Compression from scratch

Purdue Hackers
Kian Kasad

Copyright © 2025 Kian Kasad

Grab these slides!

https://puhack.horse/compression-workshop

Copyright © 2025 Kian Kasad

Before we start...

Join the Purdue Hackers Discord!

https://puhack.horse/discord

Open the Python notebook

https://puhack.horse/compression-py

Copyright © 2025 Kian Kasad

The plan

- 5 minutes on RLE theory
- 15 minutes to build RLE
- 15 minutes on Huffman theory
- 15 minutes to build Huffman encoding
- 30 minutes for you to explore!
  - Several options:
    - Implement Huffman decoding
    - Learn & implement Lempel-Ziv (LZ77)
    - Combine Huffman & LZ77 to get DEFLATE

Copyright © 2025 Kian Kasad

The plan

- 5 minutes on RLE theory
- 15 minutes to build RLE
- 15 minutes on Huffman theory
- 15 minutes on Huffman implementation
- 30 minutes for you to explore!
  - Several options:
    - Implement Huffman decoding
    - Learn & implement Lempel-Ziv (LZ77)
    - Combine Huffman & LZ77 to get DEFLATE

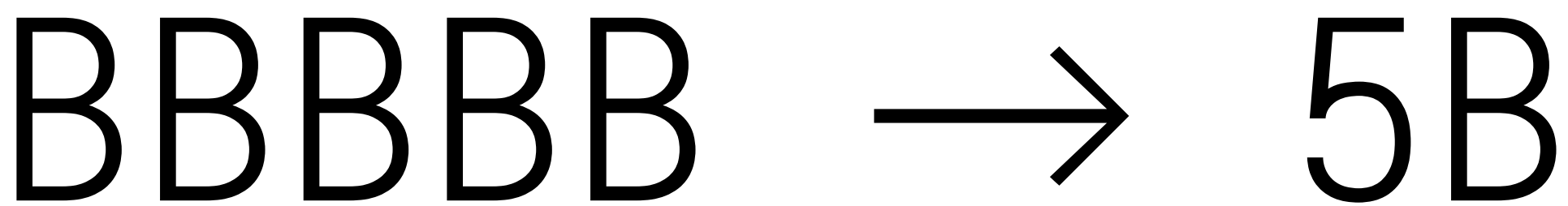
Copyright © 2025 Kian Kasad

Run-length encoding (RLE)

B B B B B

Copyright © 2025 Kian Kasad

Run-length encoding (RLE)



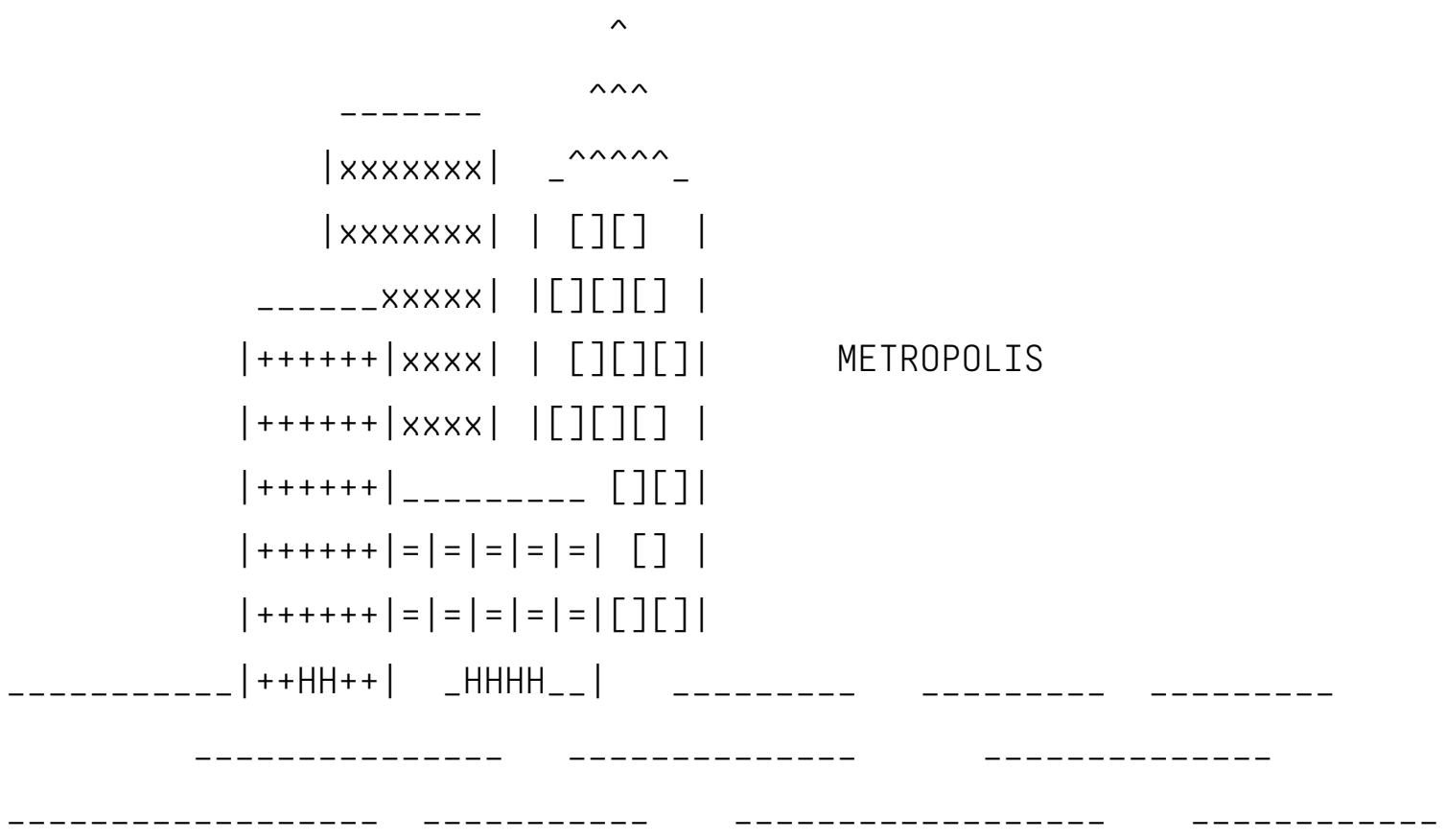
Copyright © 2025 Kian Kasad

Run-length encoding (RLE)

Woooohooooooo! → W4oh6o!

Copyright © 2025 Kian Kasad

Run-length encoding (RLE)



Copyright © 2025 Kian Kasad

Problem: overlapping alphabets

- “4o” could mean “oooo”
- Or could mean a literal “4o”
- How do we decide?

Copyright © 2025 Kian Kasad

Questions?

Copyright © 2025 Kian Kasad

The plan

- 5 minutes on RLE theory
- 15 minutes to build RLE
- 15 minutes on Huffman theory
- 15 minutes on Huffman implementation
- 30 minutes for you to explore!
  - Several options:
    - Implement Huffman decoding
    - Learn & implement Lempel-Ziv (LZ77)
    - Combine Huffman & LZ77 to get DEFLATE

Copyright © 2025 Kian Kasad

The plan

- 5 minutes on RLE theory
- 15 minutes to build RLE
- **15 minutes on Huffman theory**
- 15 minutes on Huffman implementation
- 30 minutes for you to explore!
  - Several options:
    - Implement Huffman decoding
    - Learn & implement Lempel-Ziv (LZ77)
    - Combine Huffman & LZ77 to get DEFLATE

Copyright © 2025 Kian Kasad

Huffman coding

- Plain ASCII text: each character = 1 byte = 8 bits
- Some characters are common; some aren’t
- What if we use fewer bits for common characters?

Copyright © 2025 Kian Kasad

Huffman coding

- Main idea:

  Frequent characters get fewer bits.

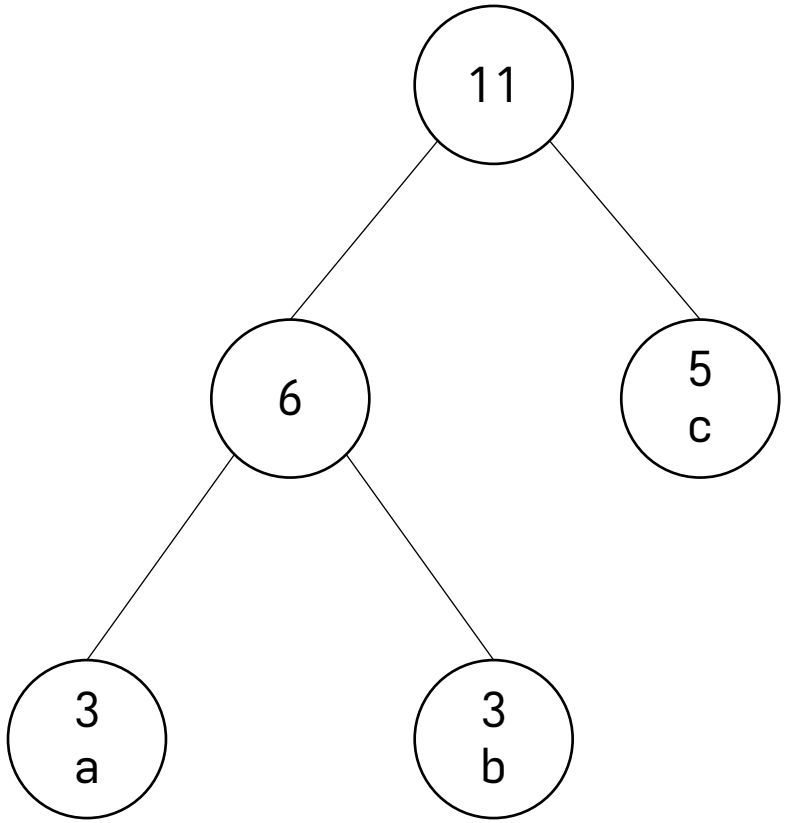
  Infrequent characters get more bits.

Copyright © 2025 Kian Kasad

How to do this?

Build a tree!

aaabbbccccc



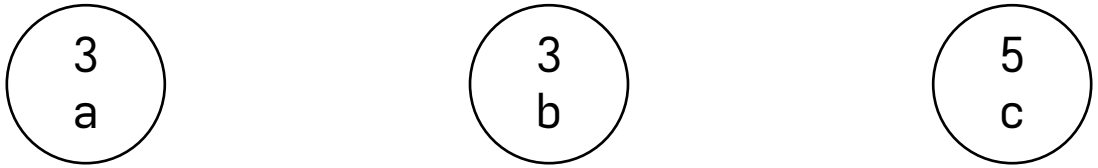
Copyright © 2025 Kian Kasad

Questions?

Copyright © 2025 Kian Kasad

How to do this?

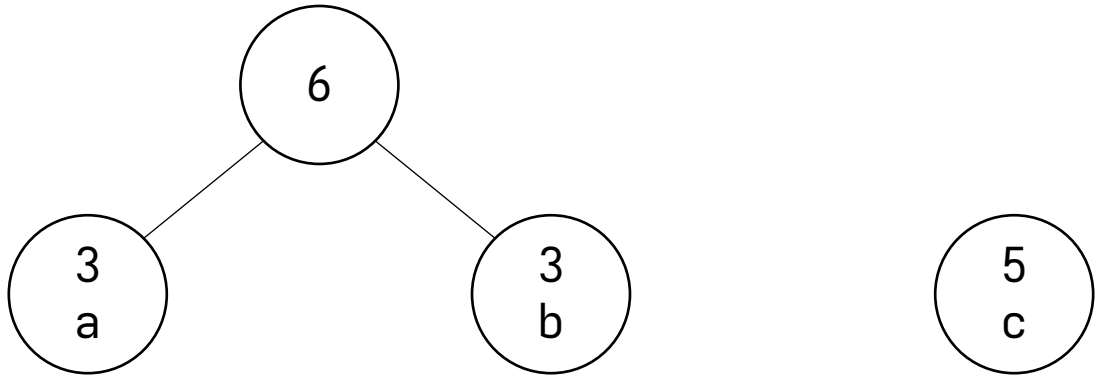
aaabbbccccc



Copyright © 2025 Kian Kasad

How to do this?

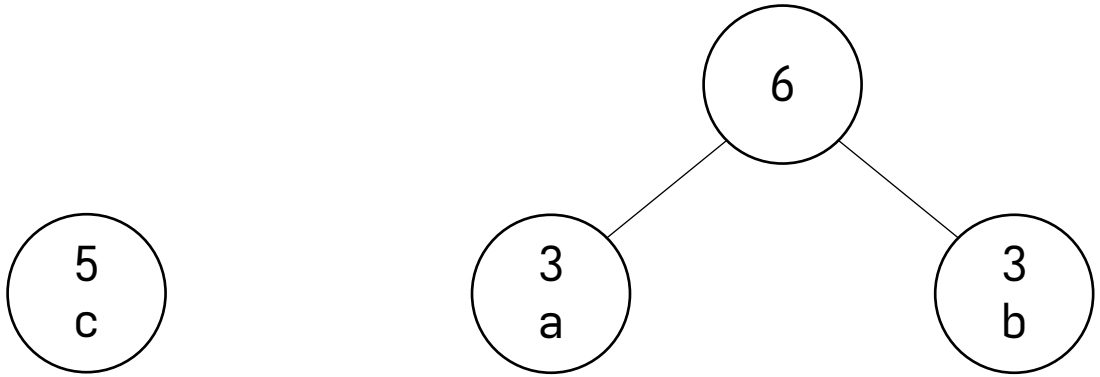
aaabbbccccc



Copyright © 2025 Kian Kasad

How to do this?

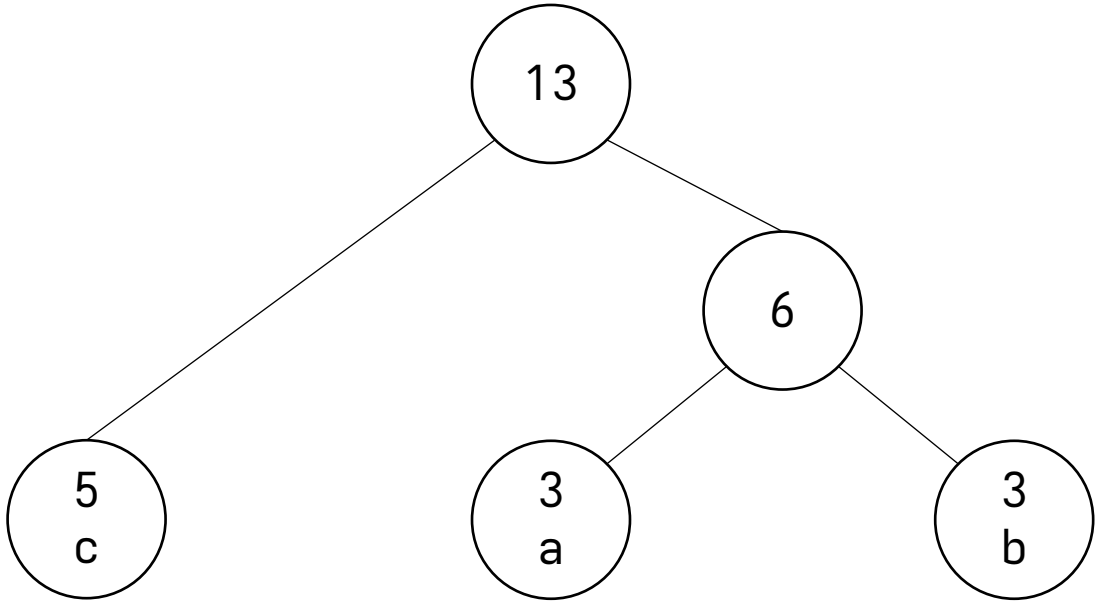
aaabbbccccc



Copyright © 2025 Kian Kasad

How to do this?

aaabbbccccc



Copyright © 2025 Kian Kasad

Questions?

Copyright © 2025 Kian Kasad

The plan

- 5 minutes on RLE theory
- 15 minutes to build RLE
- 15 minutes on Huffman theory
- **15 minutes on Huffman implementation**
- 30 minutes for you to explore!
  - Several options:
    - Implement Huffman decoding
    - Learn & implement Lempel-Ziv (LZ77)
    - Combine Huffman & LZ77 to get DEFLATE

Copyright © 2025 Kian Kasad

The plan

- 5 minutes on RLE theory
- 15 minutes to build RLE
- 15 minutes on Huffman theory
- 15 minutes on Huffman implementation
- **30 minutes for you to explore!**
  - Several options:
    - Implement Huffman decoding
    - Learn & implement Lempel-Ziv (LZ77)
    - Combine Huffman & LZ77 to get DEFLATE

Copyright © 2025 Kian Kasad

Lempel-Ziv (LZ77)

- Introduced by Lempel and Ziv in 1977
- RLE only removes immediate repetition
- What about this text:

    I like tea, I like cheese, and I like sand.

Copyright © 2025 Kian Kasad

Lempel-Ziv (LZ77)

I do not like them in a house.
I do not like them with a mouse.
I do not like them here or there.
I do not like them anywhere.

Copyright © 2025 Kian Kasad

Lempel-Ziv (LZ77)

I do not like them in a house.
I do not like them with a mouse.
I do not like them here or there.
I do not like them anywhere.

Copyright © 2025 Kian Kasad

Lempel-Ziv (LZ77)

I do not like them in a house.
<31,19>with a mouse.
<34,20>here or there.
<35,20>anywhere.

Copyright © 2025 Kian Kasad

Compression strategy

arbitrarytextthistexthasnomeaning

Copyright © 2025 Kian Kasad

Compression strategy

arbitrarytextthistexthasnomeaning

Copyright © 2025 Kian Kasad

Compression strategy

arbitrarytextthistexthasnomeaning

Copyright © 2025 Kian Kasad

Compression strategy

arbitrarytextthistexthasnomeaning

Copyright © 2025 Kian Kasad

Compression strategy

arbitrarytextthistexthasnomeaning

Copyright © 2025 Kian Kasad

Compression strategy

arbitrarytextthistexthasnomeaning

Copyright © 2025 Kian Kasad

The solutions

Fully implemented version of the workbook:

https://puhack.horse/compression-py-solved

Includes both encoding and decoding for RLE, Huffman, LZ77, and (our version of) DEFLATE.

Copyright © 2025 Kian Kasad