COMP9319 Web Data Compression and Search

LZW,

Adaptive Huffman,

Basic BWT

Dictionary coding

- Patterns: correlations between part of the data
- Idea: replace recurring patterns with references to dictionary
- LZ algorithms are adaptive:
  - Universal coding (the prob. distr. of a symbol is unknown)
  - Single pass (dictionary created on the fly)
  - No need to transmit/store dictionary

LZ77 & LZ78

- LZ77: referring to previously processed data as dictionary
- LZ78: use an explicit dictionary

Lempel-Ziv-Welch (LZW) Algorithm

- Most popular modification to LZ78
- Very common, e.g., Unix compress, TIFF, GIF, PDF (until recently)
- Read http://en.wikipedia.org/wiki/LZW regarding its patents
- Fixed-length references (12bit 4096 entries)
- Static after max entries reached

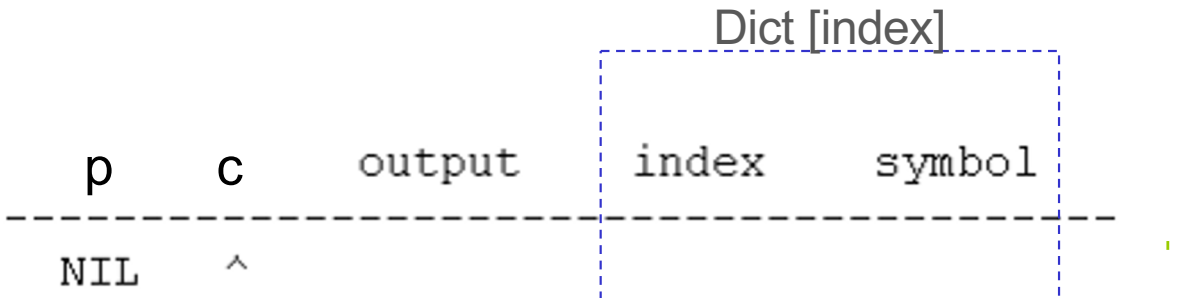
Patent issues again

From Wikipedia: “In 1993–94, and again in 1999, Unisys Corporation received widespread condemnation when it attempted to enforce licensing fees for LZW in GIF images. The 1993–1994 Unisys-Compuserve (Compuserve being the creator of the GIF format) controversy engendered a Usenet comp.graphics discussion *Thoughts on a GIF-replacement file format*, which in turn fostered an email exchange that eventually culminated in the creation of the patent-unencumbered Portable Network Graphics (PNG) file format in 1995. Unisys's US patent on the LZW algorithm expired on June 20, 2003 ...”

LZW Compression

```
p = nil;      // p for prev char
while read(c):
    if pc ∈ Dict:
        p = pc;
    else:
        add pc to Dict;
        output code(p);
        p = c;
```

Example



Input: ^WED^WE^WEE^WEB^WET

```
p = nil;      // p for prev char
while read(c):
    if pc ∈ Dict:
        p = pc;
    else:
        add pc to Dict;
        output code(p);
        p = c;
```

Example

Input: ^WED^WE^WEE^WEB^WET

Dict [index]

p	c	output	index	symbol
NIL	^			
^	W	^	256	^W

```
p = nil;     // p for prev char
while read(c):
    if pc ∈ Dict:
        p = pc;
    else:
        add pc to Dict;
        output code(p);
        p = c;
```

Example

Input: ^WED^WE^WEE^WEB^WET

Dict [index]

p	c	output	index	symbol
NIL	^			
^	W	^	256	^W
W	E	W	257	WE

```
p = nil;      // p for prev char
while read(c):
    if pc ∈ Dict:
        p = pc;
    else:
        add pc to Dict;
        output code(p);
        p = c;
```

Example

Input: ^WED^WE^WEE^WEB^WET

Dict [index]

p	c	output	index	symbol
NIL	^			
^	W	^	256	^W
W	E	W	257	WE
E	D	E	258	ED
D	^	D	259	D^
^	W			

```
p = nil;     // p for prev char
while read(c):
    if pc ∈ Dict:
        p = pc;
    else:
        add pc to Dict;
        output code(p);
        p = c;
```

Example

Input: ^WED^WE^WEE^WEB^WET

Dict [index]

p	c	output	index	symbol
NIL	^			
^	W	^	256	^W
W	E	W	257	WE
E	D	E	258	ED
D	^	D	259	D^
^	W			
^W	E	256	260	^WE

```
p = nil;      // p for prev char
while read(c):
    if pc ∈ Dict:
        p = pc;
    else:
        add pc to Dict;
        output code(p);
        p = c;
```

Example

Input: ^WED^WE^WEE^WEB^WET

Dict [index]

p	c	output	index	symbol
NIL	^			
^	W	^	256	^W
W	E	W	257	WE
E	D	E	258	ED
D	^	D	259	D^
^	W			
^W	E	256	260	^WE
E	^	E	261	E^

```
p = nil;     // p for prev char
while read(c):
    if pc ∈ Dict:
        p = pc;
    else:
        add pc to Dict;
        output code(p);
        p = c;
```

Example

Input: ^WED^WE^WEE^WEB^WET

Dict [index]

p	c	output	index	symbol
NIL	^			
^	W	^	256	^W
W	E	W	257	WE
E	D	E	258	ED
D	^	D	259	D^
^	W			
^W	E	256	260	^WE
E	^	E	261	E^
^	W			

```
p = nil;     // p for prev char
while read(c):
    if pc ∈ Dict:
        p = pc;
    else:
        add pc to Dict;
        output code(p);
        p = c;
```

Example

Input: ^WED^WE^WEE^WEB^WET

Dict [index]

p	c	output	index	symbol
NIL	^			
^	W	^	256	^W
W	E	W	257	WE
E	D	E	258	ED
D	^	D	259	D^
^	W			
^W	E	256	260	^WE
E	^	E	261	E^
^	W			
^W	E			

```
p = nil;     // p for prev char
while read(c):
    if pc ∈ Dict:
        p = pc;
    else:
        add pc to Dict;
        output code(p);
        p = c;
```

Example

Input: ^WED^WE^WEE^WEB^WET

Dict [index]

p	c	output	index	symbol
NIL	^			
^	W	^	256	^W
W	E	W	257	WE
E	D	E	258	ED
D	^	D	259	D^
^	W			
^W	E	256	260	^WE
E	^	E	261	E^
^	W			
^W	E			
^WE	E	260	262	^WEE

```
p = nil;     // p for prev char
while read(c):
    if pc ∈ Dict:
        p = pc;
    else:
        add pc to Dict;
        output code(p);
        p = c;
```

Example

Input: ^WED^WE^WEE^WEB^WET

p	c	output	Dict [index] index	Dict [index] symbol
NIL	^			
^	W	^	256	^W
W	E	W	257	WE
E	D	E	258	ED
D	^	D	259	D^
^	W			
^W	E	256	260	^WE
E	^	E	261	E^
^	W			
^W	E			
^WE	E	260	262	^WEE
E	^			

```
p = nil;     // p for prev char
while read(c):
    if pc ∈ Dict:
        p = pc;
    else:
        add pc to Dict;
        output code(p);
        p = c;
```

Example

Input: ^WED^WE^WEE^WEB^WET

```
p = nil;      // p for prev char
while read(c):
    if pc ∈ Dict:
        p = pc;
    else:
        add pc to Dict;
        output code(p);
        p = c;
```

p	c	output	Dict [index] index	symbol
NIL	^			
^	W	^	256	^W
W	E	W	257	WE
E	D	E	258	ED
D	^	D	259	D^
^	W			
^W	E	256	260	^WE
E	^	E	261	E^
^	W			
^W	E			
^WE	E	260	262	^WEE
E	^			
E^	W	261	263	E^W
W	E			
WE	B	257	264	WEB
B	^	B	265	B^
^	W			
^W	E			
^WE	T	260	266	^WET
T	EOF	T		

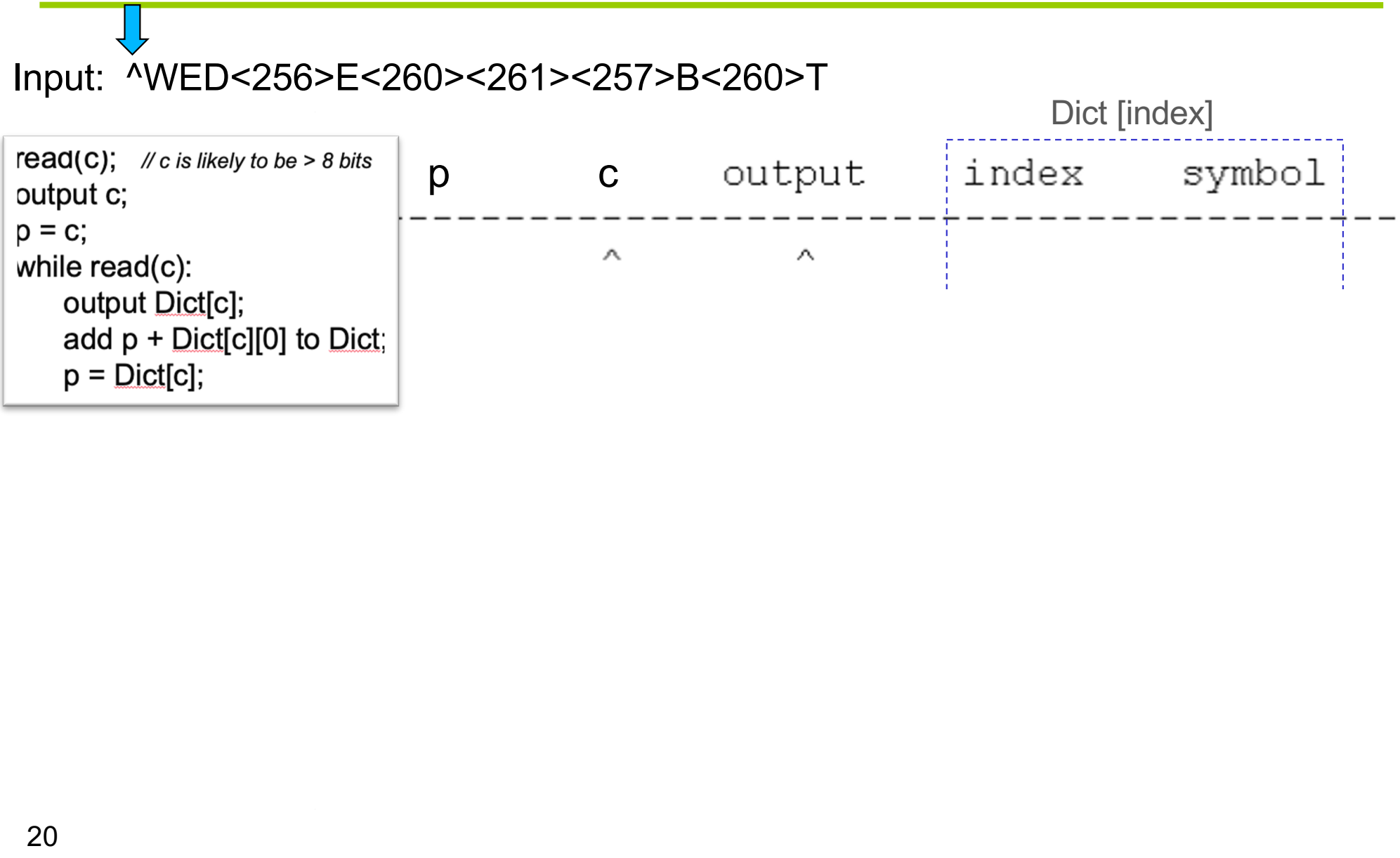
LZW Compression

- Original LZW used dictionary with 4K entries, first 256 (0-255) are ASCII codes.
- In the above example, a 19 symbols reduced to 7 symbols & 5 code. Each code/symbol will need 8+ bits, say 9 bits.
- Reference: Terry A. Welch, "A Technique for High Performance Data Compression", IEEE Computer, Vol. 17, No. 6, 1984, pp. 8-19.

LZW Decompression

```
read(c);    // c is likely to be > 8 bits
output c;
p = c;
while read(c):
      output Dict[c];
      add p + Dict[c][0] to Dict;
      p = Dict[c];
```

Example


Input: ^WED<256>E<260><261><257>B<260>T
read(c); // c is likely to be > 8 bits
output c;
p = c;
while read(c):
output Dict[c];
add p + Dict[c][0] to Dict;
p = Dict[c];
Dict [index]
p
c
output
index
symbol

Example

Input: ^WED<256>E<260><261><257>B<260>T

```
read(c);   // c is likely to be > 8 bits
output c;
p = c;
while read(c):
    output Dict[c];
    add p + Dict[c][0] to Dict;
    p = Dict[c];
```

p	c	output	Dict [index] index	Dict [index] symbol
	^	^		
^	W	W	256	^W

Example

Input: ^WED<256>E<260><261><257>B<260>T

```
read(c);   // c is likely to be > 8 bits
output c;
p = c;
while read(c):
    output Dict[c];
    add p + Dict[c][0] to Dict;
    p = Dict[c];
```

p	c	output	Dict [index] index	Dict [index] symbol
	^	^		
^	W	W	256	^W
W	E	E	257	WE

Example

Input: ^WED<256>E<260><261><257>B<260>T

```
read(c);   // c is likely to be > 8 bits
output c;
p = c;
while read(c):
    output Dict[c];
    add p + Dict[c][0] to Dict;
    p = Dict[c];
```

p	c	output	Dict [index] index	Dict [index] symbol
	^	^		
^	W	W	256	^W
W	E	E	257	WE
E	D	D	258	ED

Example

Input: ^WED<256>E<260><261><257>B<260>T

```
read(c);   // c is likely to be > 8 bits
output c;
p = c;
while read(c):
      output Dict[c];
      add p + Dict[c][0] to Dict;
      p = Dict[c];
```

p	c	output	Dict [index] index	Dict [index] symbol
	^	^		
^	W	W	256	^W
W	E	E	257	WE
E	D	D	258	ED
D	<256>	^W	259	D^

Example

Input: ^WED<256>E<260><261><257>B<260>T

```
read(c);   // c is likely to be > 8 bits
output c;
p = c;
while read(c):
      output Dict[c];
      add p + Dict[c][0] to Dict;
      p = Dict[c];
```

Dict [index]

p	c	output	index	symbol
	^	^		
^	W	W	256	^W
W	E	E	257	WE
E	D	D	258	ED
D	<256>	^W	259	D^
<256>	E	E	260	^WE

Example

Input: ^WED<256>E<260><261><257>B<260>T

```
read(c);   // c is likely to be > 8 bits
output c;
p = c;
while read(c):
      output Dict[c];
      add p + Dict[c][0] to Dict;
      p = Dict[c];
```

p	c	output	Dict [index] index	Dict [index] symbol
	^	^		
^	W	W	256	^W
W	E	E	257	WE
E	D	D	258	ED
D	<256>	^W	259	D^
<256>	E	E	260	^WE
E	<260>	^WE	261	E^

Example

Input: ^WED<256>E<260><261><257>B<260>T

```
read(c);   // c is likely to be > 8 bits
output c;
p = c;
while read(c):
    output Dict[c];
    add p + Dict[c][0] to Dict;
    p = Dict[c];
```

p	c	output	Dict [index] index	symbol
	^	^		
^	W	W	256	^W
W	E	E	257	WE
E	D	D	258	ED
D	<256>	^W	259	D^
<256>	E	E	260	^WE
E	<260>	^WE	261	E^
<260>	<261>	E^	262	^WEE

Example

Input: ^WED<256>E<260><261><257>B<260>T

```
read(c);   // c is likely to be > 8 bits
output c;
p = c;
while read(c):
    output Dict[c];
    add p + Dict[c][0] to Dict;
    p = Dict[c];
```

p	c	output
	^	^
^	W	W
W	E	E
E	D	D
D	<256>	^W
<256>	E	E
E	<260>	^WE
<260>	<261>	E^
<261>	<257>	WE
<257>	B	B
B	<260>	^WE
<260>	T	T

Dict [index]

index	symbol
256	^W
257	WE
258	ED
259	D^
260	^WE
261	E^
262	^WEE
263	E^W
264	WEB
265	B^
266	^WET

Note: LZW decoding

- There is one special case that the LZW decoding pseudocode presented is unable to handle.
- This is your exercise to find out in what situation that happens, and how to deal with it.
- I'll go through this at the live lecture.

LZW implementation

- Parsing fixed number of bits from input is easy
- Fast and efficient

Types (revision)

- Block-block
  - source message and codeword: fixed length
  - e.g., ASCII
- Block-variable
  - source message: fixed; codeword: variable
  - e.g., Huffman coding
- Variable-block
  - source message: variable; codeword: fixed
  - e.g., LZW
- Variable-variable
  - source message and codeword: variable
  - e.g., Arithmetic coding

So far

We have covered:

- Course overview
- Background
- RLE
- Entropy
- Huffman code
- Arithmetic code
- LZW

More online readings

http://www.ics.uci.edu/~dan/pubs/DC-Sec1.html

http://marknelson.us/1991/02/01/arithmetic-coding-statistical-modeling-data-compression/

Lossless compression revisited

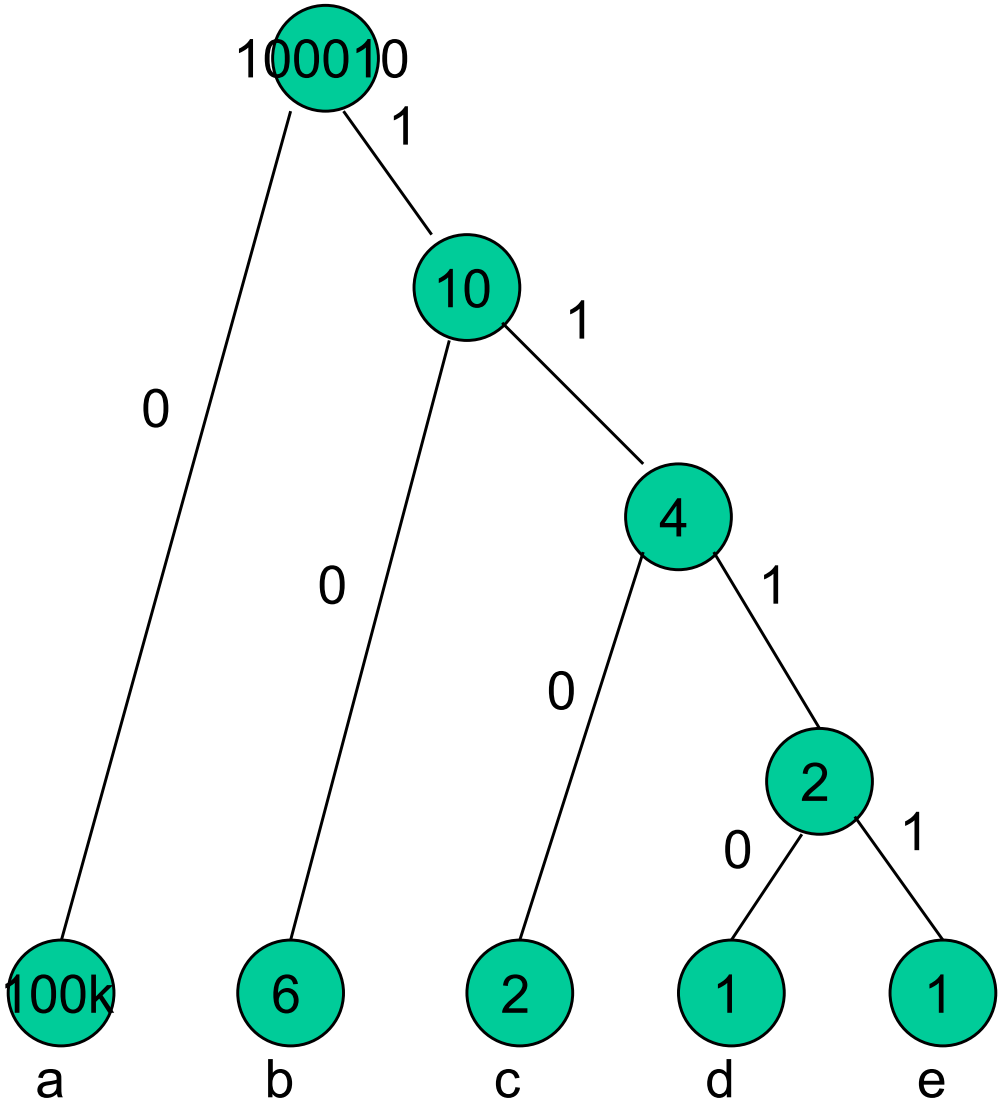
- Run-length coding
- Statistical methods
  - Huffman coding
  - Arithmetic coding
- Dictionary methods
  - Lempel Ziv algorithms

Static (Huffman, AC) vs Adaptive (LZW)

Huffman Coding (revisit)
and then
Adaptive Huffman

Huffman coding

S	Freq	Huffman
a	100000	0
b	6	10
c	2	110
d	1	1110
e	1	1111



Huffman not optimal

H = 0.9999 log 1.0001 + 0.00006 log 16668.333 + … + 1/100010 log 100010

≈ 0.00

L = (100000*1 + …)/100010

≈ 1

Problems of Huffman coding

Huffman codes have an integral # of bits.

E.g., log (3) = 1.585 while Huffman may need 2 bits

Noticeable non-optimality when prob of a symbol is high.

=> Arithmetic coding

Problems of Huffman coding

Need statistics & static: e.g., single pass over the data just to collect stat & stat unchanged during encoding

To decode, the stat table need to be transmitted. Table size can be significant for small msg.

=> Adaptive compression e.g., adaptive huffman

Adaptive compression

Encoder

Initialize the model

Repeat for each input char

(

Encode char

Update the model

)

Decoder

Initialize the model

Repeat for each input char

(

Decode char

Update the model

)

Make sure both sides have the same Initialize & update model algorithms.

Adaptive Huffman Coding (dummy)

Encoder

Reset the stat

Repeat for each input char

(

Encode char

Update the stat

Rebuild huffman tree

)

Decoder

Reset the stat

Repeat for each input char

(

Decode char

Update the stat

Rebuild huffman tree

)

Adaptive Huffman Coding (dummy)

Encoder

Reset the stat

Repeat for each input char

(

Encode char

Update the stat

Rebuild huffman tree

)

Decoder

Reset the stat

Repeat for each input char

(

Decode char

Update the stat

Rebuild huffman tree

)

This works but too slow!

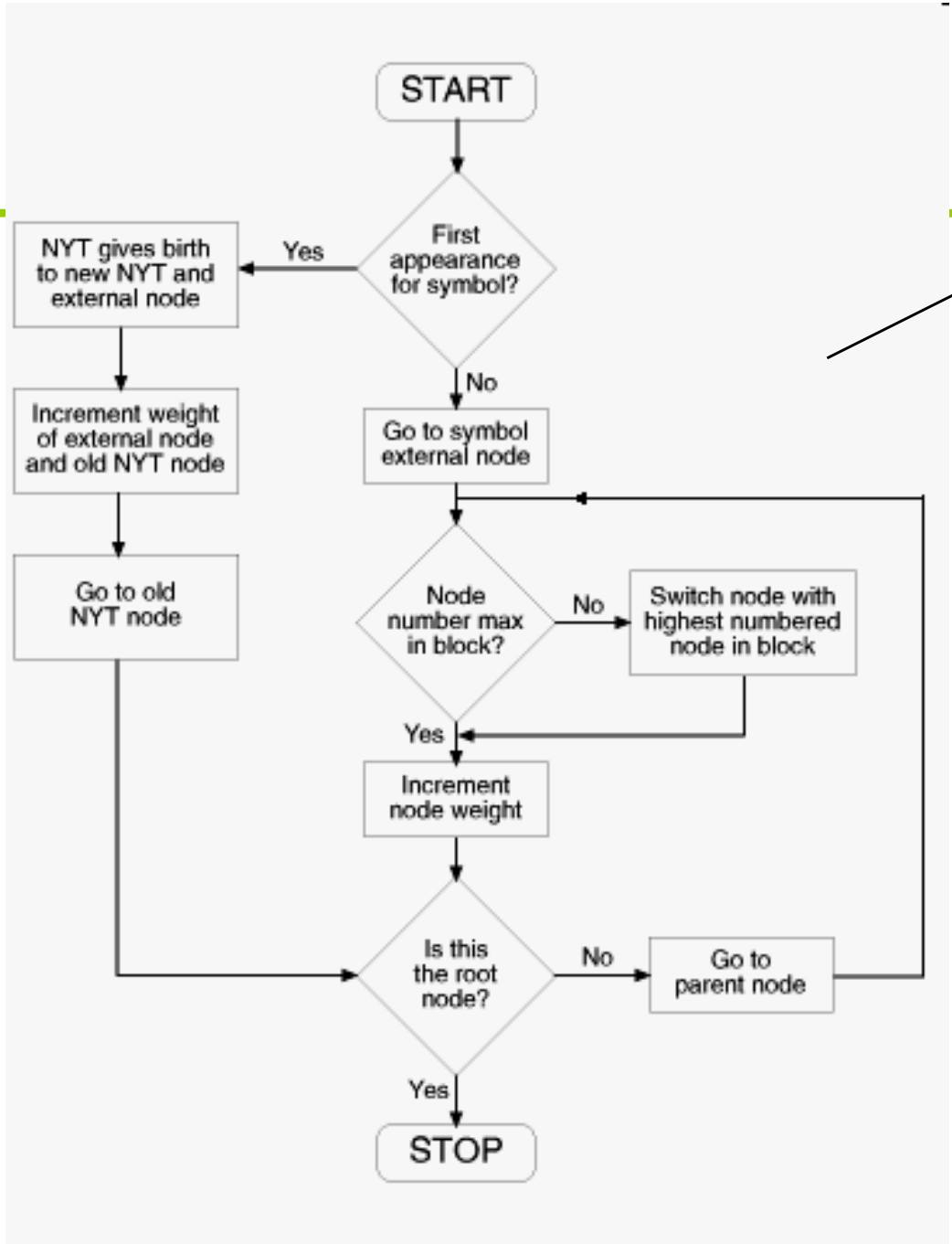
Adaptive Huffman (Algorithm outline)

1. If current symbol is NYT, add two child nodes to NYT node. One will be a new NYT node the other is a leaf node for our symbol. Increase weight for the new leaf node and the old NYT and go to step 4. If not, go to symbol's leaf node.

2. If this node does not have the highest number in a block, swap it with the node having the highest number

3. Increase weight for current node

4. If this is not the root node go to parent node then go to step 2. If this is the root, end.

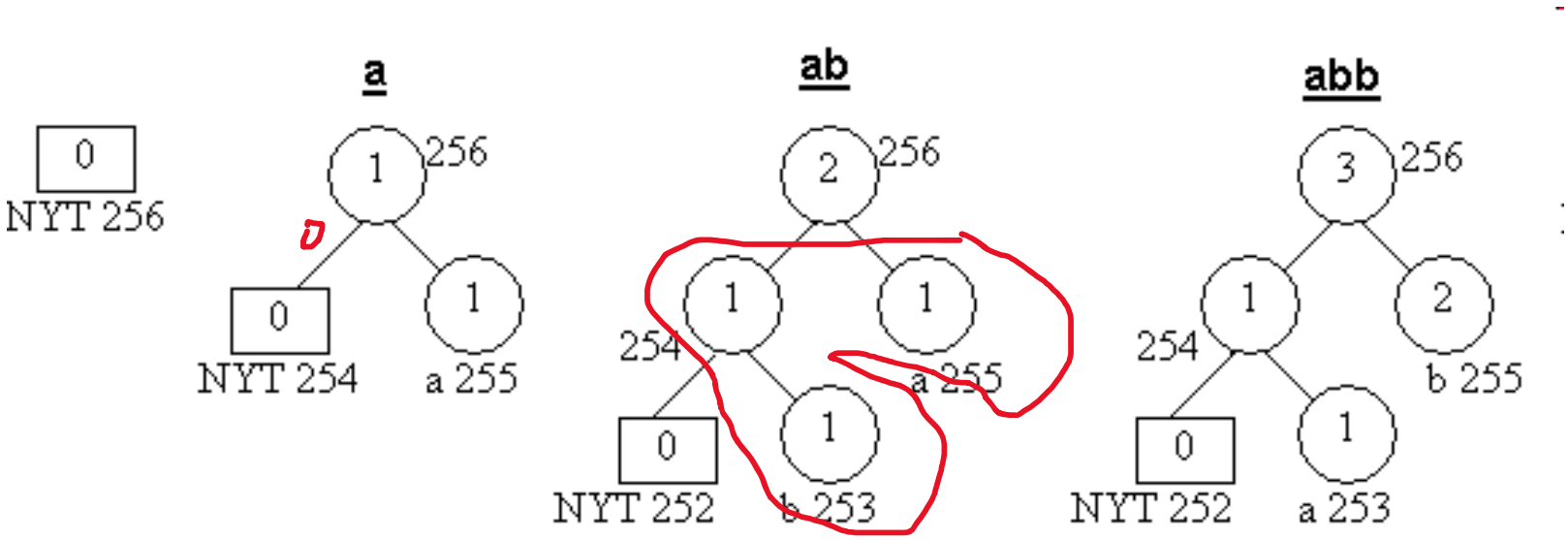


The update procedure from Introduction to Data Compression by by Sayood Khalid

Adaptive Huffman

abbbbba: 01100001011000100110001001100010011000100110001001100001

abbbbba: 0110000100110001000111101



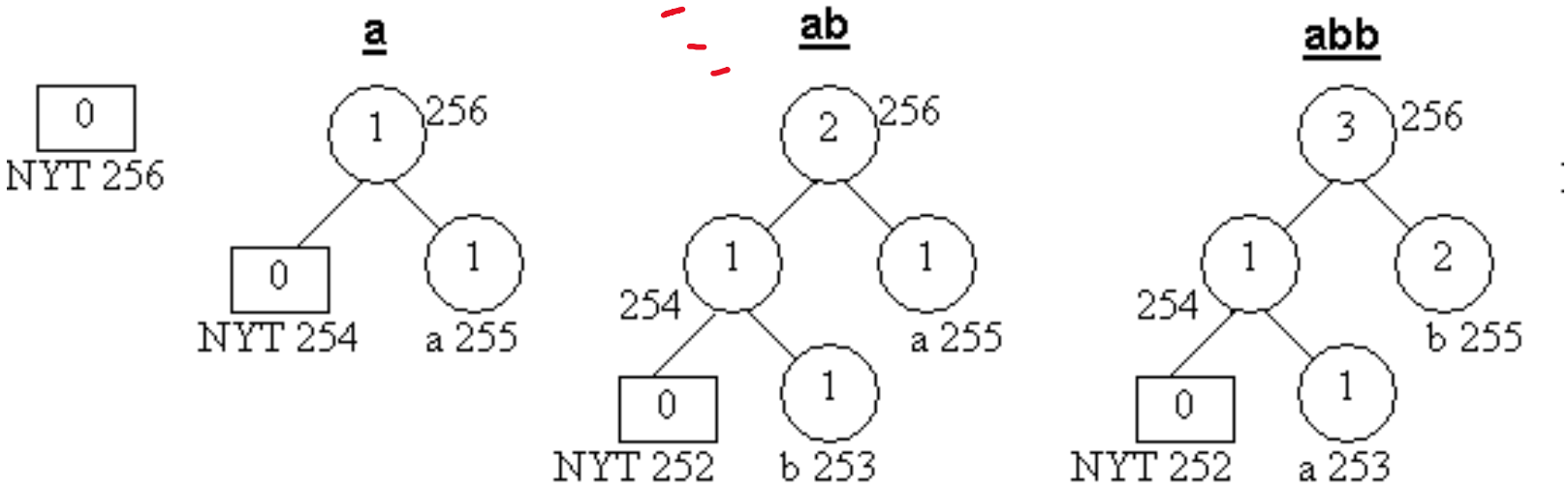
a: 01100001
b: 01100010

From an old Wikipedia page

Adaptive Huffman

abbbbba: 01100001011000100110001001100010011000100110001001100001

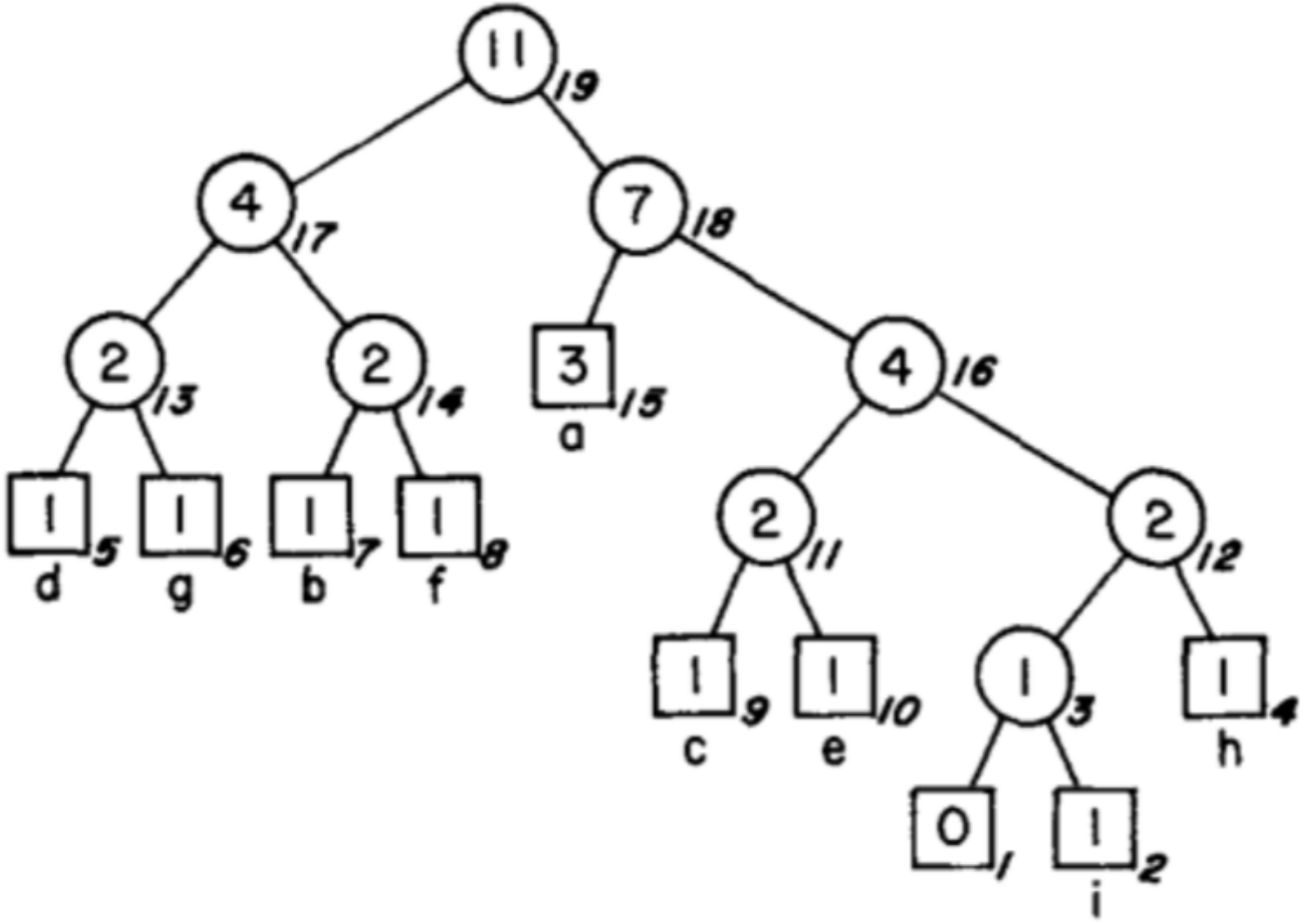
abbbbba: 01100001001100010111101



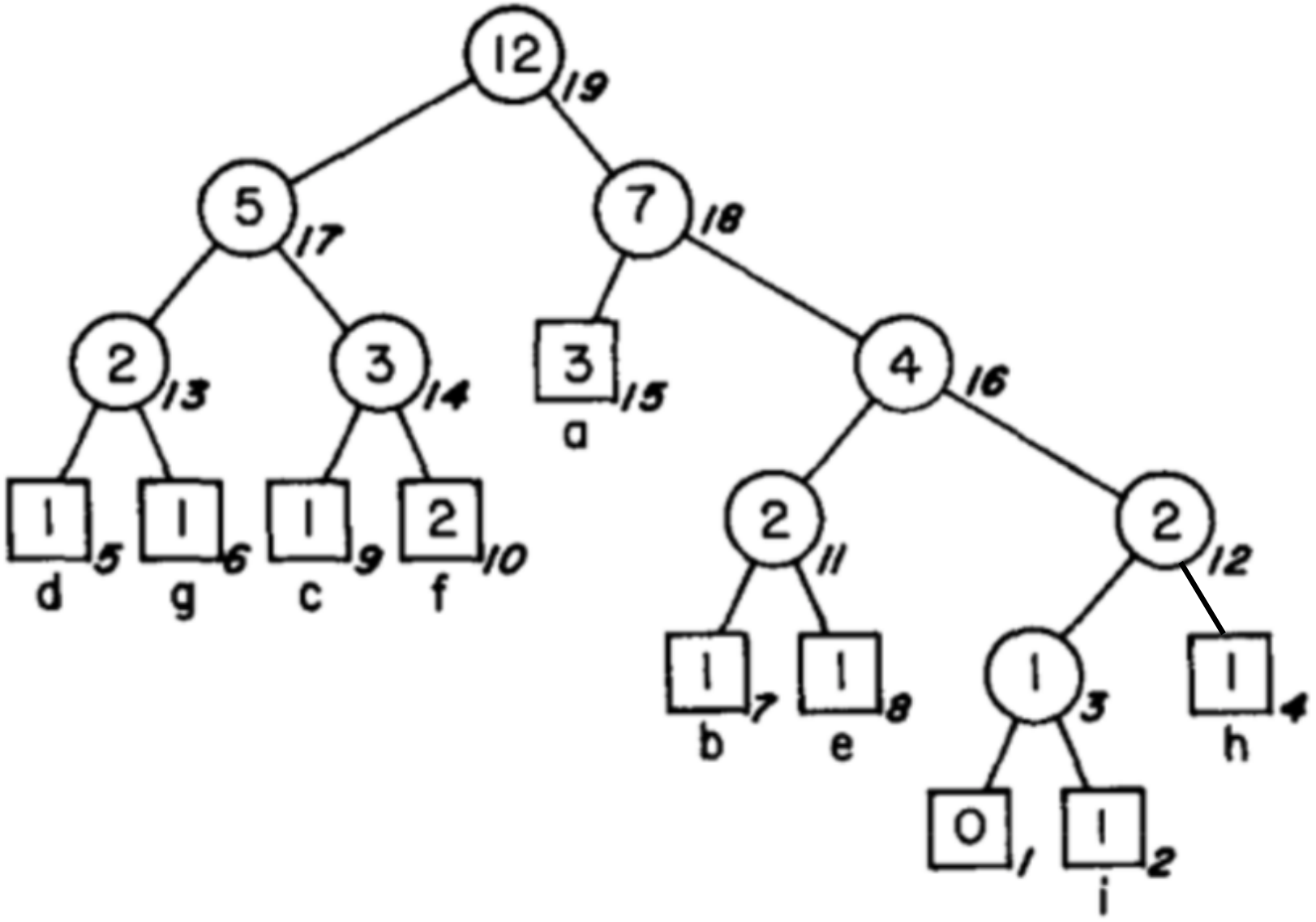
a: 01100001
b: 01100010

From an old Wikipedia page

Adaptive Huffman (FGK)



Adaptive Huffman (FGK): when f is inserted



Adaptive Huffman (FGK vs Vitter)

1.
FGK: (Explicit) node numbering
Vitter: Implicit numbering

2.
Vitter's Invariant:

(*) For each weight w, all leaves of weight w precede (in the implicit numbering) all internal nodes of weight w.

Adaptive Huffman (Vitter'87)

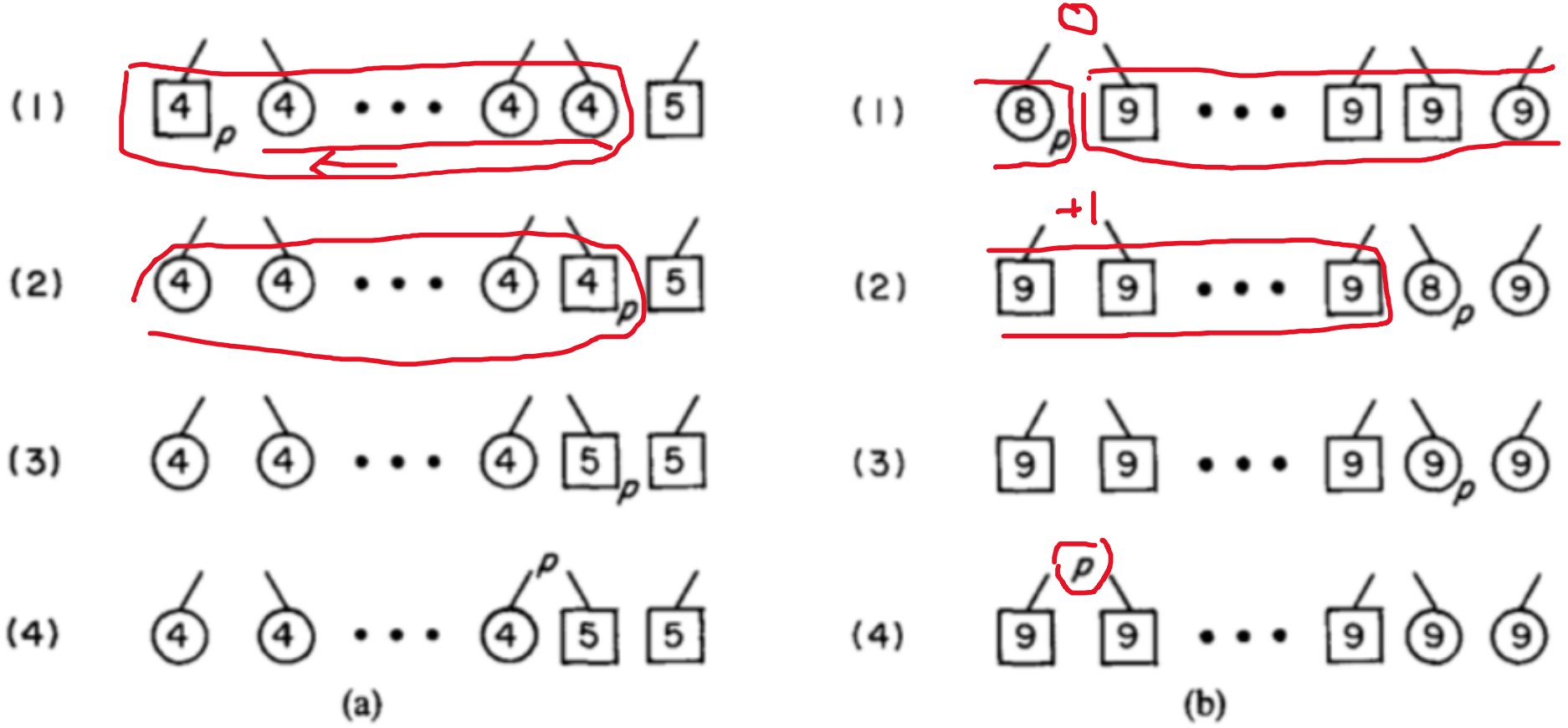
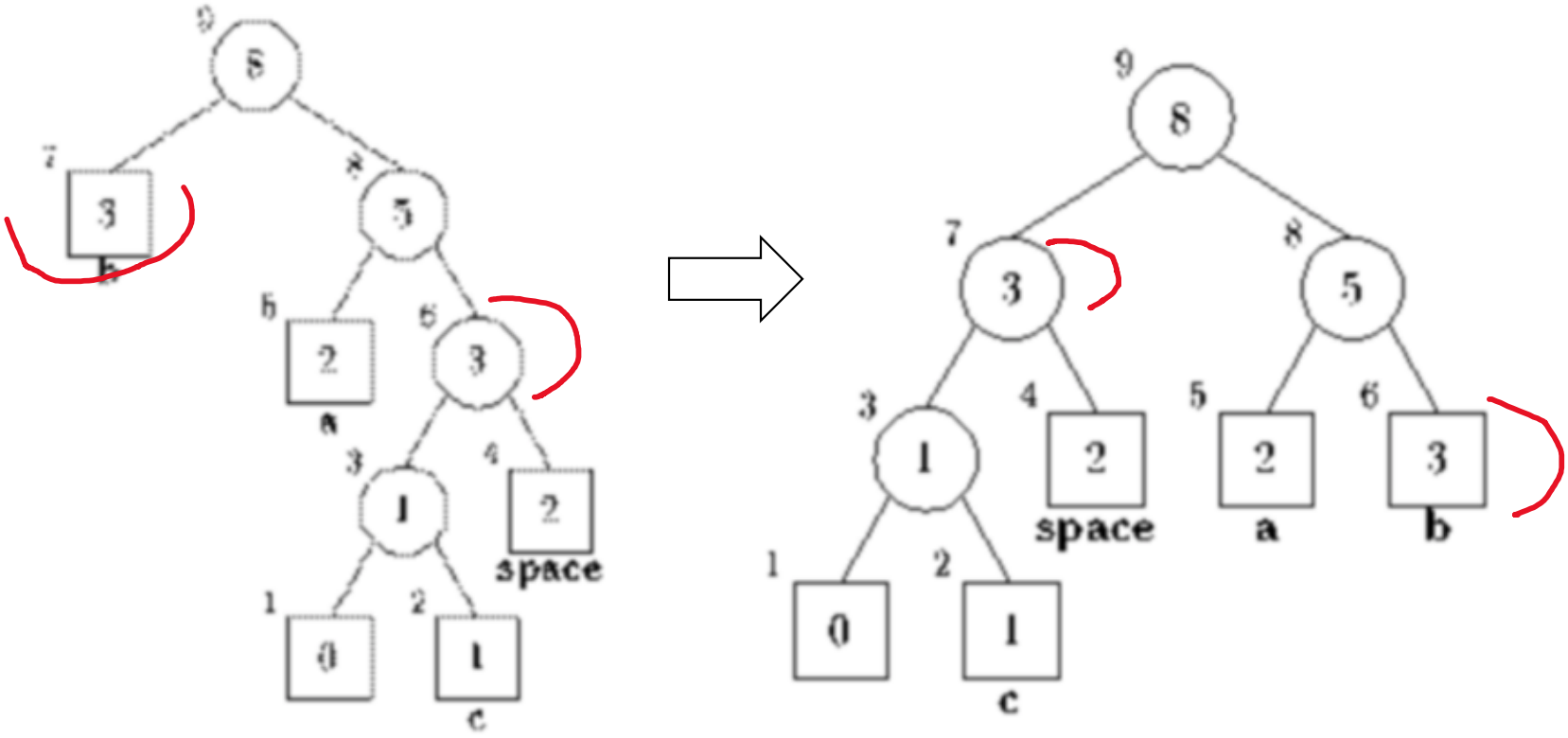


FIG. 6. Algorithm Λ's *SlideAndIncrement* operation. All the nodes in a given block shift to the left one spot to make room for node *p*, which slides over the block to the right. (a) Node *p* is a leaf of weight 4. The internal nodes of weight 4 shift to the left. (b) Node *p* is an internal node of weight 8. The leaves of weight 9 shift to the left.

Adaptive Huffman (Vitter's Invariant)



Issues with Wikipedia

en.wikipedia.org/w/index.php?title=Adaptive_Huffman_coding&oldid=699960545

at the highest-ordered node). All ancestor nodes of the node should also be processed in the same manner.

Since the FGK Algorithm has some drawbacks about the node-or-subtree swapping, Vitter proposed another algorithm to in

dit links

Vitter algorithm

Code is represented as a tree structure in which every node has a corresponding weight and a unique number.

Numbers go down, and from right to left.

Weights must satisfy the sibling property, which states that nodes must be listed in the order of decreasing weight with each a child of B, then $W(A) > W(B) > W(C)$.

The weight is merely the count of symbols transmitted which codes are associated with children of that node.

A set of nodes with same weights make a **block**.

To get the code for every node, in case of binary tree we could just traverse all the path from the root to the node, writing dc

We need some general and straightforward method to transmit symbols that are "not yet transmitted" (NYT). We could use, alphabet.

Encoder and decoder start with only the root node, which has the maximum number. In the beginning it is our initial NYT no

When we transmit an NYT symbol, we have to transmit code for the NYT node, then for its generic code.

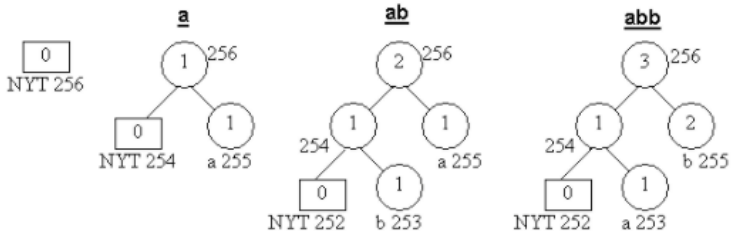
For every symbol that is already in the tree, we only have to transmit code for its leaf node.

For every symbol transmitted both the transmitter and receiver execute the update procedure:

1. If current symbol is NYT, add two child nodes to NYT node. One will be a new NYT node, the other is a leaf node for and go to step 4. If current symbol is not NYT, go to symbol's leaf node.
2. If this node does not have the highest number in a block, swap it with the node having the highest number, except if
3. Increase weight for current node
4. If this is not the root node go to parent node then go to step 2. If this is the root, end.

Note: swapping nodes means swapping weights and corresponding symbols, but not the numbers.

Example



Start with an empty tree.

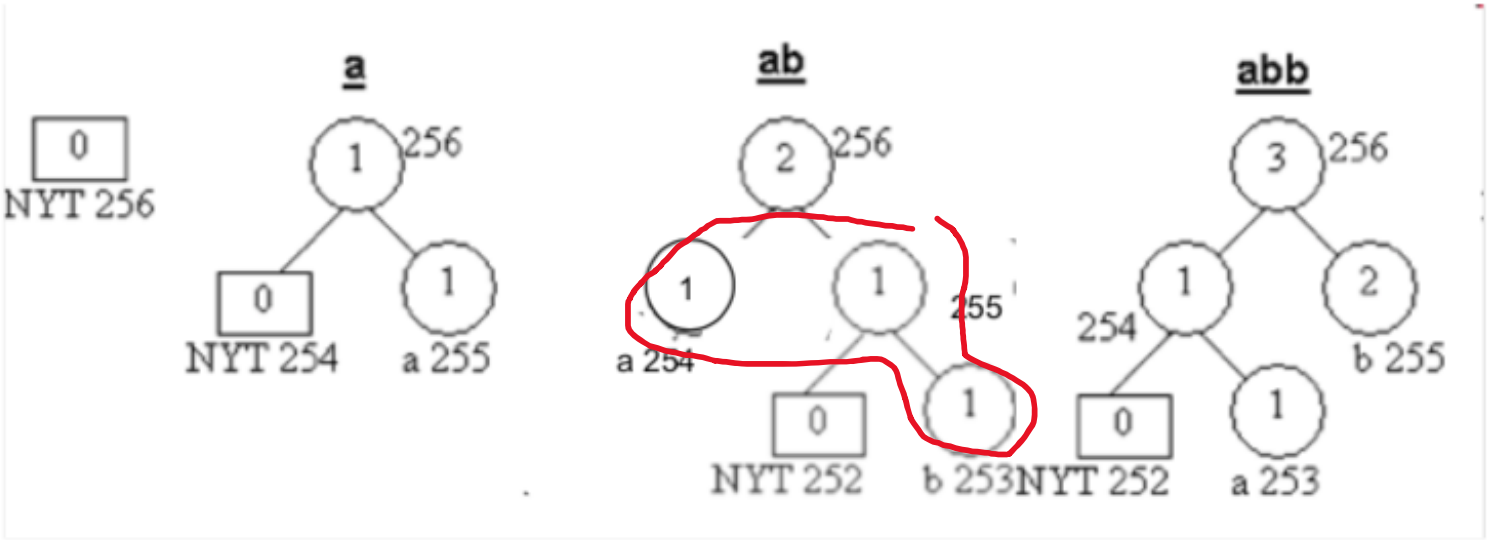
For **"a"** transmit its binary code.

COMP9319 students correcting lots of Wiki pages

Adaptive Huffman (Vitter's)

abbbbba: 01100001011000100110001001100010011000100110001001100001

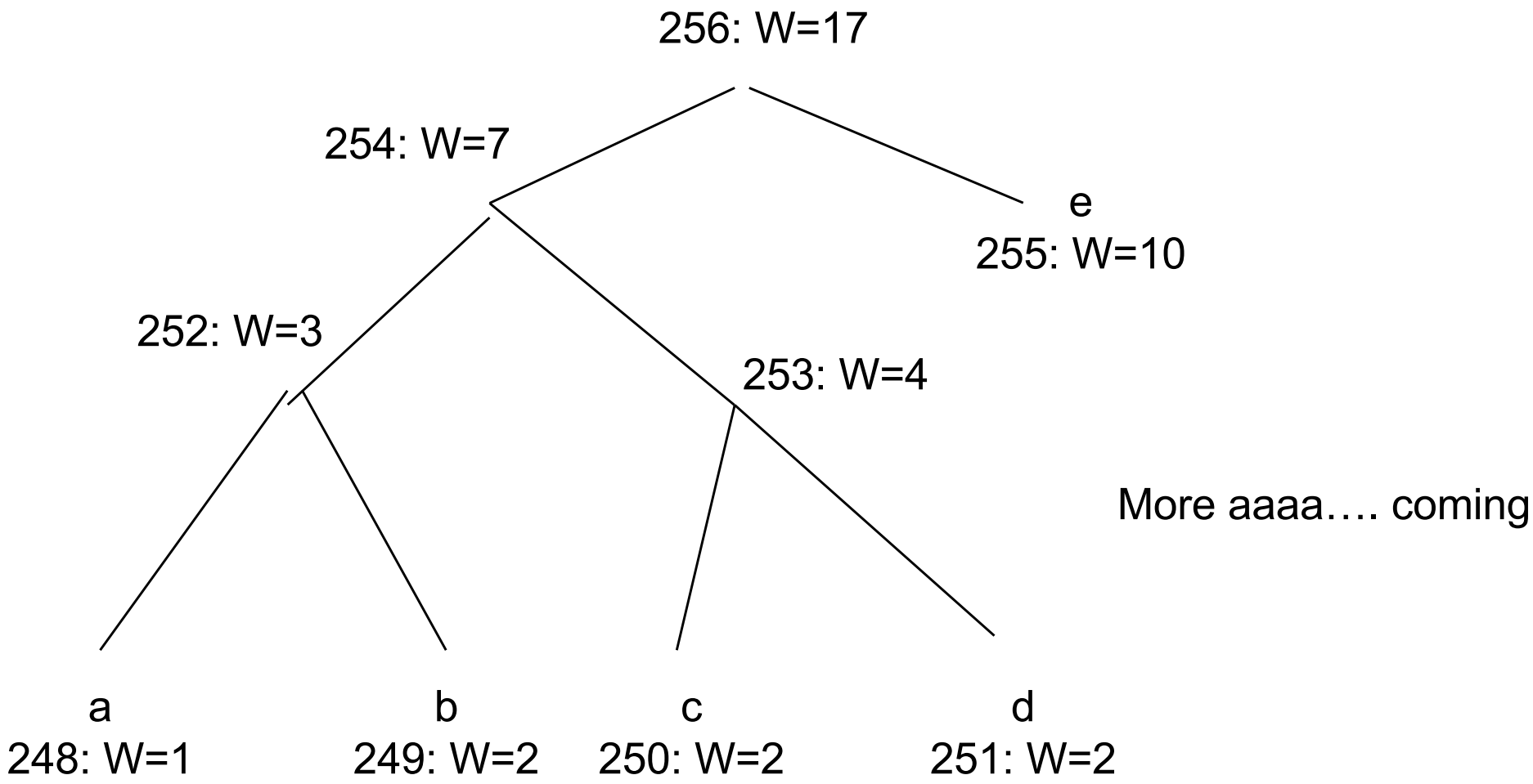
abbbbba: 01100001**0**01100010**11**111101



a: 01100001
b: 01100010

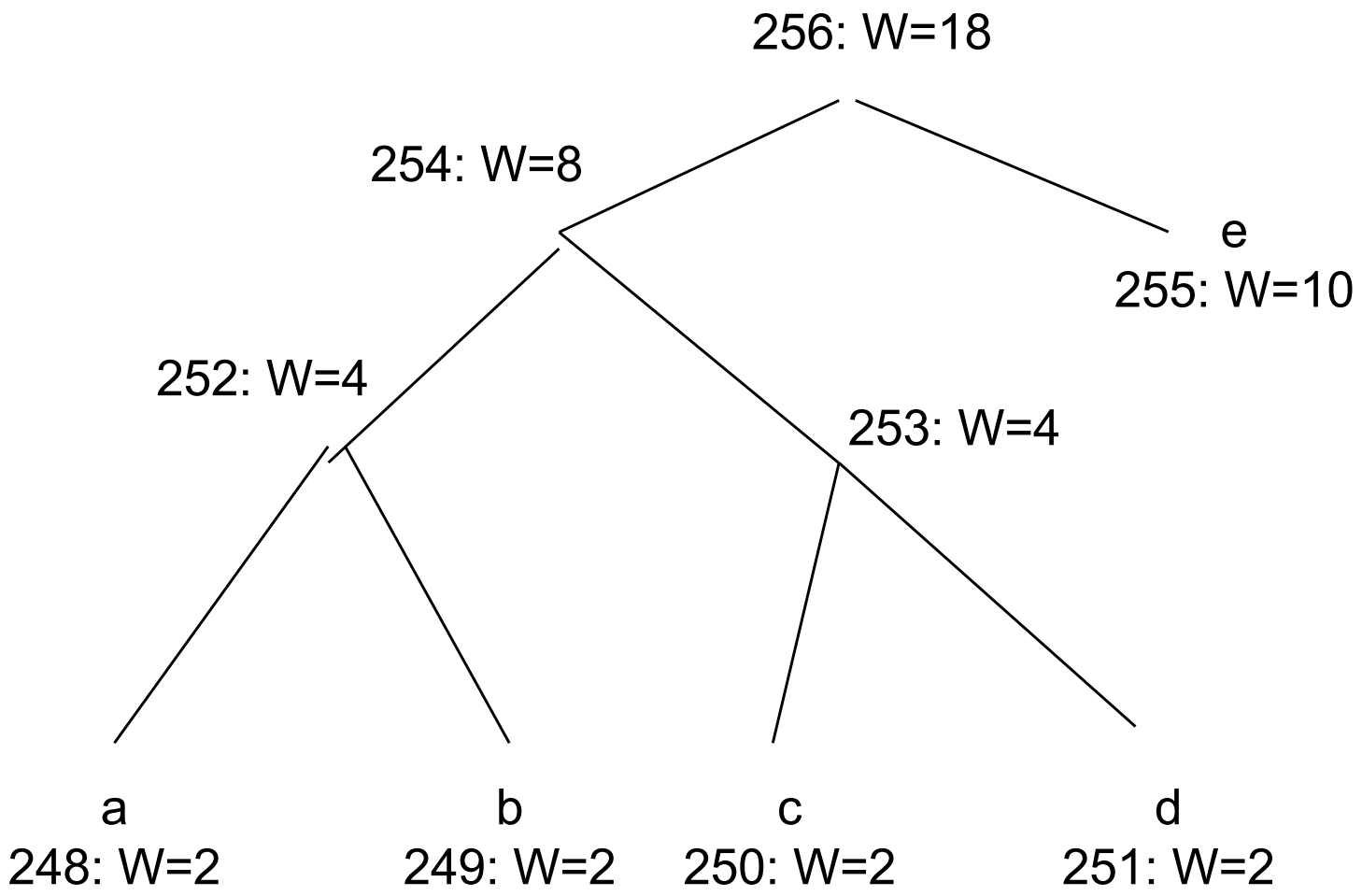
Corrected fr Wikipedia

More example on Vitter's

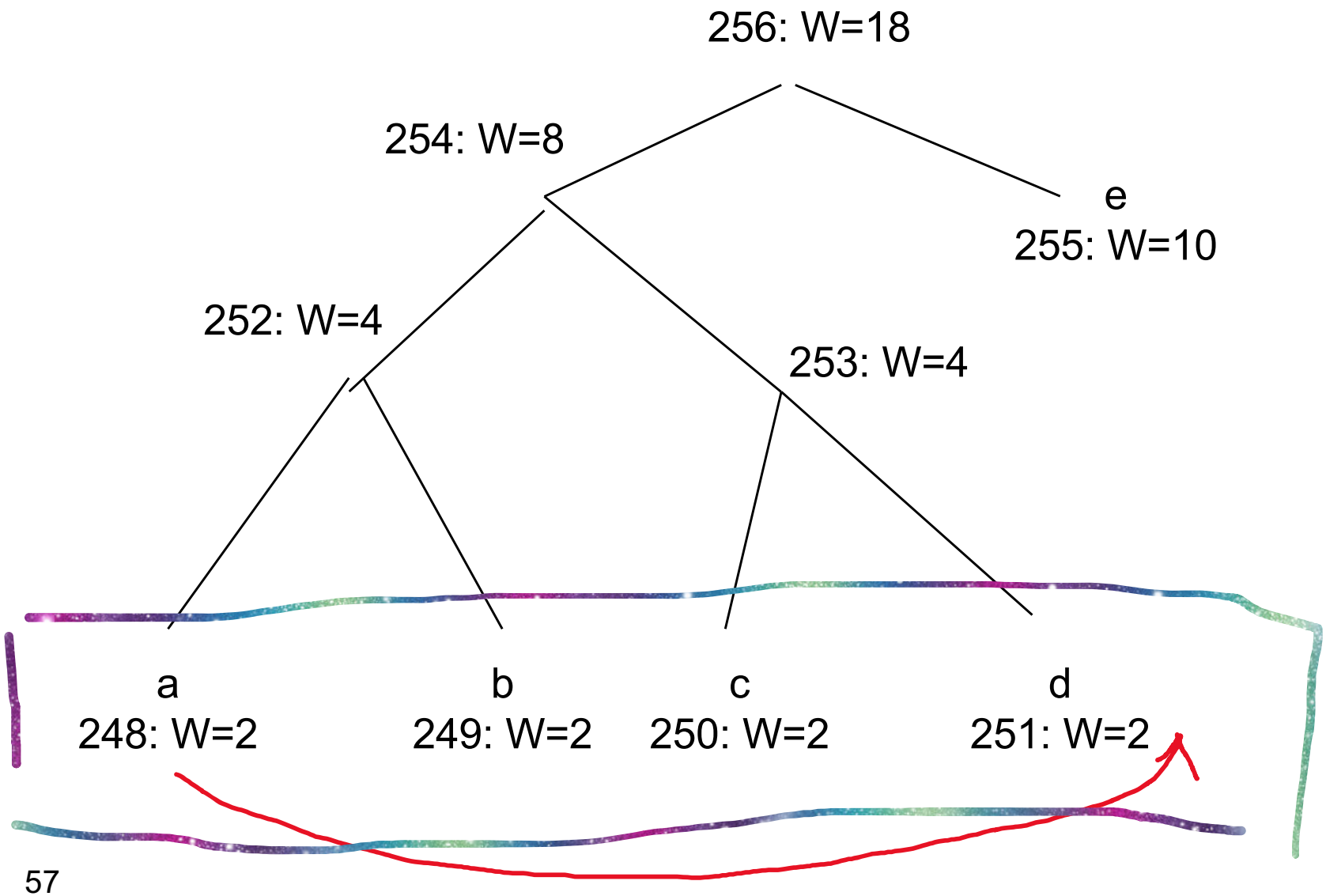


More aaaa…. coming

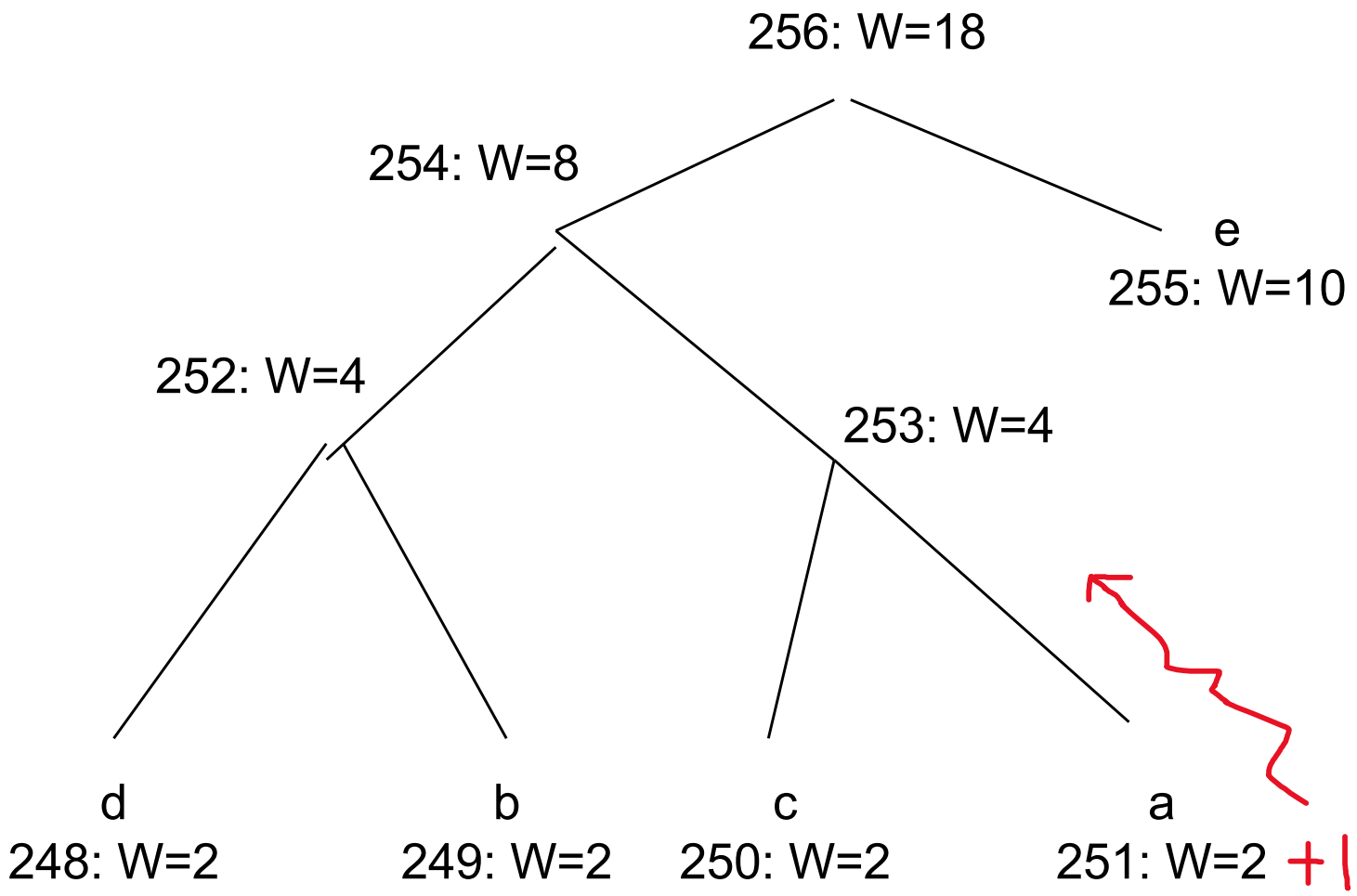
More example



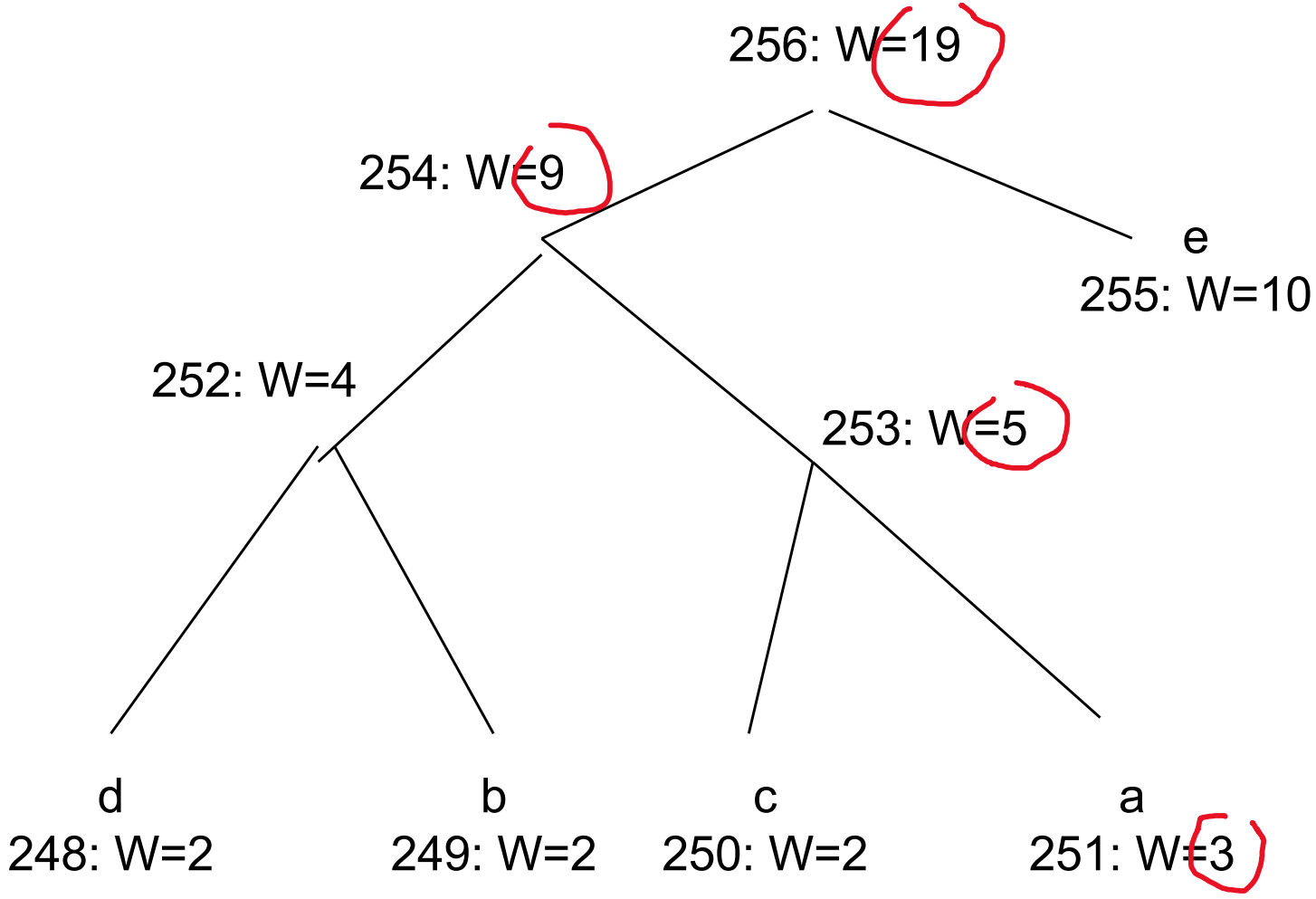
More example



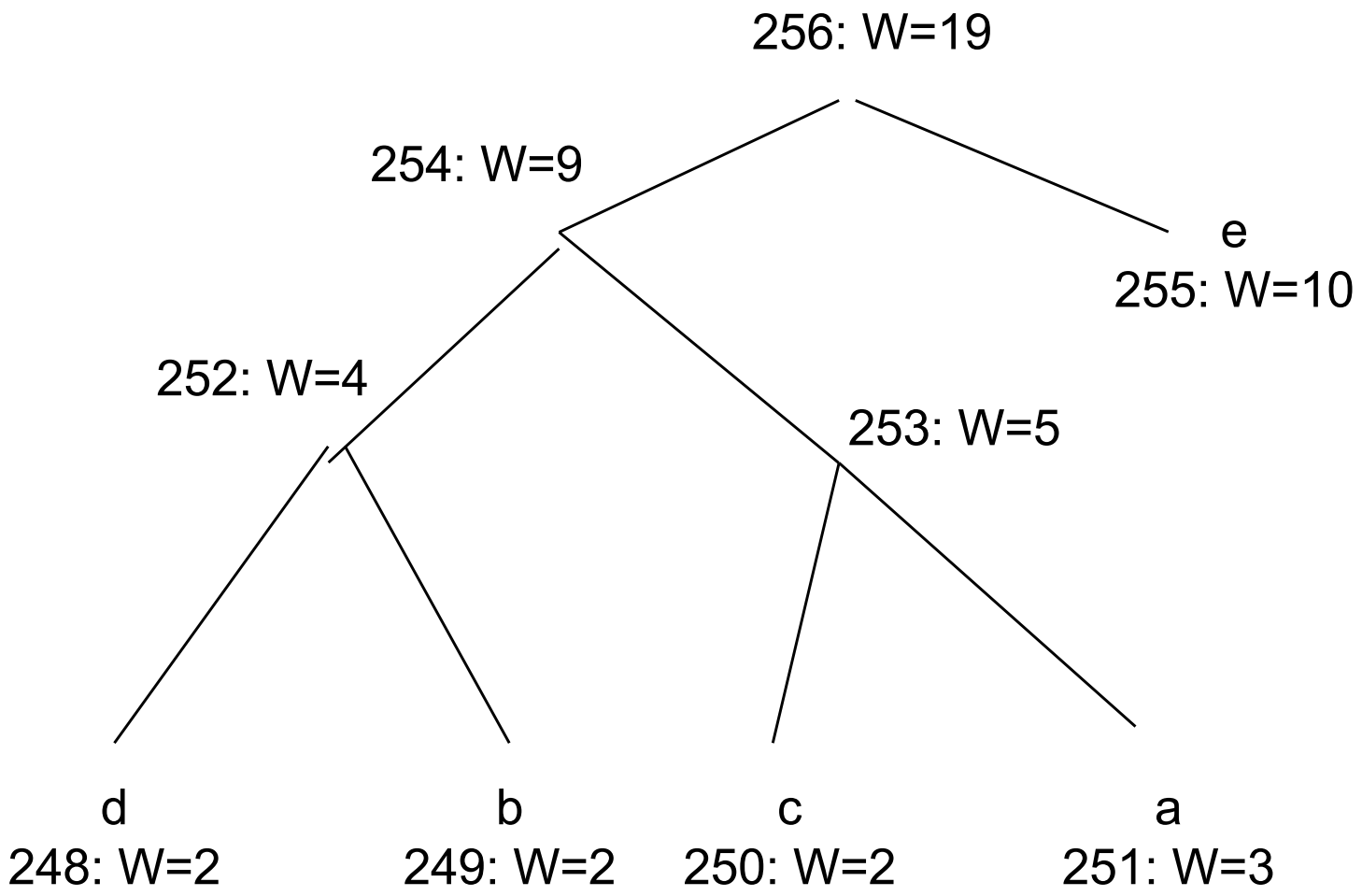
More example



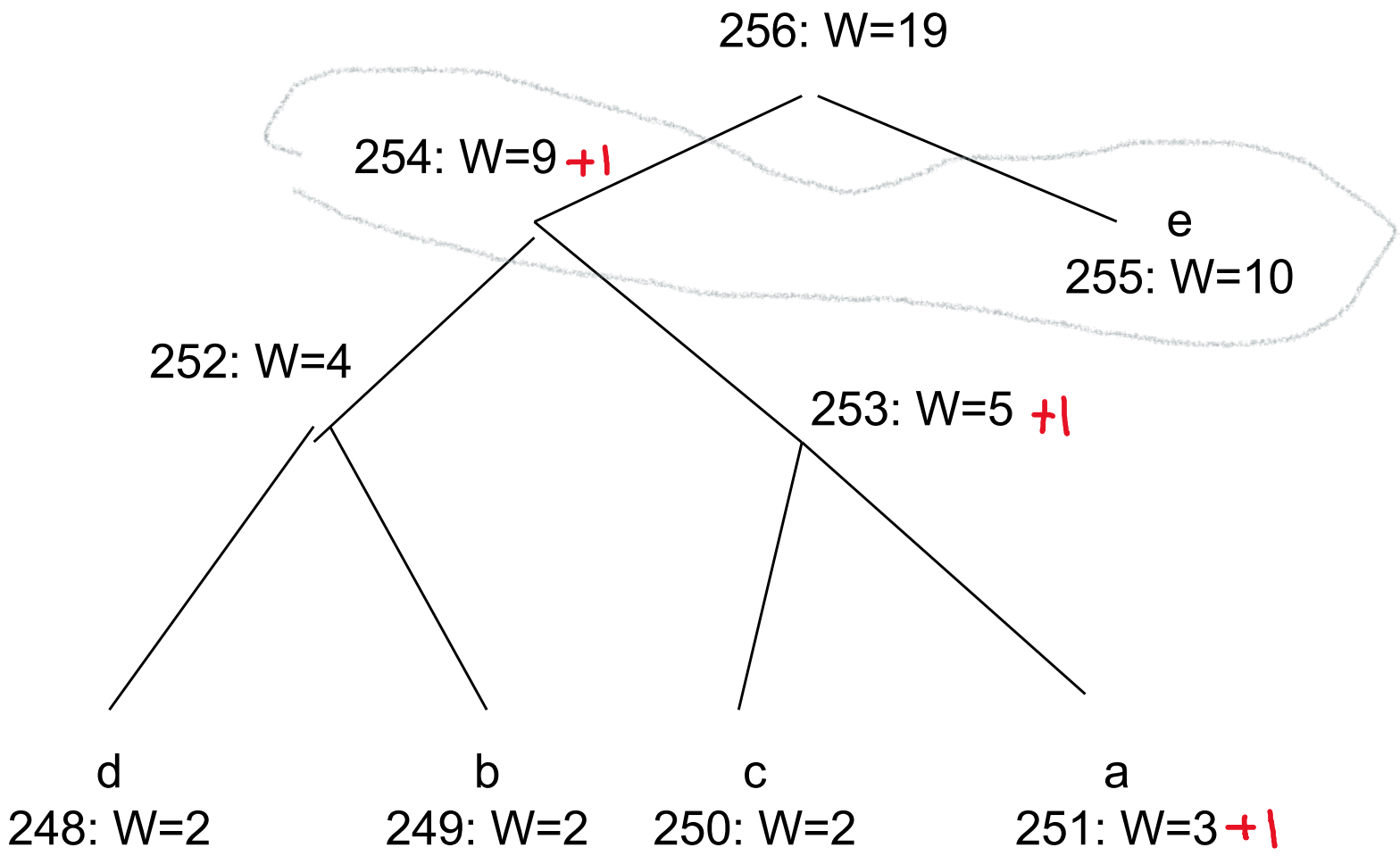
More example



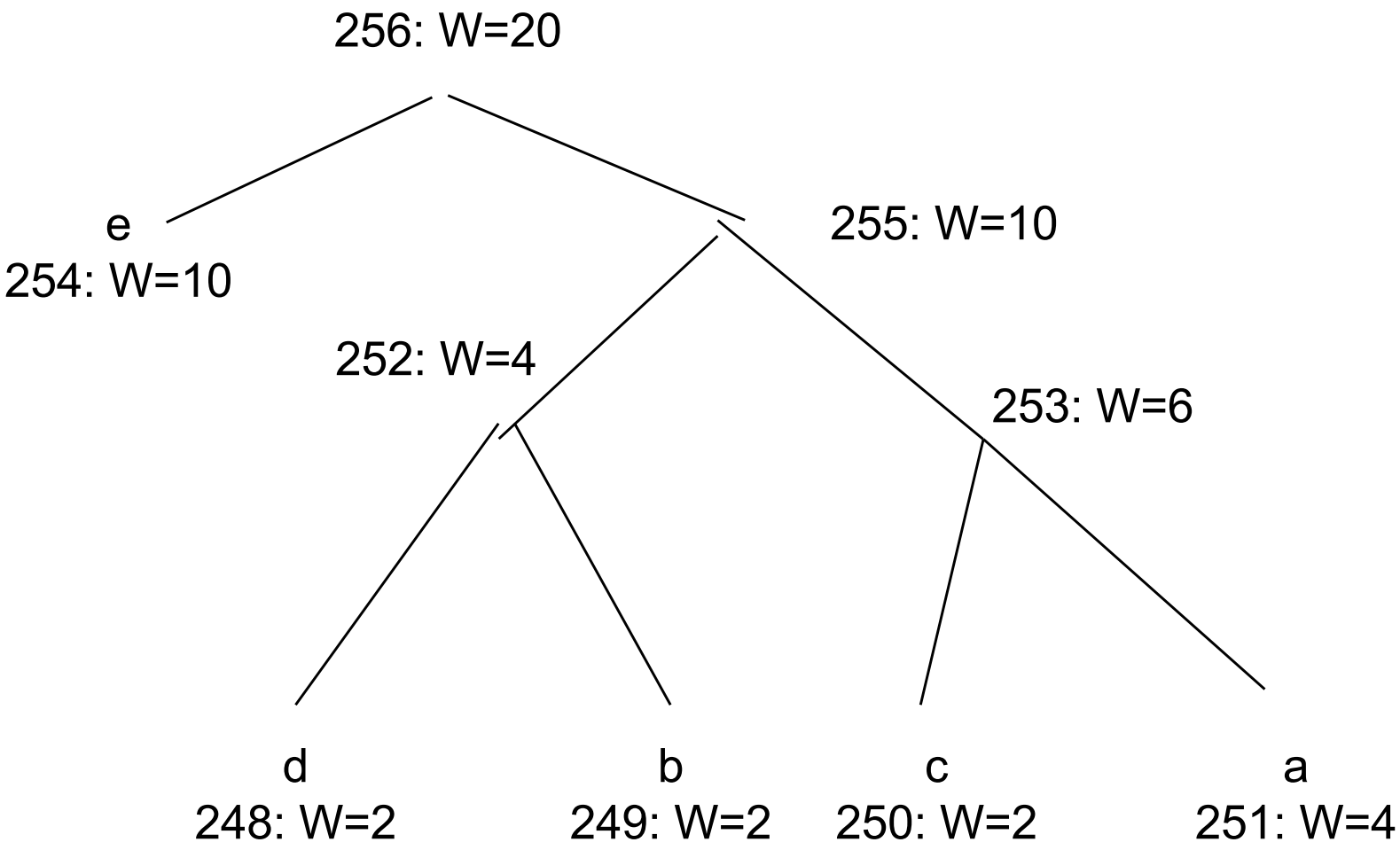
More example



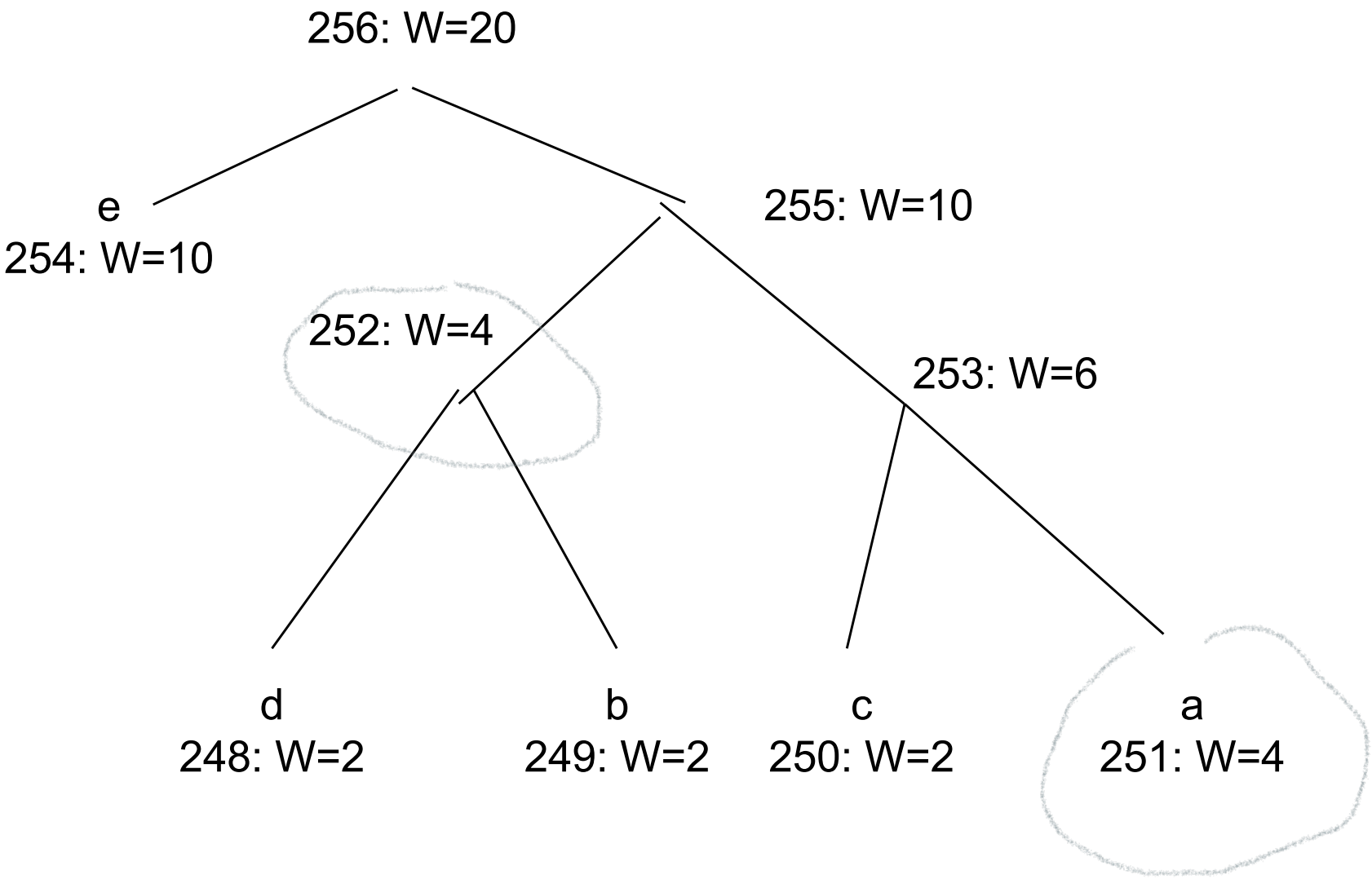
More example



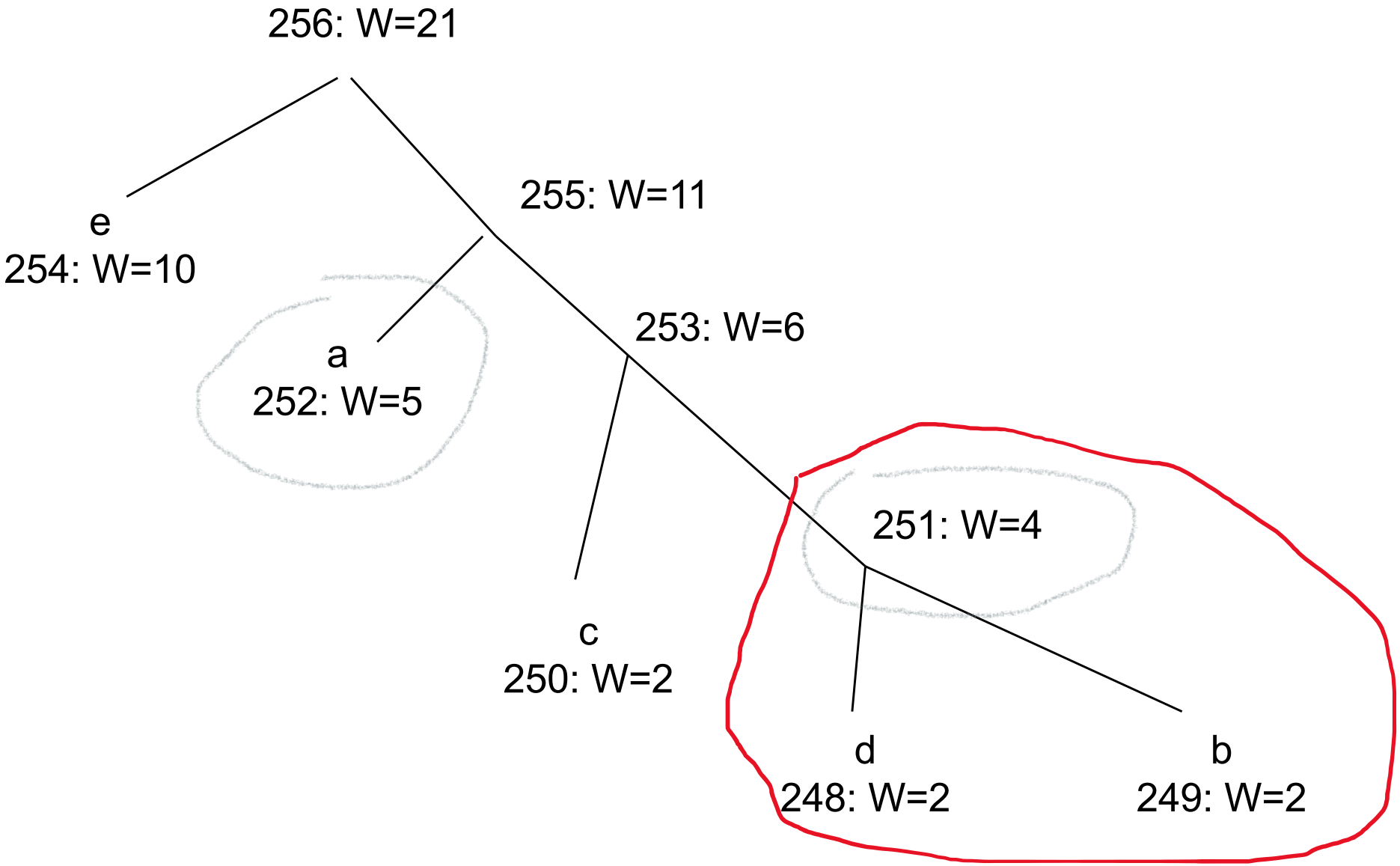
More example



More example



More example



Adaptive Huffman

Question: Adaptive Huffman vs Static Huffman

Compared with Static Huffman

Dynamic and can offer better compression (cf. Vitter's experiments next)

Works when prior stat is unavailable

Saves symbol table overhead (cf. Vitter's expt next)

Vitter's experiments

Include overheads such as symbol tables / leaf node code etc.

t	k	S_t	b/l	$D_t^{\wedge}$	b/l
100	96	664	13.1	569	10.2
500	96	3320	7.9	3225	7.4
960	96	6400	7.1	6305	6.8

Exclude overheads such as symbol tables / leaf node code etc.

95 ASCII chars + <end-of-line>

From Vitter's paper. You know where it is. ☺

More experiments

t	k	S_t	b/l	$D_t^{\wedge}$	b/l
100	34	434	7.1	420	6.3
500	52	2429	5.7	2445	5.5
1000	58	4864	5.3	4900	5.2
10000	74	47710	4.8	47852	4.8
12280	76	58457	4.8	58614	4.8

Basic BWT

(to be discussed more detailed next week)

Recall from Lecture 1's RLE and BWT example

rabcabcababaabacabcabcabcababaa$

aabbbbccacccrcbaaaaaaaaaaabbbbba$

aab4ccac3rcba10b5a$

A simple example

Input:

#BANANAS

All rotations

#BANANAS

S#BANANA

AS#BANAN

NAS#BANA

ANAS#BAN

NANAS#BA

ANANAS#B

BANANAS#

Sort the rows

#BANANAS

ANANAS#B

ANAS#BAN

AS#BANAN

BANANAS#

NANAS#BA

NAS#BANA

S#BANANA

Output

#BANANAS

ANANAS#B

ANAS#BAN

AS#BANAN

BANANAS#

NANAS#BA

NAS#BANA

S#BANANA

Exercise: you can try this example

rabcabcababaabacabcabcabcababaa$

aabbbbccacccrcbaaaaaaaaaaabbbbba$

Now the inverse, for decoding…

Input:

```
S
B
N
N
#
A
A
A
```

First add

S
B
N
N
#
A
A
A

Then sort

#
A
A
A
B
N
N
S

Add again

S#
BA
NA
NA
#B
AN
AN
AS

Then sort

#B
AN
AN
AS
BA
NA
NA
S#

Then add

S#B
BAN
NAN
NAS
#BA
ANA
ANA
AS#

Then sort

#BA
ANA
ANA
AS#
BAN
NAN
NAS
S#B

Then add

S#BA
BANA
NANA
NAS#
#BAN
ANAN
ANAS
AS#B

Then sort

#BAN
ANAN
ANAS
AS#B
BANA
NANA
NAS#
S#BA

Then add

S#BAN
BANAN
NANAS
NAS#B
#BANA
ANANA
ANAS#
AS#BA

Then sort

#BANA
ANANA
ANAS#
AS#BA
BANAN
NANAS
NAS#B
S#BAN

Then add

S#BANA
BANANA
NANAS#
NAS#BA
#BANAN
ANANAS
ANAS#B
AS#BAN

Then sort

#BANAN
ANANAS
ANAS#B
AS#BAN
BANANA
NANAS#
NAS#BA
S#BANA

Then add

S#BANAN
BANANAS
NANAS#B
NAS#BAN
#BANANA
ANANAS#
ANAS#BA
AS#BANA

Then sort

#BANANA
ANANAS#
ANAS#BA
AS#BANA
BANANAS
NANAS#B
NAS#BAN
S#BANAN

Then add

S#BANANA
BANANAS#
NANAS#BA
NAS#BANA
#BANANAS
ANANAS#B
ANAS#BAN
AS#BANAN

Then sort (???)

#BANANAS
ANANAS#B
ANAS#BAN
AS#BANAN
BANANAS#
NANAS#BA
NAS#BANA
S#BANANA