Bloom过滤器

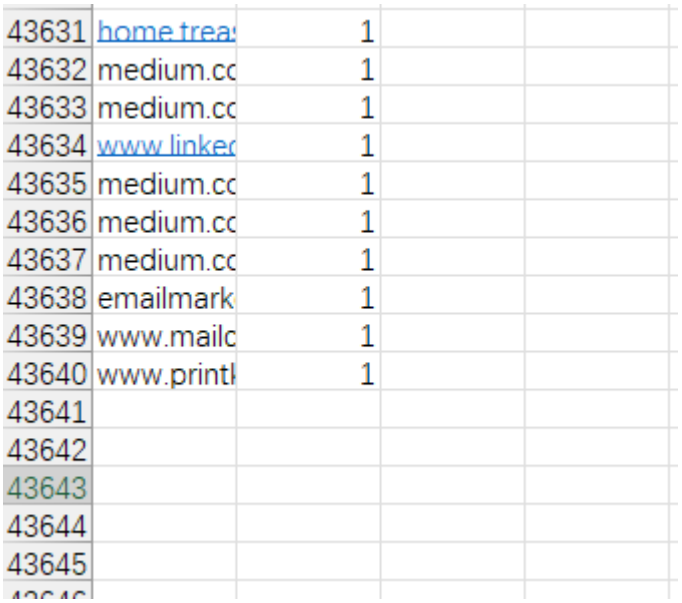
大致思路

python的hashlib标准库提供了MD5、SHA1、SHA256等密码学Hash函数

python的random库提供了产生随机数的方法

random产生的是伪随机数或者说是用一种复杂的方法计算得到的序列值，因此每次运算时需要一个不同的种子值。种子值不同，得到的序列值也不同。而种子值相同，序列值就相同。也就是说，random可以完成hash值到指定区间整数值的映射

确定参数



43631	home.trea	1
43632	medium.cc	1
43633	medium.cc	1
43634	www.linkec	1
43635	medium.cc	1
43636	medium.cc	1
43637	medium.cc	1
43638	emailmark	1
43639	www.mailc	1
43640	www.printl	1
43641		
43642		
43643		
43644		
43645		

查阅csv文件，其中的数据量为n=43639个

看样子，csv文件里的1代表无害，-1代表有害

根据公式 $k = \frac{m}{n} \ln 2$,我希望使用hash函数的个数是3-4个，那么倒推出m的值应该是200000附近。

先初始化m=200000，如果错误率过大，再调整参数

代码

由hash值映射到指定区间整数

```
def func(Length, HashValue):
    random.seed(HashValue)
    return random.randint(0, Length-1)
```

Bloom类

```
class Bloom:
    def __init__(self, length_of_bloom):
        self.vector = []
        self.length = length_of_bloom
        self.hash_algorithm_list = ['md5', 'sha1', 'sha256']
        for i in range(0, length_of_bloom):
            self.vector.append(0)
```

Bloom类拥有一个列表做成的向量vector，一个存放着此过滤器使用了哪些hash算法的列表

输入length后，vector会被初始化为有length个0

Bloom对象的向量初值设置

```
    def SetValue(self, csv_filename):
        for algorithm in self.hash_algorithm_list:
            self.SetValueWithCertainHashAlgorithm(csv_filename=csv_filename,
algorithm=algorithm)

    def SetValueWithCertainHashAlgorithm(self, csv_filename, algorithm):
        csvfile = open(csv_filename, mode='r')
        lines = csv.reader(csvfile)

        for line in lines:
            url, isMalicious = line
            if isMalicious == '-1':
                Hash = hashlib.new(name=algorithm, data=url.encode(encoding='utf-
8'))
                HashValue = Hash.hexdigest()
                index = func(Length=self.length, HashValue=HashValue)
                self.vector[index] = 1
```

两个都是Bloom类的成员函数

验证

```
    def filter(self, url):#接受一个url，返回0代表通过，1代表拦截
        for name in self.hash_algorithm_list:
            Hash = hashlib.new(name=name, data=url.encode(encoding='utf-8'))
            HashValue = Hash.hexdigest()
            index = func(Length=self.length, HashValue=HashValue)
            if self.vector[index] == 0:
                return 0
        return 1
```

再加一个verify函数做接口，对csv文件里的所有地址都验证一遍

```
    def verify(self,csv_filename):
        harmless_counter = 0#无害的总数
        harmful_counter = 0#有害的总数
        error_in_harmless = 0#无害的却过滤了的数量
        error_in_harmful = 0#有害但没过滤的数量


        csvfile = open(csv_filename, mode='r')
        lines = csv.reader(csvfile)

        for line in lines:
            url, isMalicious = line
            if isMalicious == '1':#无害
                harmless_counter += 1
                if self.filter(url) == 1:
                    error_in_harmless += 1
            elif isMalicious == '-1':#有害
                harmful_counter += 1
                if self.filter(url) == 0:
                    error_in_harmful += 1

        print('无害总数:', harmless_counter, ' 发生错误数:', error_in_harmless, '错误
率:', error_in_harmless / harmless_counter)
        print('有害总数:', harmful_counter, ' 发生错误数:', error_in_harmful, '错误
率:', error_in_harmful / harmful_counter)
```

运行结果

```
length = 200000
无害总数: 33312  发生错误数: 84 错误率: 0.002521613832853026
有害总数: 10327  发生错误数: 0 错误率: 0.0
```

调整参数

公式 $k = \frac{m}{n} \ln 2$,描述的是m和n先确定，再算出最优的k

n是不变的，恒为43639

我想当然地分析一下，

m一定时 应该是k越大，错误率越小，

k一定时 应该是m越大，错误率越小，

之所以有这个公式，是要在开销和错误之间找一个平衡

控制k=3不变，改变m

理论上来说，这样应该是m越大，错误率就越小。

恰好为公式值的情况

由理论可知，k取3，n取n=43639时，m的值是187028

```
length = 187028
无害总数: 33312  发生错误数: 110 错误率: 0.003302113352545629
有害总数: 10327  发生错误数: 0 错误率: 0.0
```

远小于公式值的情况

```
length = 10000
无害总数: 33312  发生错误数: 28592 错误率: 0.8583093179634966
有害总数: 10327  发生错误数: 0 错误率: 0.0
```

错误率极高

远大于公式值的情况

```
length = 400000
无害总数: 33312  发生错误数: 11 错误率: 0.0003302113352545629
有害总数: 10327  发生错误数: 0 错误率: 0.0

length = 2000000
无害总数: 33312  发生错误数: 0 错误率: 0.0
有害总数: 10327  发生错误数: 0 错误率: 0.0

length = 20000000
无害总数: 33312  发生错误数: 0 错误率: 0.0
有害总数: 10327  发生错误数: 0 错误率: 0.0
```

当m取足够大，已经一个错误都不发生了

控制m不变，改变k

```
self.hash_algorithm_list = ['md5', 'sha1', 'sha256']

因此，只需要在这个列表里增删就好了
```

m保持200000，k=4

```
self.hash_algorithm_list = ['md5', 'sha1', 'sha256', 'sha384']

length = 200000
无害总数: 33312  发生错误数: 36 错误率: 0.001080691642651297
有害总数: 10327  发生错误数: 0 错误率: 0.0
```

错误率降低了

m保持200000，k=5

```
self.hash_algorithm_list = ['md5', 'sha1', 'sha256', 'sha384', 'sha512']

length = 200000
无害总数: 33312  发生错误数: 21 错误率: 0.0006304034582132565
有害总数: 10327  发生错误数: 0 错误率: 0.0
```

错误率还在降低

m保持200000，k=9

```
self.hash_algorithm_list = ['md5', 'sha1', 'sha224', 'sha256', 'sha384',
'sha512', 'blake2b', 'blake2s', 'sha3_224']

length = 200000
无害总数: 33312  发生错误数: 2 错误率: 6.003842459173871e-05
有害总数: 10327  发生错误数: 0 错误率: 0.0
```