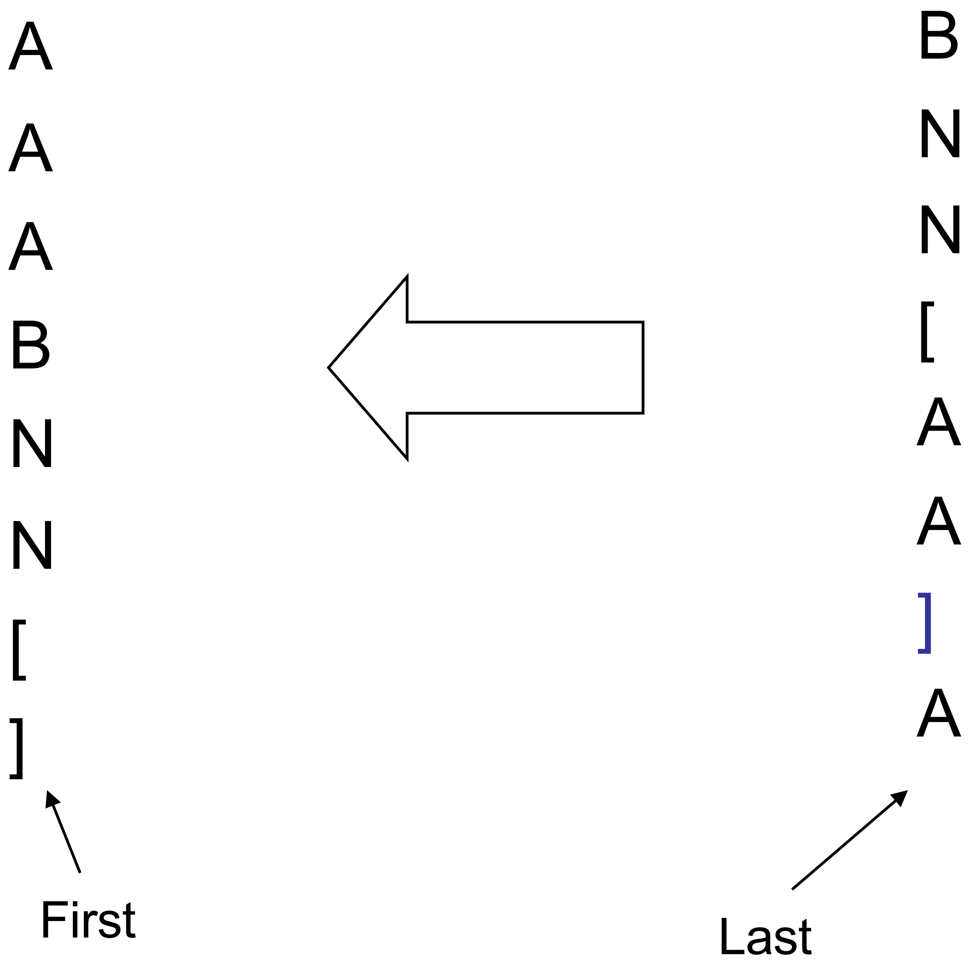
COMP9319 Web Data Compression and Search

BWT, **MTF** and Pattern Matching

BWT

- Burrows–Wheeler transform (BWT) is an algorithm used to prepare data for use with data compression techniques such as bzip2.
- It was invented by Michael Burrows and David Wheeler in 1994 at DEC SRC, Palo Alto, California.
- It is based on a previously unpublished transformation discovered by Wheeler in 1983.

Recall: Last column = BWT



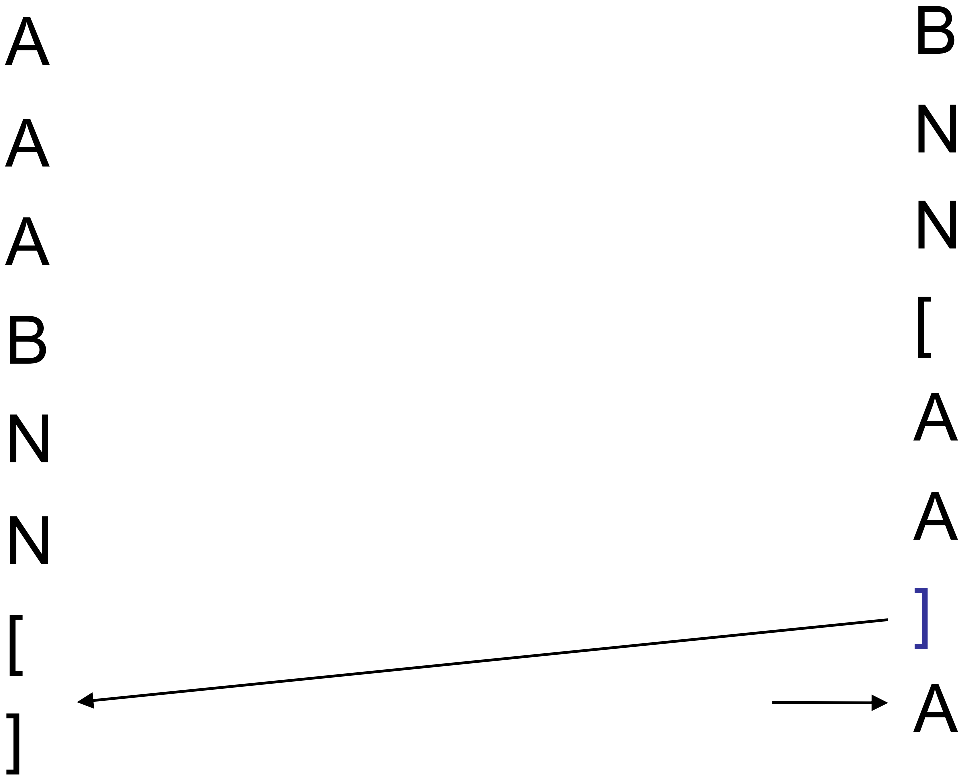
A
A
A
B
N
N
[
]

First

B
N
N
[
A
A
]
A

Last

A]



A

A

A

B

N

N

[

]

B

N

N

[

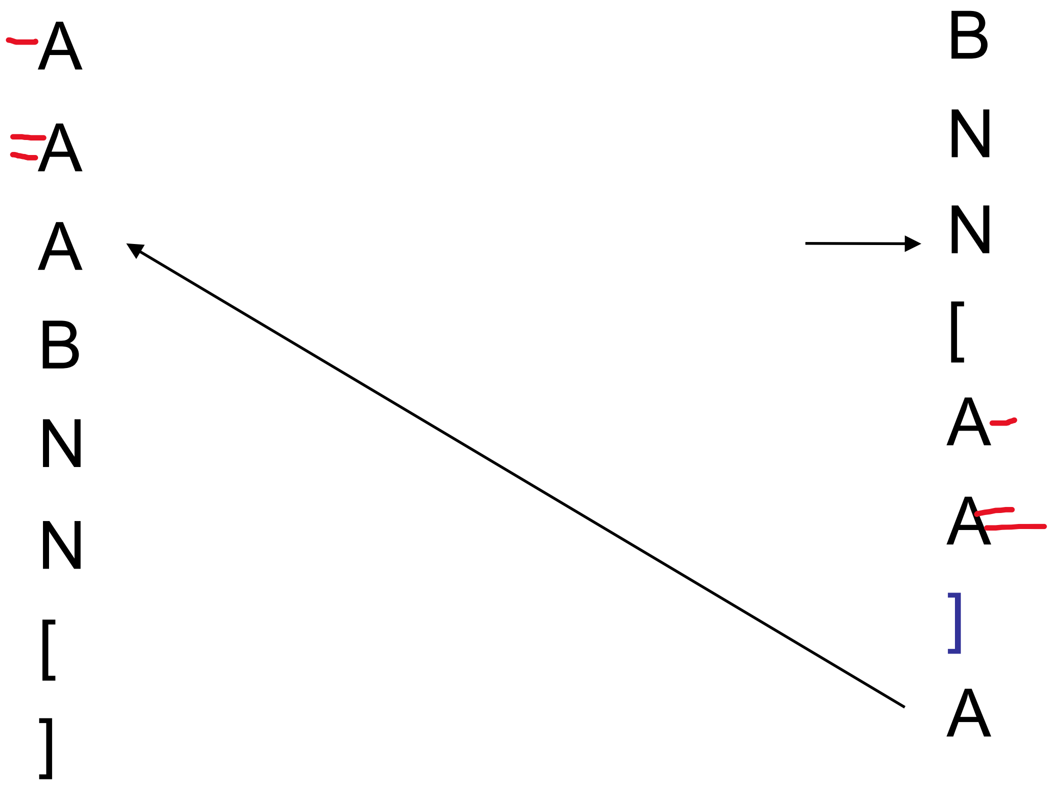
A

A

]

A

NA]



A

A

A

B

N

N

[

]

B

N

N

[

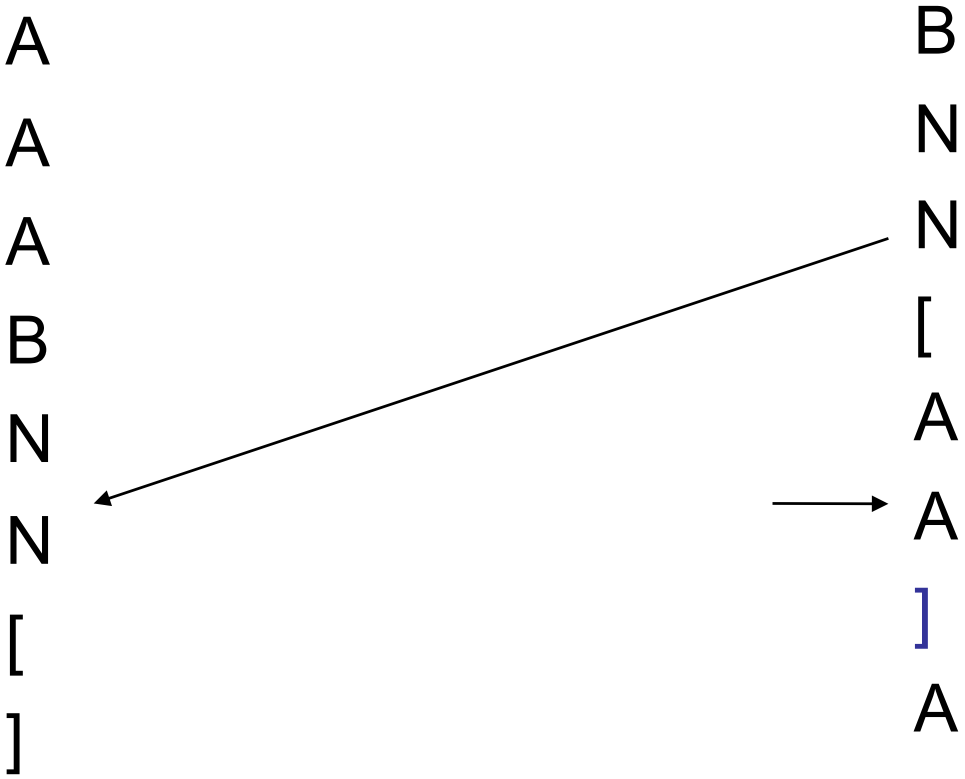
A

A

]

A

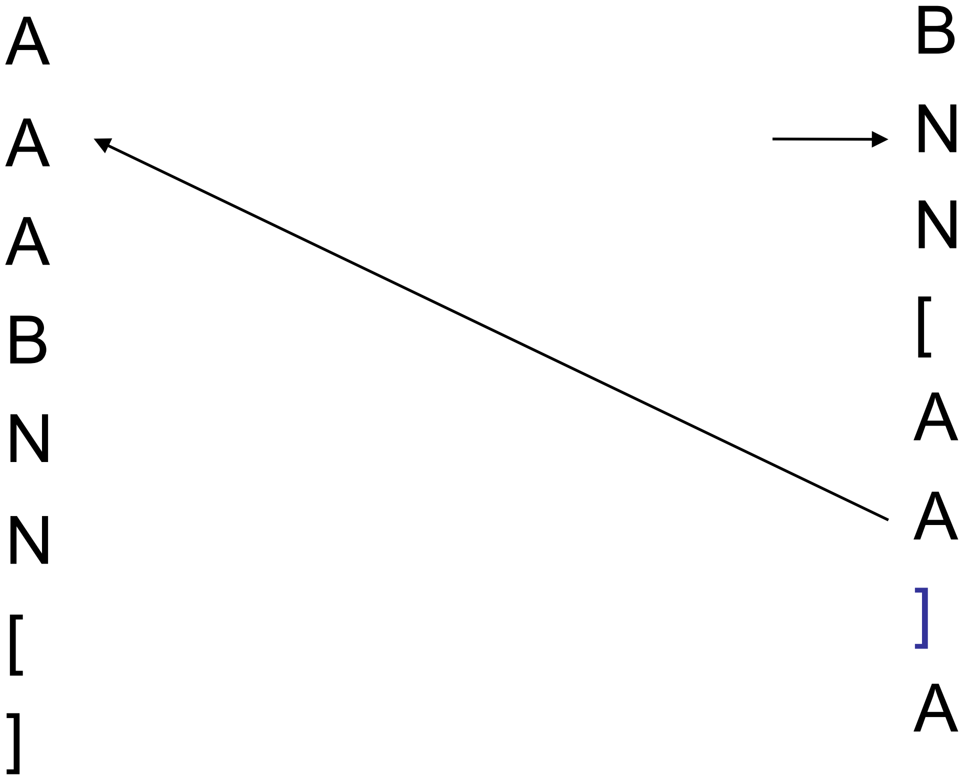
ANA]



A
A
A
B
N
N
[
]

B
N
N
[
A
A
]
A

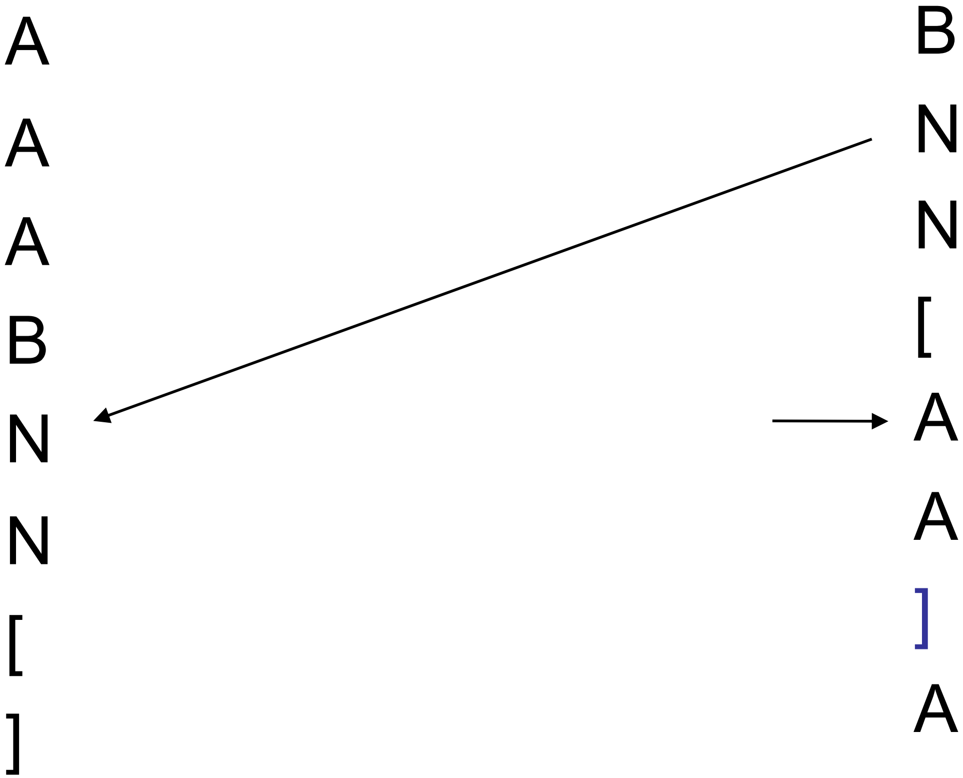
NANA]



A
A
A
B
N
N
[
]

B
N
N
[
A
A
]
A

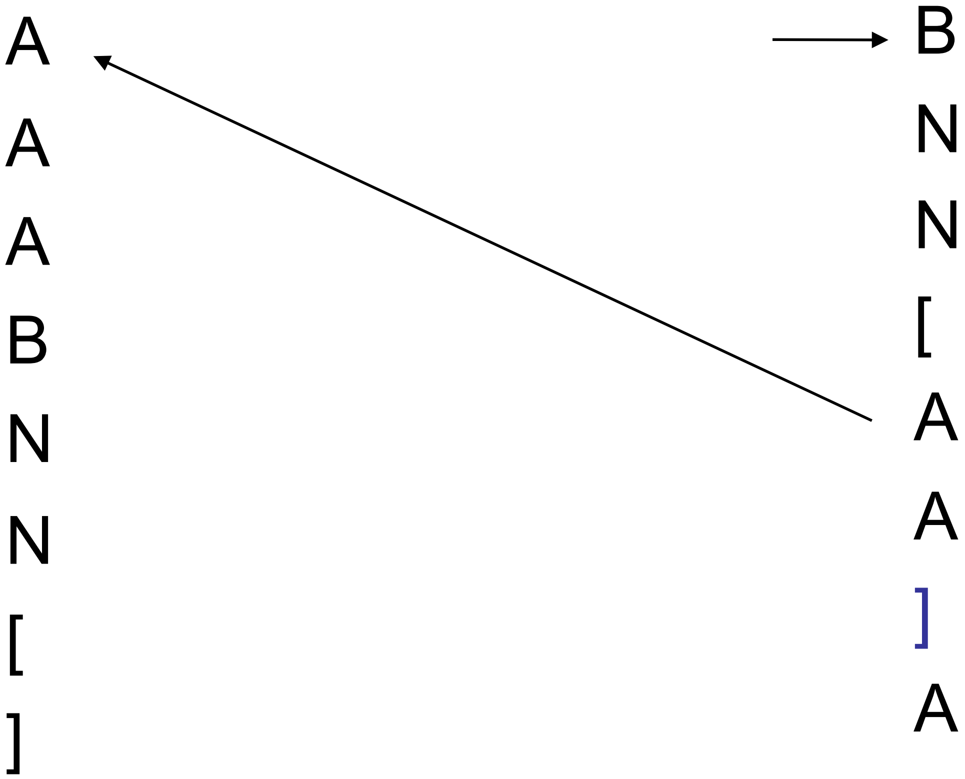
ANANA]



A
A
A
B
N
N
[
]

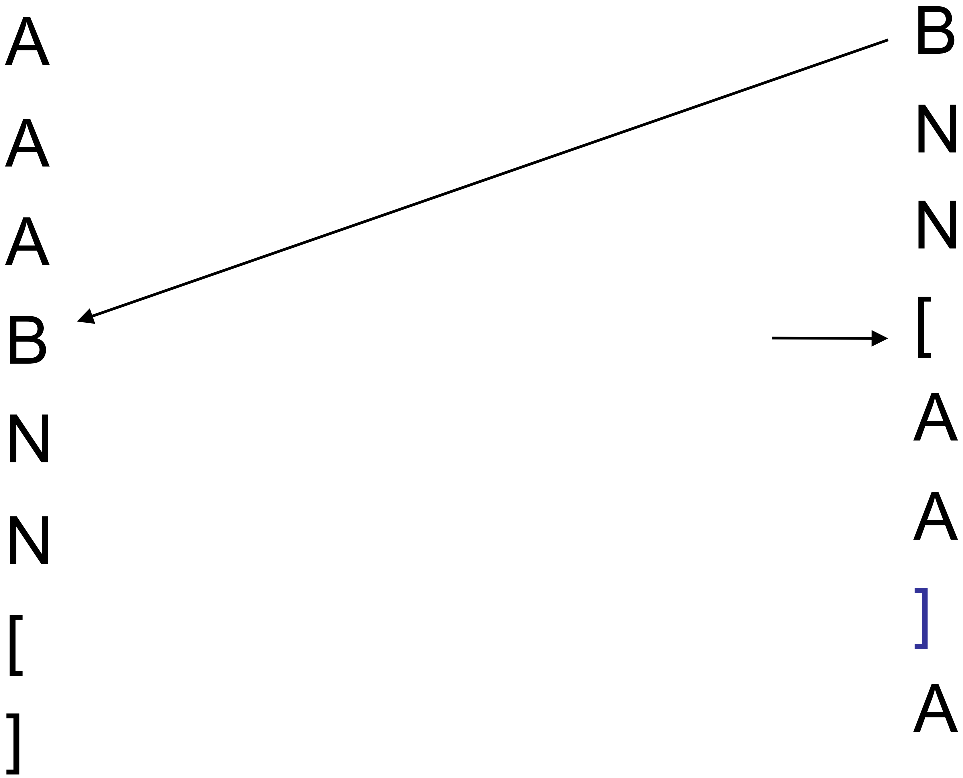
B
N
N
[
A
A
]
A

BANANA]



A	B
A	N
A	N
B	[
N	A
N	A
[	]
]	A

[BANANA]



A	B
A	N
A	N
B	[
N	A
N	A
[	]
]	A

Example using C[ ] & Occ[ ]

Position	Symbol	# Matching
0	B	0
1	N	0
2	N	1
3	[	0
4	A	0
5	A	1
6	]	0
7	A	2

Symbol	# LessThan
A	0
B	3
N	4
[	6
]	7

???????]

Position	Symbol	# Matching
0	B	0
1	N	0
2	N	1
3	[	0
4	A	0
5	A	1
6	]	0
7	A	2

Symbol	# LessThan
A	0
B	3
N	4
[	6
]	7

??????A]

Position	Symbol	# Matching
0	B	0
1	N	0
2	N	1
3	[	0
4	A	0
5	A	1
6	]	0
7	A	2

Symbol	# LessThan
A	0
B	3
N	4
[	6
]	7

?????NA]

Position	Symbol	# Matching
0	B	0
1	N	0
2	N	1
3	[	0
4	A	0
5	A	1
6	]	0
7	A	2

Symbol	# LessThan
A	0
B	3
N	4
[	6
]	7

????ANA]

Position	Symbol	# Matching
0	B	0
1	N	0
2	N	1
3	[	0
4	A	0
5	A	1
6	]	0
7	A	2

Symbol	# LessThan
A	0
B	3
N	4
[	6
]	7

???NANA]

Position	Symbol	# Matching
0	B	0
1	N	0
2	N	1
3	[	0
4	A	0
5	A	1
6	]	0
7	A	2

Symbol	# LessThan
A	0
B	3
N	4
[	6
]	7

??ANANA]

Position	Symbol	# Matching
0	B	0
1	N	0
2	N	1
3	[	0
4	A	0
5	A	1
6	]	0
7	A	2

Symbol	# LessThan
A	0
B	3
N	4
[	6
]	7

?BANANA]

Position	Symbol	# Matching
0	B	0
1	N	0
2	N	1
3	[	0
4	A	0
5	A	1
6	]	0
7	A	2

Symbol	# LessThan
A	0
B	3
N	4
[	6
]	7

[BANANA]

Position	Symbol	# Matching
0	B	0
1	N	0
2	N	1
3	[	0
4	A	0
5	A	1
6	]	0
7	A	2

Symbol	# LessThan
A	0
B	3
N	4
[	6
]	7

[BANANA]

Position	Symbol	# Matching
0	B	0
1	N	0
2	N	1
3	[	0
4	A	0
5	A	1
6	]	0
7	A	2

Occ / Rank

Symbol	# LessThan
A	0
B	3
N	4
[	6
]	7

C [ ]

Move to Front (MTF)

Reduce entropy based on local frequency correlation

Usually used for BWT before an entropy-encoding step

Author and detail:

- Original paper at webcms3
- http://www.arturocampos.com/ac_mtf.html

Example: abaabacad

Symbol	Code	List
a	0	abcde…..
b	1	bacde…..
a	1	abcde…..
a	0	abcde…..
b	1	bacde…..
a	1	abcde…..
c	2	cabde…..
a	1	acbde…..
d	3	dacbe…..

To transform a general file, the list has 256 ASCII symbols.

Example: abaaabbbccddddcc

Symbols: abaaabbbccddddcc
Codes (in ASCII binary): 01100001, 01100010, 01100001, 01100001, ..., 01100100, 01100011, 01100011
Codes (in ASCII dec): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99

Example: abaaabbbccddddcc

Symbols: abaaabbbccddddcc
Codes (in ASCII binary): 01100001, 01100010, 01100001, 01100001, ..., 01100100, 01100011, 01100011
Codes (in ASCII dec): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99

Recall that Shannon's entropy reaches the max when there is max uncertainly, i.e., equal probability, like the example above (4 "97"s, 4 "98"s, 4 "99"s, 4 "100"s).

e.g., Entropy H = 2.00

Example: abaaabbbccddddcc

Symbols: abaaabbbccddddcc
Codes (in ASCII binary): 01100001, 01100010, 01100001, 01100001, ..., 01100100, 01100011, 01100011
Codes (in ASCII dec): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99

List

Index:	0	1	2	3	4	97	98	99	100	101	102	...	255
Value:	..	..	..	..	..	a	b	c	d	e	f	...	..

Codes (in ASCII): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99
Codes (in MTF): 97

Example: abaaabbbccddddcc

Symbols: abaaabbbccddddcc

Codes (in ASCII binary): 01100001, 01100010, 01100001, 01100001, ..., 01100100, 01100011, 01100011

Codes (in ASCII dec): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99

List

Index:	0	1	2	3	4	97	98	99	100	101	102	...	255
Value:	..	..	..	..	..	a	b	c	d	e	f	...	..

Codes (in ASCII): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99

Codes (in MTF): 97

List

Index:	0	1	2	3	4	97	98	99	100	101	102	...	255
Value:	..	..	..	..	..	a	b	c	d	e	f	...	..

move to front

Example: abaaabbbccddddcc

List

Index:	0	1	2	3	4	97	98	99	100	101	102	...	255
Value:	a	..	..	..	..	..	b	c	d	e	f	...	..

Codes (in ASCII): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99

Codes (in MTF): 97, 98

List

Index:	0	1	2	3	4	97	98	99	100	101	102	...	255
Value:	b	a	..	..	..	..	..	c	d	e	f	...	..

Example: abaaabbbccddddcc

List

Index:	0	1	2	3	4	97	98	99	100	101	102 ...	255
Value:	b	a	..	..	..	..	..	c	d	e	f ...	..

Codes (in ASCII): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99

Codes (in MTF): 97, 98, 1

List

Index:	0	1	2	3	4	97	98	99	100	101	102 ...	255
Value:	b	a	..	..	..	..	..	c	d	e	f ...	..

Example: abaaabbbccddddcc

<u>List</u>

Index:	0	1	2	3	4	97	98	99	100	101	102	...	255
Value:	a	b	..	..	..	..	..	c	d	e	f	...	..

Codes (in ASCII): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99

Codes (in MTF): 97, 98, 1, 0,

<u>List</u>

Index:	0	1	2	3	4	97	98	99	100	101	102	...	255
Value:	a	b	..	..	..	..	..	c	d	e	f	...	..

Example: abaaabbbccddddcc

List

Index:	0	1	2	3	4	97	98	99	100	101	102	...	255
Value:	a	b	..	..	..	..	..	c	d	e	f	...	..

Codes (in ASCII): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99

Codes (in MTF): 97, 98, 1, 0, 0, 1, 0, 0, 99, 0, 100, 0, 0, 0, 1, 0

Example: MTF decoding

List

Index:	0	1	2	3	4	97	98	99	100	101	102	...	255
Value:	..	..	..	..	..	a	b	c	d	e	f	...	..

Codes (in MTF): 97, 98, 1, 0, 0, 1, 0, 0, 99, 0, 100, 0, 0, 0,
1, 0

Symbols: **a**, **b**

List

Index:	0	1	2	3	4	97	98	99	100	101	102	...	255
Value:	b	a	..	..	..	..	..	c	d	e	f	...	..

Example: MTF decoding

List

Index:	0	1	2	3	4	97	98	99	100	101	102	...	255
Value:	b	a	..	..	..	..	..	c	d	e	f	...	..

Codes (in MTF): 97, 98, 1, 0, 0, 1, 0, 0, 99, 0, 100, 0, 0, 0,
1, 0
Symbols: **a**, **b, a**

List

Index:	0	1	2	3	4	97	98	99	100	101	102	...	255
Value:	a	b	..	..	..	..	..	c	d	e	f	...	..

Example: MTF decoding

List

Index:	0	1	2	3	4	97	98	99	100	101	102	...	255
Value:	a	b	..	..	..	..	..	c	d	e	f	...	..

Codes (in MTF): 97, 98, 1, 0, 0, 1, 0, 0, 99, 0, 100, 0, 0, 0,
1, 0
Symbols: **a**, **b, a, a**

List

Index:	0	1	2	3	4	97	98	99	100	101	102	...	255
Value:	a	b	..	..	..	..	..	c	d	e	f	...	..

Example: MTF decoding

List

Index:	0	1	2	3	4	97	98	99	100	101	102	...	255
Value:	a	b	..	..	..	..	..	c	d	e	f	...	..

Codes (in MTF): 97, 98, 1, 0, 0, 1, 0, 0, 99, 0, 100, 0, 0, 0,
1, 0

Symbols: **a**, **b, a, a, a, b, b, b, c, c, d, d, d, d,**
c, c

Example: MTF decoding

List

Index:	0	1	2	3	4	97	98	99	100	101	102	...	255
Value:	a	b	..	..	..	..	..	c	d	e	f	...	..

Codes (in MTF): 97, 98, 1, 0, 0, 1, 0, 0, 99, 0, 100, 0, 0, 0,
1, 0
Symbols: **a, b, a, a, a, b, b, b, c, c, d, d, d, d,**
c, c

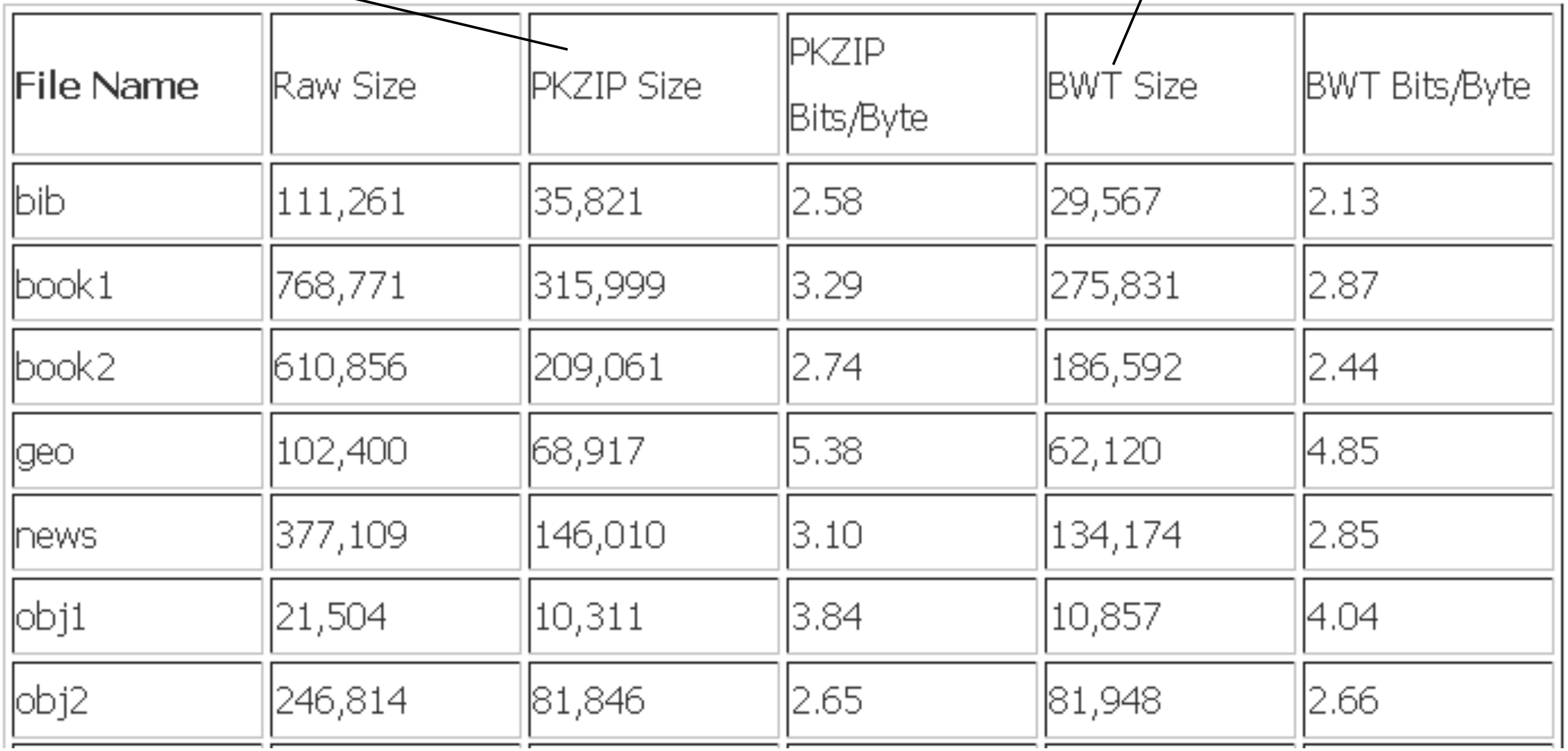
The distribution of symbols is changed, with more *local* references (1 “97”, 1 “98”, 1 “99”, 1 “100”, 9 “0”s, 3 “1”s). => Reduced entropy

H = 1.92

BWT compressor vs ZIP

ZIP (i.e., LZW based)

BWT+RLE+MTF+AC



File Name	Raw Size	PKZIP Size	PKZIP Bits/Byte	BWT Size	BWT Bits/Byte
bib	111,261	35,821	2.58	29,567	2.13
book1	768,771	315,999	3.29	275,831	2.87
book2	610,856	209,061	2.74	186,592	2.44
geo	102,400	68,917	5.38	62,120	4.85
news	377,109	146,010	3.10	134,174	2.85
obj1	21,504	10,311	3.84	10,857	4.04
obj2	246,814	81,846	2.65	81,948	2.66

From http://marknelson.us/1996/09/01/bwt/