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 باکت شامل دو عدد یک است. اولین باکت قطعا یک عدد یک دارد.
۲. غلط؛ چهارمین باکت با صفر شروع شده است. باکت‌ها با یک شروع می‌شوند.

۳. غلط؛ دومین باکت با صفر شروع شده است. باکت‌ها با یک شروع می‌شوند.
۴. صحیح

سوال ۳

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

من ادعا می‌کنم امکان ندارد باکت‌های بزرگتر از ۲۵۶ ظاهر شود و برای اثبات از برهان خلف استفاده می‌کنم. فرض کنید حداقل یک باکت بزرگ‌تر از ۲۵۶ وجود داشته باشد. با توجه به آنکه از هر نوع باکت ۱-تایی تا بزرگترین باکت باید یک یا دو تای آن را داشته باشیم و از آنجایی که وجود دارد باکتی که بزرگتر از ۲۵۶ باشد، پس حداقل یک باکت ۱-تایی، حداقل یک باکت ۲-تایی تا حداقل یک باکت 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 باکتی که امکان ظاهر شدن دارد ۲۵۶ است.

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

ج) فرض کنید به ترتیب باکت‌های 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.

بخش دوم: سوالات پیاده‌سازی

سوال ۱

الف و ب) در جدول زیر موارد خواسته‌شده آورده شده است:

بخش جریان	تعداد بیت‌های یک
۱۰۰۰ بیت آخر	۳۹۱
۵۰۰ بیت آخر	۲۰۱
۲۰۰ بیت آخر	۹۰

ج) ابتدا مجموعه‌داده ارائه‌شده در سوال را درنظر گرفتم. سه پارامتر را برای ارزیابی و تحلیل درنظر گرفتم: زمان مورد نیاز برای پردازش یک بیت از داده موقع خوانده جریان، زمان مورد نیاز برای پیش‌بینی یا شمارش تعداد بیت‌های یک موجود در پنجره در انتهای جریان و تعداد بیت پیش‌بینی‌شده یا شمارش‌شده. نتایج موجود در جدول زیر حاصل شد:

	زمان خواندن (µs)	زمان پیش‌بینی (ms)	تعداد بیت یک
DGIM	۲/۸۰۳	۰/۰۴۶	۵۰۸
دقیق	۰/۴۷۷	۰/۱۵۶	۳۹۱

همانطور که از نتایج بر می‌آید، DGIM نیاز به ۵/۸ برابر زمان برای خواندن یک بیت نیاز دارد ولی در زمان پیش‌بینی می‌تواند با ۰/۲۹ برابر زمان شمارش دقیق پیش‌بینی انجام دهد. پیش‌بینی انجام‌شده برای مجموعه‌داده موجود و برای لحظه آخر جریان نزدیک به ۳۰٪ خطا دارد.

باتوجه به اینکه جریان‌داده داده‌شده و اندازه پنجره هر دو کوچک هستند، من سه مجموعه‌داده دیگر با یک میلیون بیت و اندازه پنجره صد هزار در نظر گرفتم. بدین ترتیب می‌توان ارزیابی از میزان مقیاس‌پذیری DGIM هم داشته باشیم. یکی از این مجموعه‌ها دارای تعداد برابر صفر و یک است؛ یکی دارای تعداد بیت یک سه برابر تعداد بیت صفر و دیگری دارای تعداد بیت صفر سه برابر تعداد بیت یک. هر سه مجموعه‌داده

به صورت تصادفی ساخته شده است. مقدار هر سه پارامتر معرفی‌شده برای مجموعه‌داده اصلی را روی این سه مجموعه‌داده محاسبه کردم و نتایج آن در جدول زیر آورده شده است:

مجموعه داده	روش	زمان خواندن (µs)	زمان پیش‌بینی (ms)	تعداد بیت یک
عادی	DGIM	۳/۹۴۳	۰/۰۲۱	۵۷۲۶۵
	دقیق	۰/۴۴۸	۲۲۷/۰۰۱	۵۰۲۵۳
یک بیشتر	DGIM	۵/۲۷۹	۰/۰۲۲	۷۸۷۰۶
	دقیق	۰/۴۴۳	۲۳۳/۱۵۶	۷۵۰۶۹
صفر بیشتر	DGIM	۲/۰۲۸	۰/۰۱۹	۲۹۰۳۸
	دقیق	۰/۴۵۸	۲۳۳/۳۸۲	۲۴۶۵۳

برای سه مجموعه‌داده به ترتیب خطای ۱۴٪، ۵٪ و ۱۸٪ داشتیم که از خطای مجموعه‌اصلی کمتر است. تسریع پیش‌بینی به ترتیب ۱۰۸۰۹، ۱۰۵۹۸ و ۱۲۲۸۳ برابر بوده است. این مسئله نشان می‌دهد که الگوریتم DGIM برای داده‌های کلان می‌توان تسریع جدی‌ای در زمان پیش‌بینی داشته باشد. زمان خواندن DGIM به ترتیب ۸/۸، ۱۱/۹۱۶ و ۴/۴ برابر حالت دقیق بوده است. این موضوع نشان می‌دهد الگوریتم DGIM از این جنبه هم مقیاس‌پذیر است و باتوجه به پیش‌بینی سریعی که داشتیم به صرفه است تا در موقع خواندن داده پردازش بیشتری انجام گیرد. در عین حال و مطابق انتظار می‌بینیم DGIM برای جریان داده با تعداد بیت یک بیشتر به دلیل پردزاش بیشتر، کندتر بوده است.

سوال ۲

برای کاربر ۵۴۶۱ بازی‌های جدول زیر پیشنهاد می‌شود:

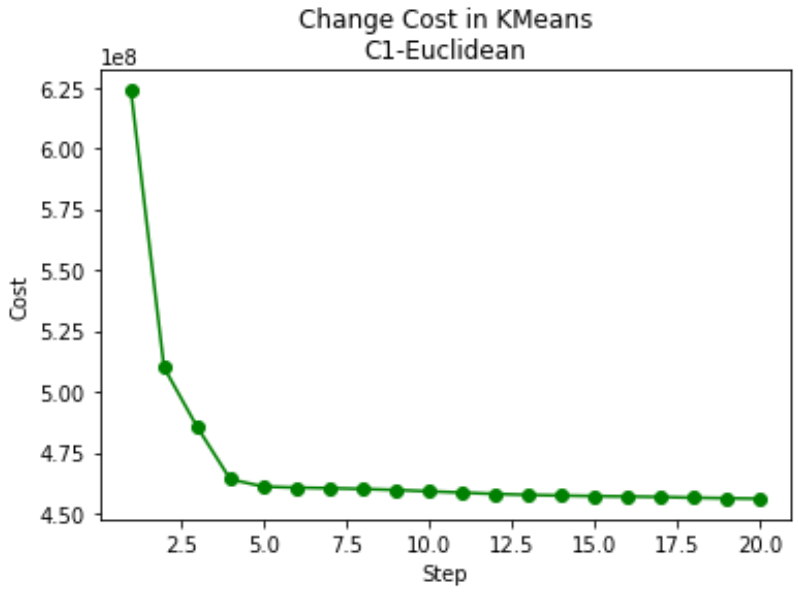
ردیف	آیدی بازی	نام بازی	امتیاز پیشنهادی
۱	۱۴۳	Forza Horizon 4	۵
۲	۵۳۲	DUSK	۵
۳	۸۰۳	NASCAR 2005: Chase for the Cup	۵
۴	۱۵۲۷	FTL: Faster Than Light	۵
۵	۱۹۶۷	The Book of Unwritten Tales	۵

برای کاربر ۱۰۱۴۰ بازی‌های جدول زیر پیشنهاد می‌شود:

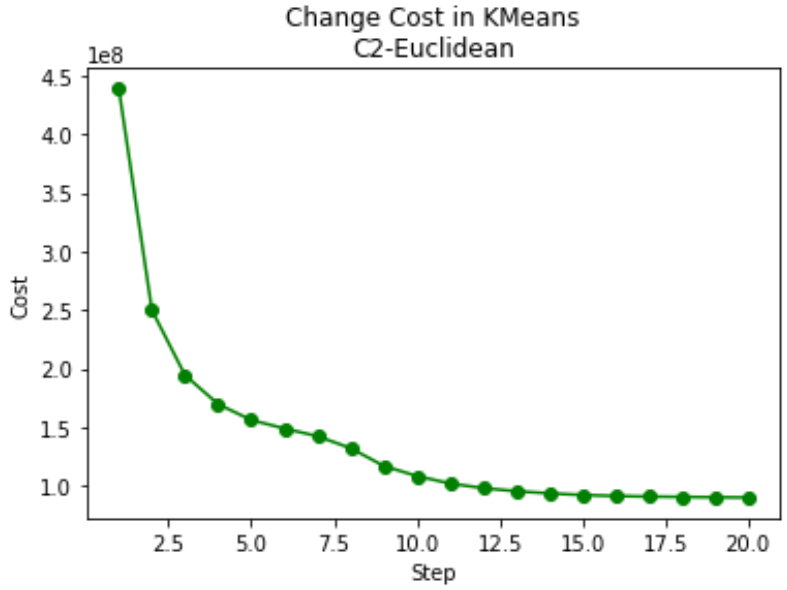
ردیف	آیدی بازی	نام بازی	امتیاز پیشنهادی
۱	۱۴۳	Forza Horizon 4	۵
۲	۵۳۲	DUSK	۵
۳	۸۰۳	NASCAR 2005: Chase for the Cup	۵
۴	۱۰۱۰	Katamari Damacy	۵
۵	۱۳۴۹	Shantae: Half-Genie Hero - Ultimate Edition	۵

سوال ۳

الف) برای مراکز اولیه C1 و با در نظرگرفتن فاصله اقلیدسی نمودار زیر حاصل می‌شود:



برای مراکز اولیه C2 و با درنظر گرفتن فاصله اقلیدسی نمودار زیر حاصل می‌شود:

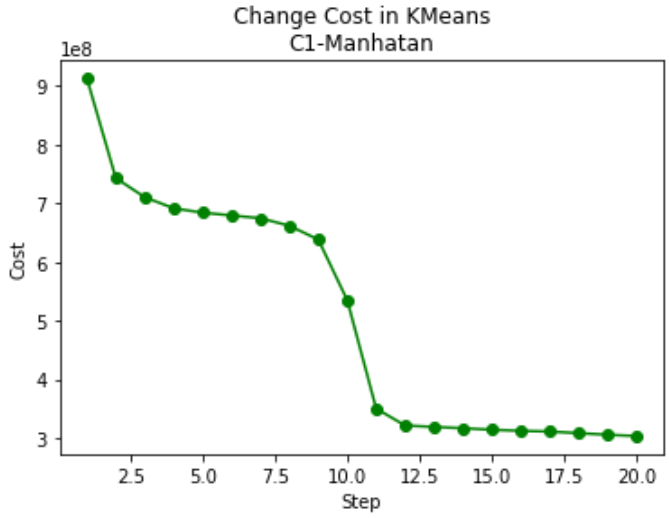


ب) در جدول زیر مورد خواسته‌شده آورده شده است:

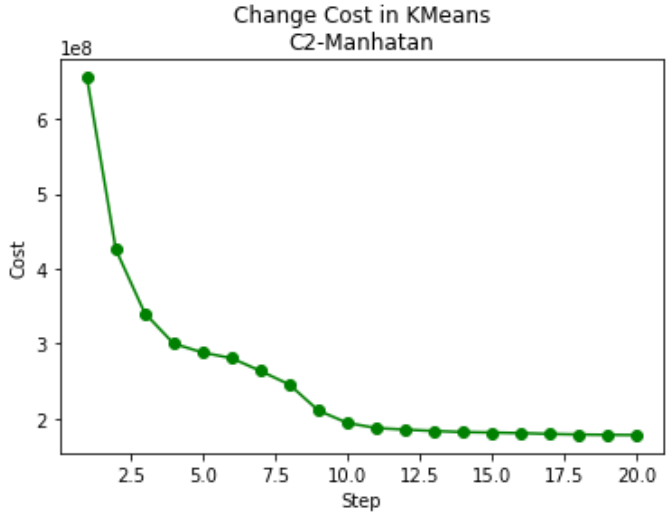
	مراکز اولیه C1	مراکز اولیه C2
درصد تغییر هزینه	۲۶٪	۷۶٪

با توجه به آنکه برای C2 درصد کاهش هزینه خیلی بیشتر از درصد کاهش هزینه C1 بوده است نشان می‌دهد که C2 مراکز اولیه را مکان مناسبی میان داده‌ها قرار داده است؛ جدای از این جدول و با بررسی نمودارهای قسمت الف می‌بینیم که هزینه گام اولیه C1 به قدری بالاست که در انتها تازه به هزینه ابتدای مراکز C2 می‌رسد که این مسئله هم نشان می‌دهد مراکز C2 مراکز بهتری هستند.

ج) برای مراکز اولیه C1 و با درنظر گرفتن فاصله منهتن نمودار زیر حاصل می‌شود:



برای مراکز اولیه C2 و با درنظر گرفتن فاصله منهتن نمودار زیر حاصل می‌شود:



د) در جدول زیر مورد خواسته‌شده آورده شده است:

	مراکز اولیه C1	مراکز اولیه C2
درصد تغییر هزینه	۶۱٪	۷۱٪

برای این فاصله هم تحلیل‌ها مشابه قسمت ب است و باز با دلایل مشابه متوجه می‌شویم که مراکز اولیه C2 مراکز بهتری نسبت C1 هستند. تنها تفاوت مهم این است که استفاده از فاصله منهتن توانسته است خطای زیاد اولیه C1 را بیشتر از فاصله اقلیدسی کاهش دهد. در نمودار هم مشخصا می‌بینیم که در گام ۹ مدل از مینیمم محلی خارج شده است و به جواب‌های خیلی بهتری رسیده است. چنین پدیده‌ای برای C2 هم رخ داده است ولی شدت تاثیر کمتر بوده است.