BIG DATA

Homework #2

MALIHE GHAZALIAN
830402056
Spring 2024

سوال 1)

فیلتر بلوم با تعداد 1000 کد ملی معتبر که از دیتاست به صورت رندم انتخاب شد و 1000 کد ملی نامعتبر که به غیر از این 1000 تای اول بود از دیتاست سمپل شدند سپس با طول بیت 10000 و تعداد 10 تابع هش، الگوریتم پیاده سازی شد و تعداد کد ملی های معتبر و نامعتبر به صورت زیر تخمین زده شد:

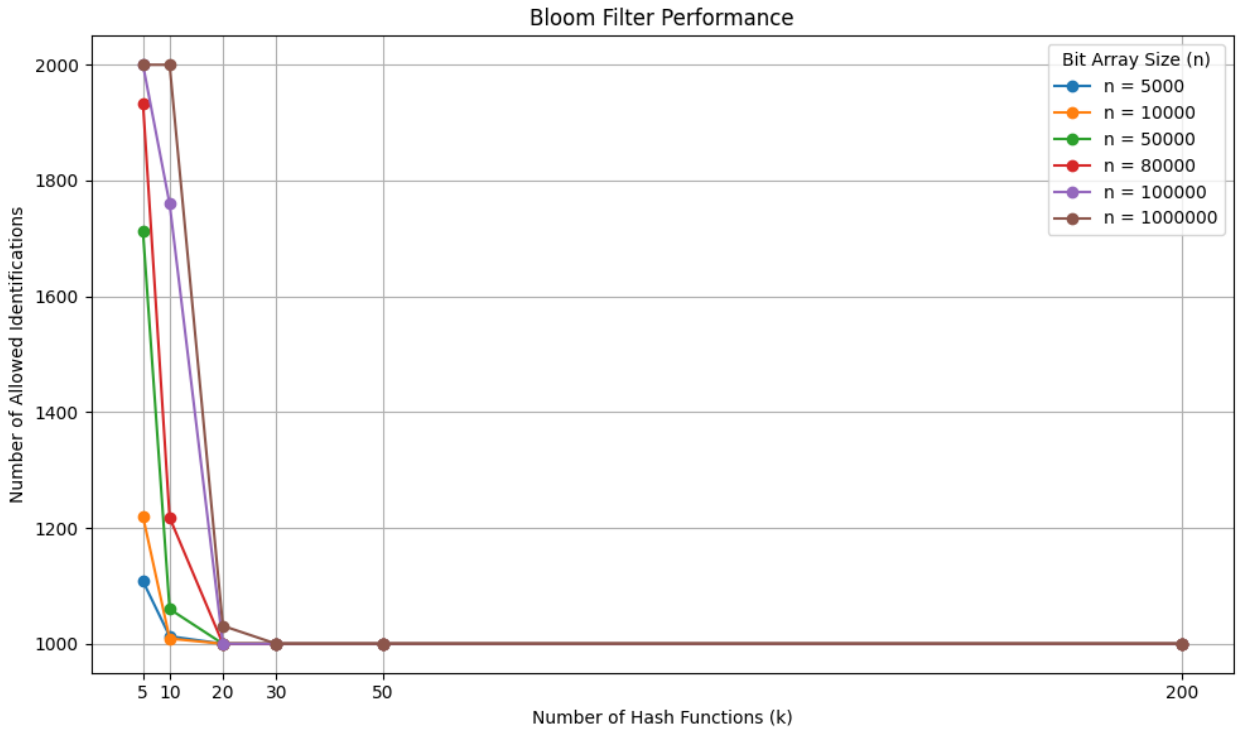
| Count of Allowed IDs: 1009

| Count of Not Allowed IDs: 991

در مرحله بعد با تعداد تابع هش متفاوت و طول های متفاوت بردار بیتی به شرح زیر الگوریتم اجرا شد :

```
num_hash_funcs = [5, 10, 20, 30, 50, 200]
bit_arr_sizes = [5000, 10000, 50000, 80000, 100000, 1000000]
```

نتایج تغییرات تعداد کدهای معتبر تشخیص داده شده به تعداد تابع هش در طول های متفاوت بردار بیتی به صورت نمودار زیر است:



بردار های بیتی در تعداد توابع کم تعداد false positive زیادی را نشان می دهد که بیان گر این است که استریم های جدید به خانه هایی هش می شوند که از قبل 1 بودند بدون آنکه عضو آن مجموعه

باشند ولی با افزایش تعداد هش و بیت های بردار بیتی این خطا کاهش می یابد و در تعداد 10 تا 20 هش و بردار بیتی با اندازه 5000 تا 10000 بهینه ترین جواب را می دهد.

اگر بخواهیم که احتمال fp برابر 0.01 باشد در فورمول احتمال یک بودن هر خانه که احتمال fp هم هست احتمال را قرار داده و از فرمول طول بیت را بدست می آوریم که فورمول به صورت زیر در می آیدو با قرار دادن احتمال و تعداد المان ها که 1000 تا است:

M = -nln(P)/(ln2)^2

جوابش می شود : 9585.74 حدود 10000 هزار طول بردار بیتی باشد

سپس از فورمول k را استخراج می کنیم که به صورت : k=n/m ln(2) در می آید و با جایگزاری m می شود: 6.91 یعنی حدود 7 تابع هش نیاز است تا احتمال fp بشود 0.01

سپس با مقادیر فوق یکبار دیگر الگوریتم را اجرا کرده که به جواب زیر می رسیم که همان حدود 0.01 fp را می دهد:

Count of Allowed IDs: 1008 |

Count of Not Allowed IDs: 992 |

سوال 2)

برای اینکه 35 تخمین داشته باشیم 35 تابع هش میخواهیم و با تخمین های مختلف بر اساس نتایج که به ازای هر استریم ورودی ثبت می شود روش ترکیبی نتایج به واقعیت نزدیک تری می دهد مثلا برای استریم اول روش ترکیبی 1 می شود ولی برای دو روش میانه و میانگین اعداد متفاوت است مخصوصا برای میانگین که کمی پرت می شود.

در قسمت بعد با طول های مختلف و با تعداد توابع مختلف الگوریتم اجرا شد و میانگین مقادیر پرتی را در استریم های بالاتر نشان می دهد همچنین با افزایش مقادیر میانگین بیشترین تغییرات را دارد.

سوال 3)

الف)

به این علت که بعضی از آیتم ها ممکن است در خیلی از بسکت ها ظاهر شوند و دلیل اصلی ظاهر شدن آیتم دوم آیتم اول نباشد به طوری که آیتم دوم مستقل از آیتم اول در بسیاری از بسکت ها ظاهر می شود بنابرین باید یک طوری این وابستگی را اثر دهیم یعنی تعداد بسکت هایی که آیتم دوم در آن ظاهر شده اگر آیتم های گروه ا در آن وجود دارند تقریبا برابر باشد بنابرین باید اثر ظاهر شدن کلی آیتمی که نتیجه می شود نیز ظاهر شود در دو متریک دیگر به دلیل اینکه در lift داریم مقدار کانفیدنس را به ساپورت آیتم دوم که به طور مستقل ظاهر می شود تقسیم میکنیم پس اگر آیتم دوم خیلی ظاهر شده باشد به طور مستقل مقدار لیفت کم می شود در conv نیز به طور معکوس عمل می کنیم یعنی اگر تعداد باری که دومی ظاهر می شود کم باشد صورت زیاد می شود و مخرج هم که میتوانیم ثابت فرض کنیم زیرا میخواهیم اثر وجود آیتم دوم به طور مستقل را بررسی کنیم پس با ثابت بودن کانفیدنس اگر آیتم دوم کمتر به طور مستقل ظاهر شود مخرج بزرگ شده و مقدار conv بیشتر می شود.

ب)

① confidence $\rightarrow (A \rightarrow B) = Pr(B|A) = \dfrac{S(A,B)}{S(A)}$

$$\frac{\text{support}(A,B)}{\text{support}(A)}$$

② lift $\rightarrow (A \rightarrow B) = \dfrac{\text{conf}(A \rightarrow B)}{S(B)} = \dfrac{S(A,B) \times N}{S(A)\,S(B)}$

③ Conviction $\rightarrow (A \rightarrow B) = \dfrac{1 - S(B)}{1 - \text{conf}(A \rightarrow B)} = \dfrac{1 - \dfrac{\text{support}(B)}{N}}{1 - \dfrac{\text{support}(A,B)}{\text{support}(A)}}$

① $\text{conf}(A \rightarrow B) = \dfrac{\text{support}(A,B)}{\text{support}(A)}$

$\text{conf}(B \rightarrow A) = \dfrac{\text{support}(A,B)}{\text{support}(B)}$

چون صورت یکی است ولی مخرج
یکی در طرفین بیشتر در صورت مستقل
ظاهر شده باشد پس این
دو مقدار برابر نمی‌شود.
در نتیجه یک معیار متقارن نیست

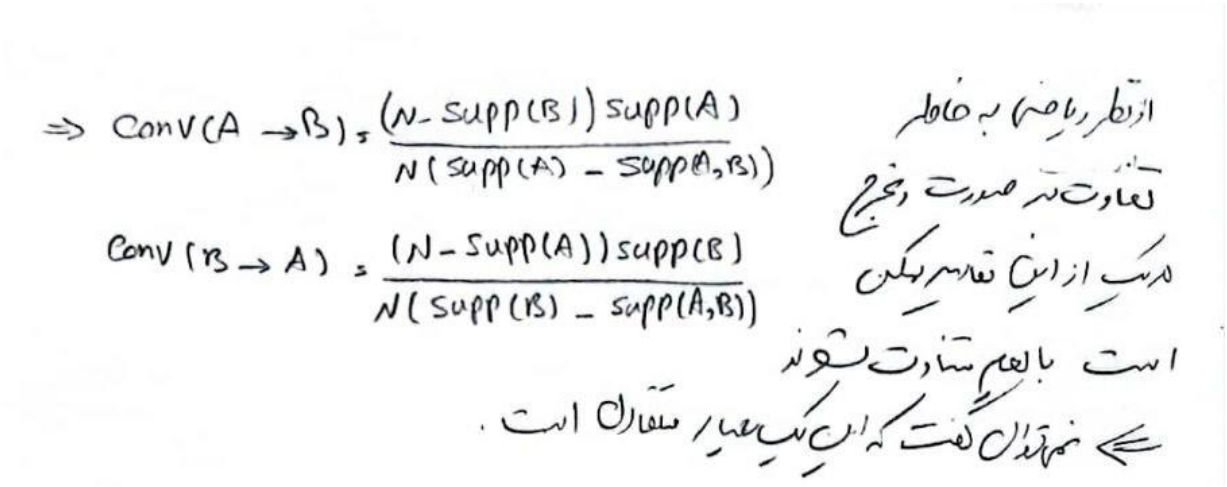
② $\text{lift}(A \rightarrow B) = \dfrac{\text{support}(A,B) \times N}{\text{support}(A)\,\text{support}(B)}$

$\text{lift}(B \rightarrow A) = \dfrac{\text{support}(A,B) \times N}{\text{support}(B) \times \text{support}(A)}$

که هر ۲ مقدار برابر است پس
این یک معیار دو طرفه است

③ $\text{conv}(A \rightarrow B) = \dfrac{1 - \dfrac{\text{support}(B)}{N}}{1 - \dfrac{\text{support}(A,B)}{\text{support}(A)}} = \dfrac{\dfrac{N - \text{supp}(B)}{N}}{\dfrac{\text{supp}(A) - \text{supp}(A,B)}{\text{supp}(A)}} = \dfrac{N - \text{supp}(B)}{}$

$\text{conv}(B \rightarrow A) = \dfrac{1 - \text{support}(A)}{1 - \dfrac{\text{support}(A,B)}{\text{support}\,B}} = \dfrac{\dfrac{N - \text{supp}(A)}{N}}{\dfrac{\text{supp}(B) - \text{supp}(A,B)}{\text{supp}(A)}}$



ج)

Confidence صورت دلالت کامل، مقدارش برابر 1 است، بنابراین Confidence به حداکثر مقدار خود می‌رسد.

Lift در صورت دلالت کامل مقدار صورت که کانفیدنس است به یک میرسد ولی چون وابسته به ساپورت آیتم دوم است بنابرین به حداکثر مقدار نمی رسد و مطلوب نیست.

Conviction در صورت دلالت کامل چون مخرج برابر 1-1 می شود مقدارش برابر بینهایت است ولی اگر بینهایت را مقدار حداکثر آن در نظر بگیریم پس به حداکثر خود می رسد و مطلوب است

در قسمت بعدی سوال نیز الگوریتم apriori طبق خواسته های سوال اجرا شده و قوانین انجمنی برای جفت ها و سه تایی ها محاسبه شده و 5 تای برتر شناسایی شدند:

پنج قانون برتر دوتایی :

[(('DAI93865', 'FRO40251'), 1.0), (('GRO85051', 'FRO40251'), 0.99918), (('GRO38636', 'FRO40251'), 0.99065), (('ELE12951', 'FRO40251'), 0.99057), (('DAI88079', 'FRO40251'), 0.98673)]

پنج قانون برتر سه تایی:

[(('ELE17451', 'DAI83733'), 'DAI62779'), 0.9279279279279279), ((('DAI85309', 'ELE92920'), 'DAI62779'), 0.8984375), ((('ELE59028', 'ELE92920'), 'DAI62779'), 0.874251497005988), ((('FRO85978', 'ELE92920'), 'DAI62779'), 0.9502487562189055), ((('GRO81087', 'ELE92920'), 'DAI62779'), 0.9571428571428572)]

سوال 4)

با همان دیتا الگوریتم های اصلاح شده ی a-priori اجرا شد که نتایج نشان می دهد که جوابهایی که الگوریتم ها به دست می دهند با هم متفاوت است مثلا در الگوریتم pcy با تغییر مقادیر سایز هش تعداد زوج ها تغییری میکند.