2/22/23 Data Compression
"The Bit Player" Movie Claude Shannon

Communication
- Source - Produces message
- Transmitter - where compression occurs
- Channel - Medium
- Receiver - inverse operation of transmitter, decompresses
- Destination - target

Entropy - uncertainty - avg # of correct guess
- measure of randomness
- opposite of uniformness

ex) AAAA vs ABCD
uniform has entropy
100% of guess 25% of guesses

Run length code:
exploit low entropy file
show letter & # of times it appears
ex) aaaabb
↓
a3 b2

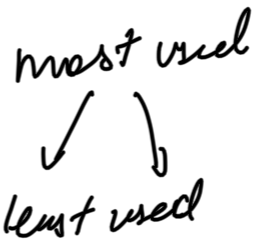
Huffman coding:
- create histogram of message
aka:

bananaaas

a	3
b	1
⋮	⋮

priority Queeue:
- can use heap or queue

least popular last, most popular front



now make a huffman tree

first go through priority queue, add items together & create parent node until only 1 value left, then go through list & put into tree

go left = 0 right = 1

so can represent a most used letter w/ 1 bit

ex) code for A = 0
code for n = 11

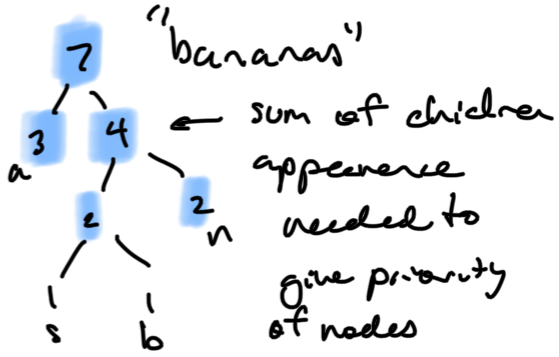
in ascii. each char is 8 bits so you cut down significantly
don't need to separate chars bc all letters are leaf nodes

leaf ↓ interior node

a tree can be rebuilt via the following code: L a L s L b I I n I I

"Dumping the Tree"

tree must be written to front of file



use a stack to traverse tree,
push as you go down, copy stack to file

need to know length of message, bc there may be extra bits @ the end of the buffer