به نام خدا



تحلیل داده های حجیم

تمرین سری سوم

استاد: دکترغلامپور

دانشجو:سجاد هاشم بیکی(98107077)

پاییز1401

**سوال1:**

**الف)** اگر همه سطرهای انتخاب شده مقدار صفر داشته باشند مقدار هش نامشحص خواهد بود.

احتمال اینکه همه سطرهای انتخاب شده صفر باشد برابر بدین صورت است:

عبارات بالا همگی کوچک تر مساوی با میباشند بنابراین داریم:

**ب)**

***ج)***

|  |  |
| --- | --- |
| S2 | S1 |
| 0 | 0 |
| 1 | 1 |
| 0 | 1 |

*:*

با توجه به جایگشت تناوبی در یک سوم حالت ها مقدار هش یکسان میشود.که این مقدار مخالف مقدار similarity میباشد

**سوال2:**

**الف:**

با استفاده از نامساوی مارکوف داریم:

به عنوان مثال W1 را انتخاب میکنیم.

توابع H به صورت هستند بنابراین داریم:

با فرض استقلال هش ها داریم:

*اگر فاصله همه نقاط از z بیشتر از باشد،بنابراین n هستند که به احتمال 1/n به باکت یکسان میروند.نقاط آن باکت توزیع* binomial(n,1/n) *دارند. که امید ریاضی ان یک میباشد.بنابراین داریم:*

**ب:**

با توجه به توابع H داریم:

با فرض استقلال هش ها داریم:

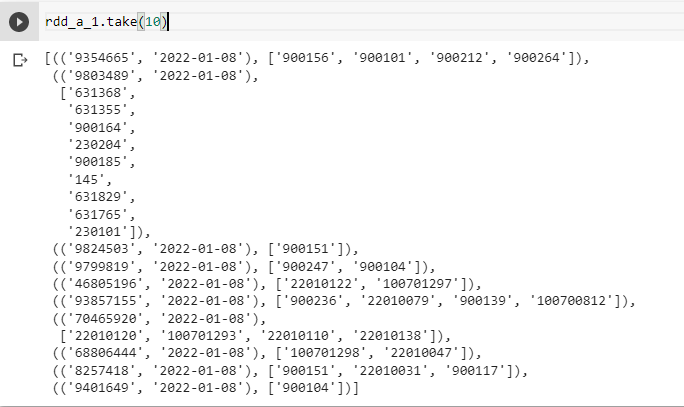
*با فرض استقلال* داریم:

*اگر مقدار n را بزرگ در نظر بگیریم داریم:*

**سوال3:**

الف:

با همان قالب گفته شده در سوال((key=(plate ,date),value = [Device Code List])) rdd ساخته شده بدین شکل میباشد:



ب:

دیتای sample:

یک مسیر فرضی انتخاب میکنیم.بدین صورت که چند دوربین را به صورت رندوم انتخاب کرده و آن را بردار مسیر فرضی در نطر میگیریم.

حال کسینوس زاویه میان دو بردار مسیر فرضی و مسیر خودروها را محاسبه میکنیم.

هر چه مقدار کسینوس به یک نزدیک تر باشد به معنای زاویه کمتر میان ان دو بردار و شباهت بیشتر آنهاست.

در اینجا یک مسیر فرضی شامل 4 دوربین را در نظر گرفتیم.

پنج مقدار بزرگتر برابر 0.5 و0.354 و0.289 و0.25 و 0.224 میباشد.

مولفه های هر زوج بدین صورت میباشد: (key =cos(theta), value = ((plate,date),[Device Code List]))

بیشترین شباهت ها به ازای پنج مقدار اول:

(**0.5**, (('11064269', '2022-01-08'), ['22010058'])

(**0.354**, (('12094907', '2022-01-08'), ['22010057', '22010058']))

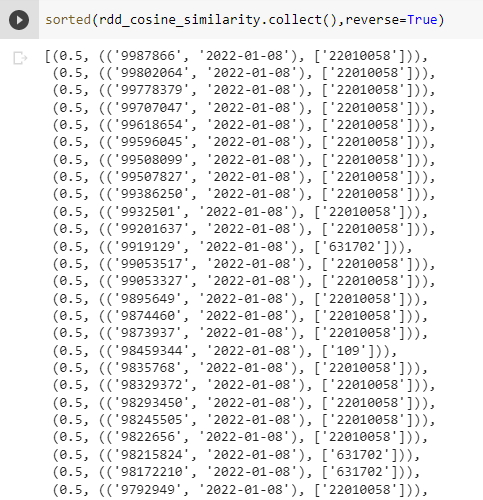
(**0.289**, (('10238525', '2022-01-08'), ['900155', '22010057', '22010058']))

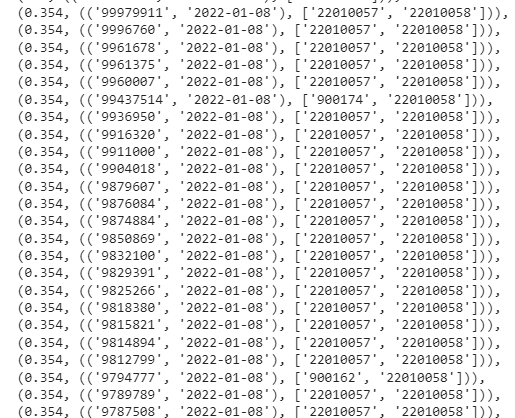
(**0.25**, (('68635642', '2022-01-08'), ['22010058', '22010057', '100701150', '100701145']))

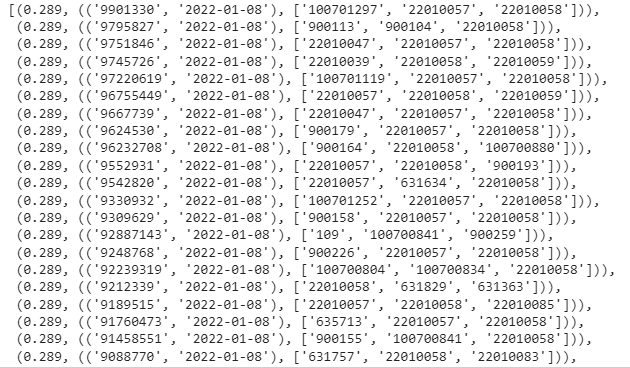
(**0.224**, (('69018599', '2022-01-08'), ['631633', '900273', '100701252', '22010058', '22010057']))

**مقادیر بالا به ازای مسیر فرضی**

**{['22010058'],['631702'],['22009911'],['109']} میباشد.**







**کل دیتا:**

نتایج این قسمت برای **کل دیتا** و مسیر فرضی (['22009806'], ['1001001'], ['900195'], ['100701298']]) بدین شکل میباشد:

