۶-تایی است. باتوجه به این اثبات و نتیجه قسمت قبل، حتی می‌توان نتیجه گرفت که اندازه بزرگترین بلاک در این حالت دقیقا برابر با ۲۵۶ است.

ج) فرض کنید به ترتیب بلاک‌های زیر را داشته باشیم:

۱، ۱، ۲، ۴، ۸، ۸، ۱۶، ۳۲، ۳۲، ۶۴، ۶۴، ۱۲۸، ۱۲۸، ۲۵۶، ۲۵۶

سوال ۴

الف) نمودار ۳؛ مقدار false positive در فیلتر Bloom بر اساس تعداد تابع هش برابر است با $\left(1-e^{-\frac{km}{n}}\right)^k$. این تابع دارای یک نقطه مینیمم محلی است. پس باید نموداری را انتخاب کنیم که یک مینیم محلی در میانه‌ی آن داشته باشد که چنین چیزی تنها در نمودار ۳ دیده می‌شود.

به طور شهودی هم می‌توان حدس زد که نمودار باید دارای مینیمم محلی باشد. چراکه فرض کنید k برابر با یک باشد. در این صورت اگر مقدار هش یک ورودی با هش یکی از عناصر مجموعه S یکسان باشد به عنوان نمونه مثبت درنظر گرفته می‌شود. باتوجه به اینکه تنها یک شرط داریم، محتمل است false positive بالایی داشته باشیم ولی اگر مقدار k تعدادی کمی بیشتر باشد (بدون آنکه فضای B اشباع شود) می‌توان شرایط بیشتری را چک کرد و false positive کم می‌شود. از طرفی هم می‌دانیم اگر به صورت افراطی k را بسیار زیاد بگیریم، تقریبا به ازای تمام داده‌ها کل فضای B روشن می‌شود و تقریبا تمام داده‌ها مثبت تشخیص داده می‌شود که false positive را بسیار زیاد می‌کند. پس یک k نه خیلی کوچک و نه خیلی بزرگ کمترین false positive را خواهد داشت.

ب) نمودار ۵؛ مقدار false negative برای هر تعداد تابع هش برابر با صفر است. چراکه اگر یک ورودی برابر با یکی از اعضای مجموعه S باشد، در زمان تشکیل B به ازای آن و تمام توابع هش یک نقطه از B را برابر با یک قرار دادیم و تحت هیچ شرایط مقدار آن نقطه برابر با صفر نخواهد شد. حال موقع دیدن ورودی جدید به ازای هر تابع هش محل مورد نظر در B حداقل یک بار تبدیل به یک شده است. بدین ترتیب تمام شرایط برقرار خواهد بود و امکان ندارد این ورودی نمونه منفی شناخته شود.