Resevoir Sampling

Problem: Sampling

Lines from a large text file

Sample search engine queries, updated live

The Simple Way

1. Scan the text file, counting lines
2. Generate random line numbers $[0, |\text{lines}|)$
3. Sort the line numbers
4. Scan the text file, outputting selected lines

The Simple Way

1. Scan the text file, counting lines
2. Generate random line numbers $[0, |\text{lines}|)$
3. Sort the line numbers
4. Scan the text file, outputting selected lines

Cost: two scans

Sample One Line

Flip a coin at each line.
If it's heads, record the line (and forget the others).

Sample One Line

Flip a coin at each line.
If it's heads, record the line (and forget the others).

```python
#!/usr/bin/env python
import sys
import random
resevoir = sys.stdin.readline().strip()
for line in sys.stdin:
  if random.randint(0,1) == 0:
    resevoir = line.strip()
print(resevoir)
```

This is biased. The last line has probability 0.5.

Sample One Line

Flip a coin at each line.
If it's heads, record the line (and forget the others).

```
#!/usr/bin/env python
import sys
import random
resevoir = sys.stdin.readline().strip()
for line in sys.stdin:
  if random.randint(0,1) == 0:
    resevoir = line.strip()
print(resevoir)
```

This is biased. The last line has probability 0.5.
It should be $\frac{1}{|lines|}$.

Uniformly Sample One Line

```python
#!/usr/bin/env python
import sys
import random
line_number = 0
for line in sys.stdin:
  if random.randint(0, line_number) == 0:
    resevoir = line.strip()
  line_number += 1
print(resevoir)
```

Line n overwrites the resevoir with probability $\frac{1}{n}$

$\implies$ Uniform sampling

Proof Sketch: Induction

Base One line with probability 1.

Inductive Assume n lines were sampled with probability $\frac{1}{n}$ each. When the $n+1$th line is added, the resevoir is kept with probability $\frac{n}{n+1}$. Thus the first n lines each have probability

$$\frac{1}{n} \cdot \frac{n}{n+1} = \frac{1}{n+1}$$

And the $n+1$th line also has probability $\frac{1}{n+1}$ by construction.

Sample Multiple Lines
Without Replacement

First few lines: Fill the resevoir

Afterwards: Substitute an entry with probability $\frac{|\text{samples}|}{|\text{lines}|}$

Summary

Efficiently sample streaming data
Small memory