```
!pip install bitarray
!pip install mmh3

!wget https://bit.ly/3umrYIV -O positive-words.txt
!wget https://bit.ly/39HgJmk -O negative-words.txt
```

```
f = open("positive-words.txt")
S = f.readlines()
S = [x.strip() for x in S if not x.startswith(';')]
print(S)

f = open("negative-words.txt")
T = f.readlines()
T = [x.strip() for x in T if not x.startswith(';')]
print(T)
```

```
['', 'a+', 'abound', 'abounds', 'abundance', 'abundant', 'accessable', 'accessible', 'ac
['', '2-faced', '2-faces', 'abnormal', 'abolish', 'abominable', 'abominably', 'abominate
```

```
#insert

import math
import mmh3
from bitarray import bitarray

m = len(S)

n = 8*m

k = 5

B = bitarray(n)
B.setall(False)

for s in S:
  for i in range(k):
    # with differet seed, hash function is different
    h_i_s = mmh3.hash(s, i) % n
    B[h_i_s] = 1

print("Fraction of 1's in B", B.count()/n)
```

```
Fraction of 1's in B 0.4624439461883408
```

```
# lookup (filtering the stream)

import random
```

```
4
5 S_sample = random.sample(S, 100)
6 T_sample = random.sample(T, 100)  # all these words aren't in S
7
8 print(S_sample)
9 print(T_sample)
10
11 # Let's check these words against the Bloom filter
12 cnt_not_in = 0
13 for x in T_sample:
14   not_in = False
15
16   for i in range(k):
17     h_i_x = mmh3.hash(x, i) % n
18     if B[h_i_x] == 0:
19       not_in = True
20       cnt_not_in += 1
21       break
22
23 # how many elements did not pass the filter
24 print('We discarded', cnt_not_in)
```

```
['rejoicingly', 'dreamland', 'refunded', 'envy', 'enthuse', 'precise', 'sumptuously', 'c
['maladjustment', 'infamously', 'shun', 'gossip', 'hopelessness', 'cramping', 'dastard',
We discarded 95
```

✓ 0s completed at 6:51 PM