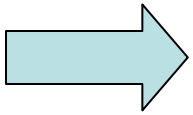
2nd Workshop on Compression, Text, and Algorithms 2007

Santiago de Chile. Nov 1st, 2007

A Compressed Self-Index on Words

Databases Laboratory
University of A Coruña (Spain)
+ Gonzalo Navarro

Outline



- Introduction
- Some previous Concepts
- **W**ord **C**ompressed **S**uffix **A**rray
  - Experimental results
- Presentation Layer + WiCSA
  - Motivation
  - Structure
  - Preliminary Results

Introduction

- Very large text collections of data
  - Need of IR
  - Aiming at:
    - Reducing space needs: compression
    - Improving efficiency of retrieval

- String matching
  - Searches that imply a sequential scan.
    - Over compressed text? → **0,3** · |T|
  - Indexed searches.
    - Inverted indexes, suffix arrays, …

T = M**IS**S**IS**SIPPI

P = **IS**

Introduction

*A Word-Oriented **Self-index***

- Traditional index (with or without compression)



COUNT & **LOCATE** operations

- Self-index



COUNT, **LOCATE** & **DISPLAY** operations

Introduction

*A **Word-Oriented** Self-index*

- Full-text indexes for Nat. Lang (English) text…
  - Are they actually needed?
    - Flexible and powerful
    - Reasonable small AF-index, LZ-index … (pizza&chili)
  - But…
    - a word-oriented index has less symbols
      - Smaller ?
      - Faster ?
      - Poses enough flexibility (i.e. inverted indexes)
  - So lets try a word-oriented approach

Introduction

WCSA? Presentation layer + WCSA ?

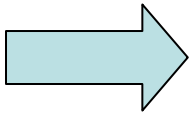
- WCSA?
  - Self-Index based on Sadakane's CSA
    - Word-based Suffix array vs char-based.

- Presentation layer + W_iCSA?
  - How to retrieve documents containing …

    "time", "Time", "times", "TIME", "tiME" …

    1. Indexing canonical words (*Separators & stopwords are not indexed*)
    2. Adding a presentation layer that permits to recover the source text

Outline

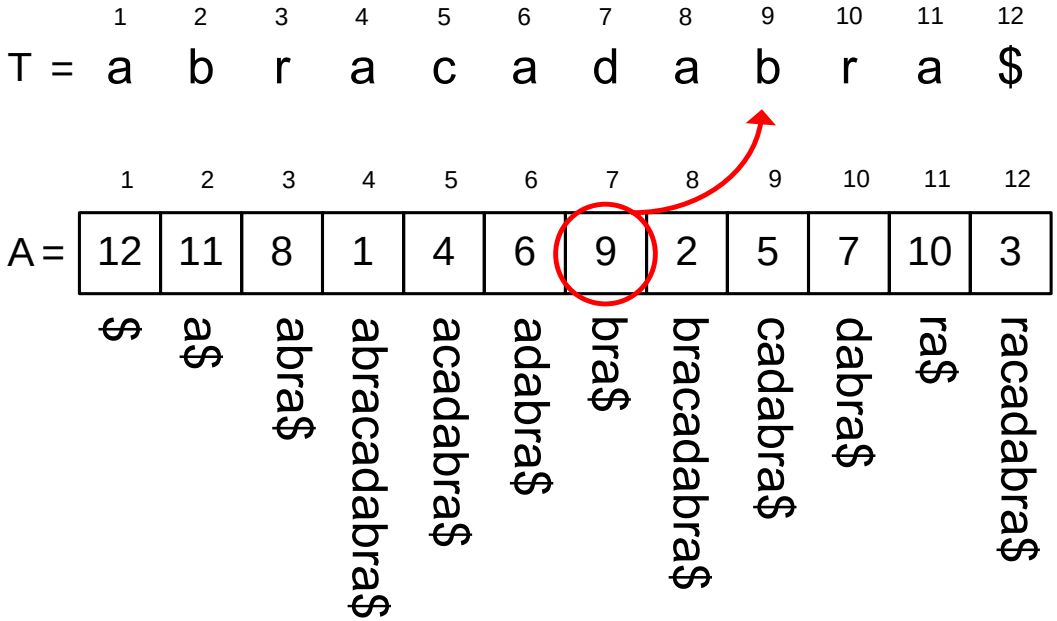


- Introduction
- Some previous Concepts
- **W**ord **C**ompressed **S**uffix **A**rray
  - Experimental results
- Presentation Layer + WiCSA
  - Motivation
  - Structure
  - Preliminary Results

Previous concepts

Suffix Array

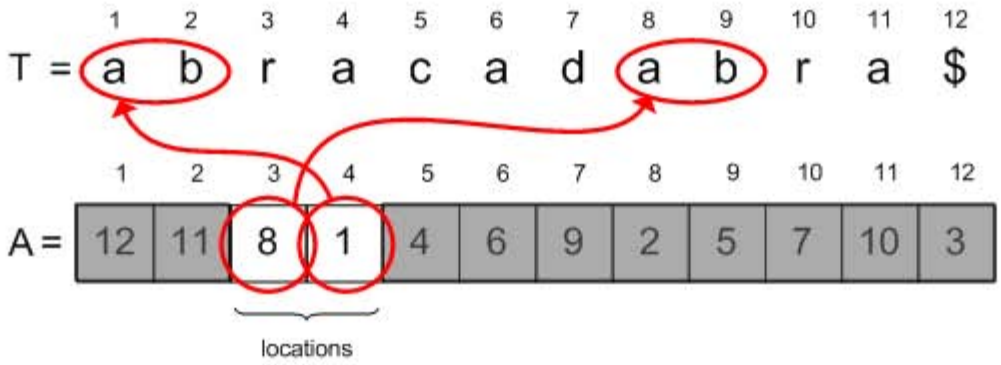
– Suffixes of the text sorted alphabetically



Previous concepts

Suffix Array

– Binary search: i.e Pattern “ab”

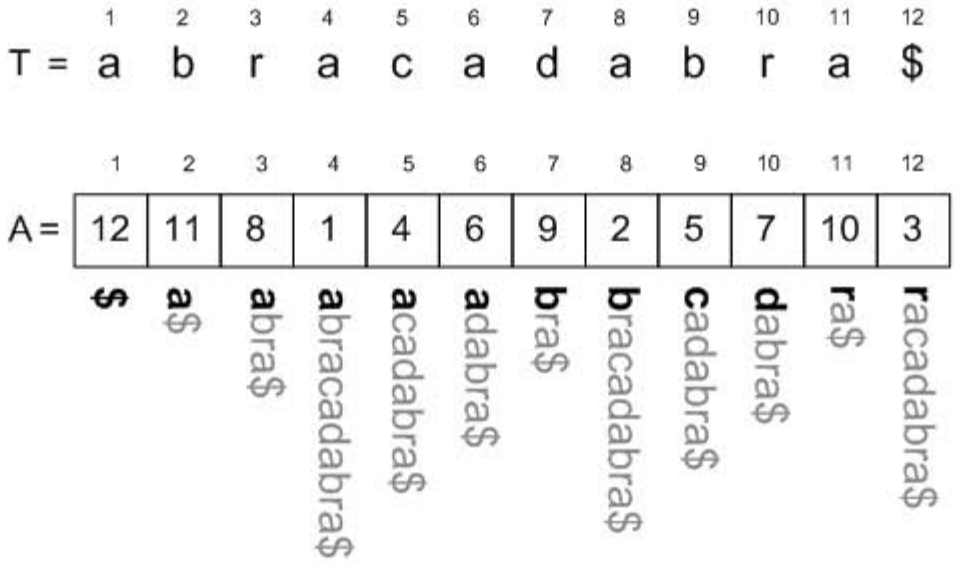


Fast

Space

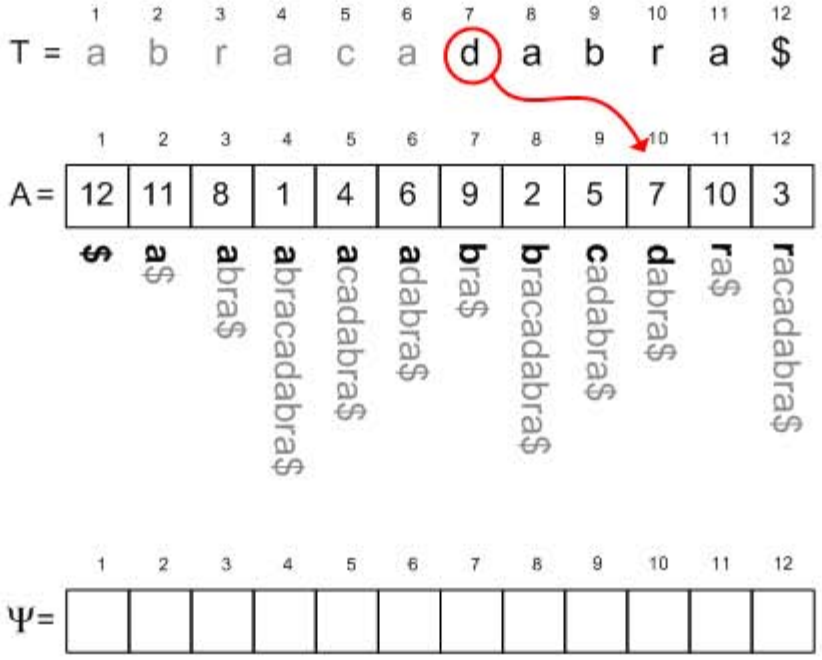
Conceptos Previos

Sadakane's CSA



Previous concepts

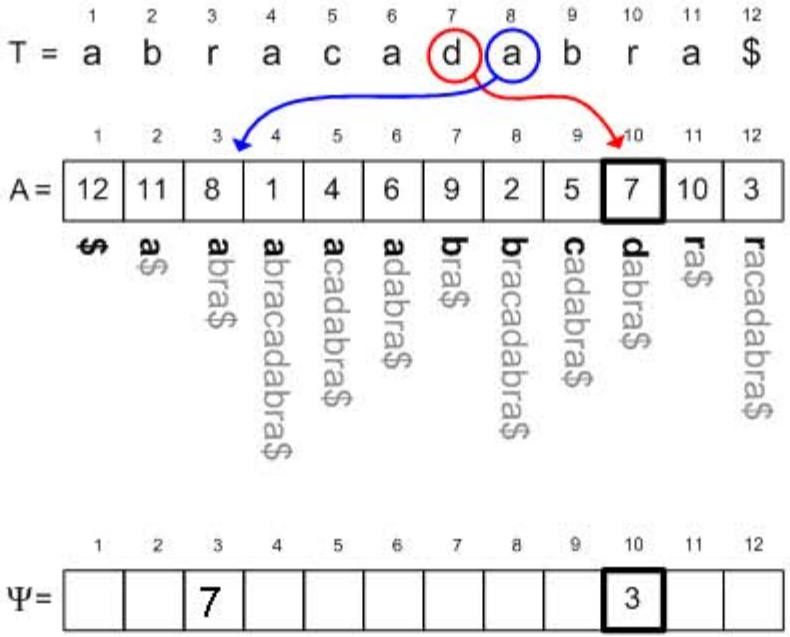
Sadakane's CSA: Ψ function



– ***A[Ψ(i)] = A[i] +1***

Previous concepts

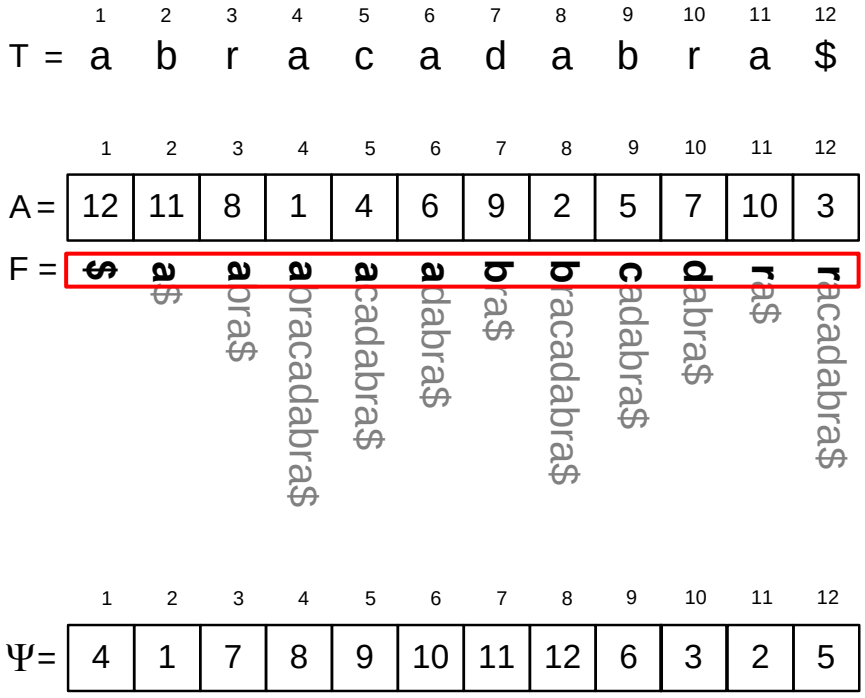
Sadakane's CSA: Ψ function



– ***A[Ψ(10)] = A[3] = A[10] +1 = 8***

Previous concepts

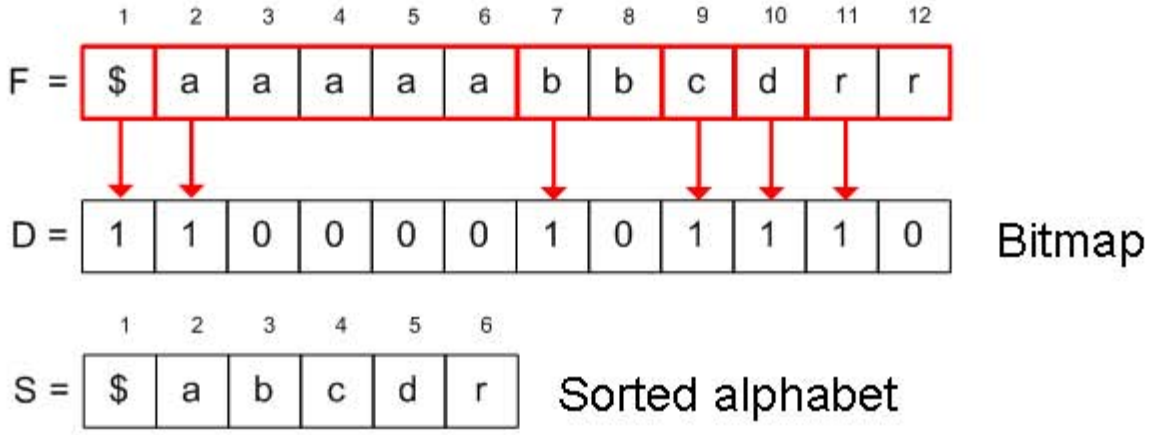
Sadakane's CSA: Ψ & F



– ***Ψ and F suffices for performing binary search***

Previous concepts

Sadakane's CSA: Ψ & F



- *Example:* $F[8] = S[rank_1(D, 8)] = S[3] = b$

Previous concepts

Sadakane's CSA: Summary

	1	2	3	4	5	6	7	8	9	10	11	12
T =	a	b	r	a	c	a	d	a	b	r	a	$

	1	2	3	4	5	6	7	8	9	10	11	12
A =	12	11	8	1	4	6	9	2	5	7	10	3
A^{-1} =	4	8	12	5	9	6	10	3	7	11	2	1
Ψ =	4	1	7	8	9	10	11	12	6	3	2	5
D =	1	1	0	0	0	0	1	0	1	1	1	0

S =	$	a	b	c	d	r

⇐ sampled

⇐ sampled

⇐ Absolute samples + gap encoding
- delta codes*
- Huffman-based

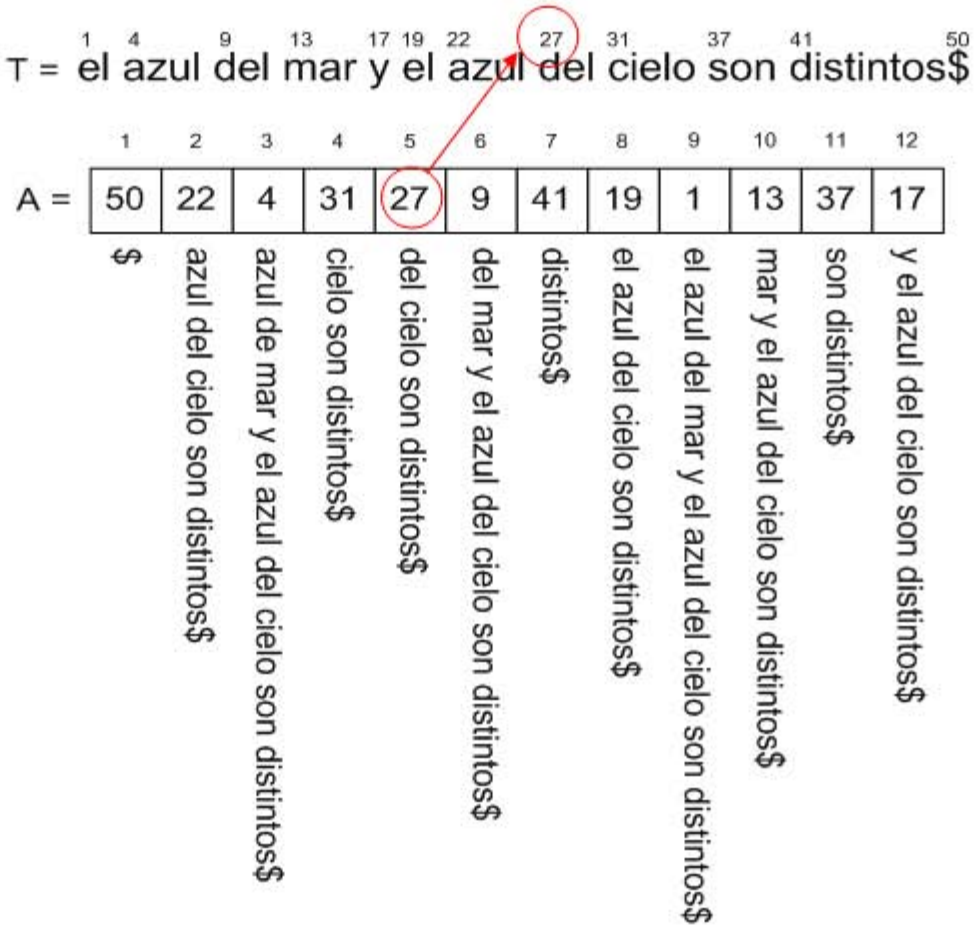
- ***Ψ, D, S*** → ***count***
- ***A*** → ***locate***
- ***A^{-1}*** → ***display***

Outline

- Introduction
- Some previous Concepts
- **W**ord **C**ompressed **S**uffix **A**rray
  - Experimental results
- Presentation Layer + WiCSA
  - Motivation
  - Structure
  - Preliminary Results

Word Compressed Suffix Array

A Suffix array on words (WSA)



- ***As expected, just indexing words instead of characters***
- ***Less suffixes need to be indexed → A is smaller***

Word **C**ompressed **S**uffix **A**rray

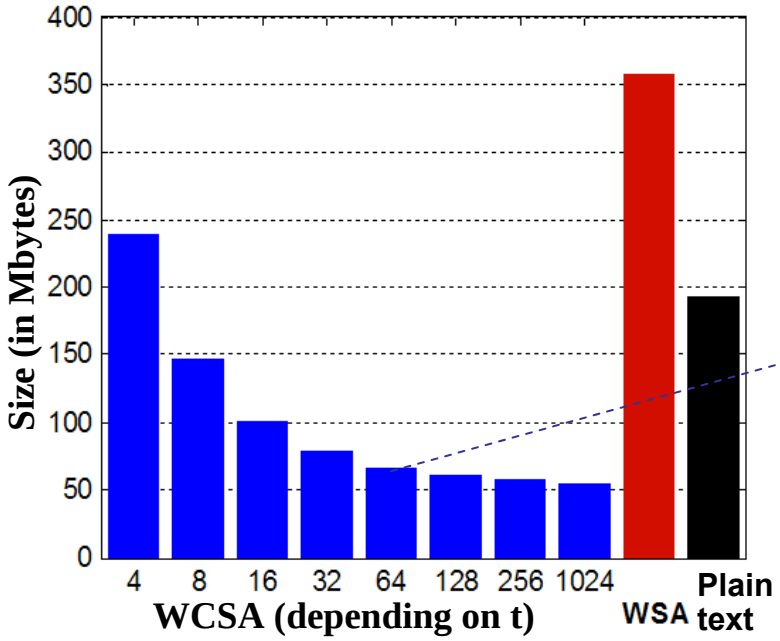
Word-based CSA (WCSA)

- **sA** Samples of A
- **sA^{-1}** Samples of A^{-1}
- **S** **Vocabulary of words** sorted alphabetically
- **D** Bitmap
- **Ψ** psi array (with absolute samples & gap encoding)

Word Compressed Suffix Array

Some experimental results: Space

- **WCSA – WSA**
  - *Financial Times 1994*, from *TREC-4 (200Mb English text)*
  - Pentium®IV@3.0GHz, 1GB, Debian (kernel 2.4.27), gcc 3.3.5 (-O9)

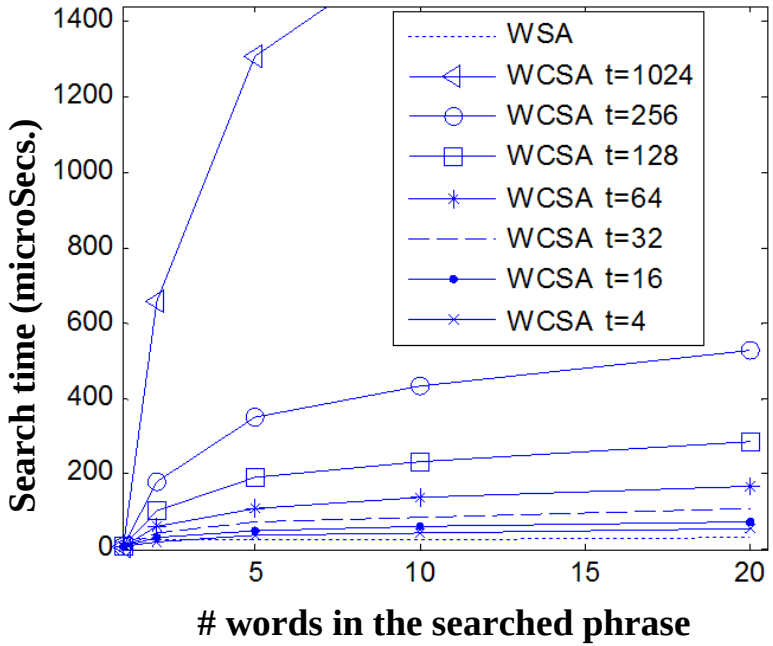


	WCSA (t = sample period)							WSA
	t=4	t=16	t=32	t=64	t=128	t=256	t=1024	
Mbytes	240,6	101,4	78,2	66,6	60,8	57,5	55,7	358,5
% cr	123,8	52,2	40,2	34,3	31,3	29,6	28,7	184,5

Word Compressed Suffix Array

Some preliminary experimental results: count time

- **Counting** occurrences of phrases with #words each
  - user-time in microseconds.



# words	WCSA (t = sample period)							WSA
	t=4	t=16	t=32	t=64	t=128	t=256	t=1024	
1	7	7	7	7	7	7	7	31
2	22	30	40	60	100	180	660	27
5	36	51	71	110	191	351	1308	27
10	43	62	85	136	234	435	1617	28
20	52	75	105	164	284	526	1975	29

- 30 - 40% space
- Not optimized but still good results

Outline

- Introduction
- Some previous Concepts
- **W**ord **C**ompressed **S**uffix **A**rray
  - Some experimental results
- Presentation Layer + WiCSA
  - Structures
  - Preliminary Results

Presentation Layer + WiCSA

Motivation

- Self-indexes on words (i.e. WCSA)
  - Good compression
  - Enough flexibility
- Indexing only “valid-words” (neither stopwords nor separators)
  - Number of indexed words much smaller
  - Searches refer only to meaningful words (in a phrase)
    - “time series” ← “time series” “time-series” “time of series”
- Indexing a **canonical** form of a word
  - “time” ← “Time”, “TIME”, “Times”,…
  - Searches obtain all the **variants** at once
- A sort of **presentation layer** permit us to recover the source text

Presentation Layer + WiCSA

General Structure

Source text

The	Blue	Color	of the	sea	and the	blue	COLOR	of the	sky	are	different	colors
1	5	10	16	19	23	27	31	35	40	46	49	53	57	61	71

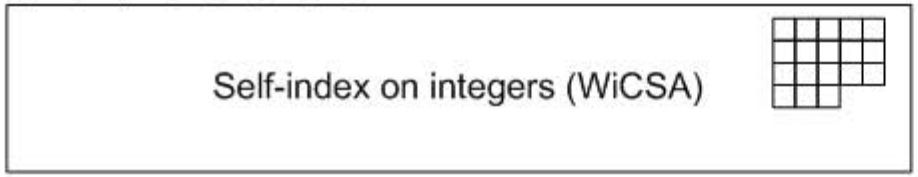


Canonical words mapping

$	blue	color	different	sea	sky
1	2	3	4	5	6

Sequence of ids of valid words

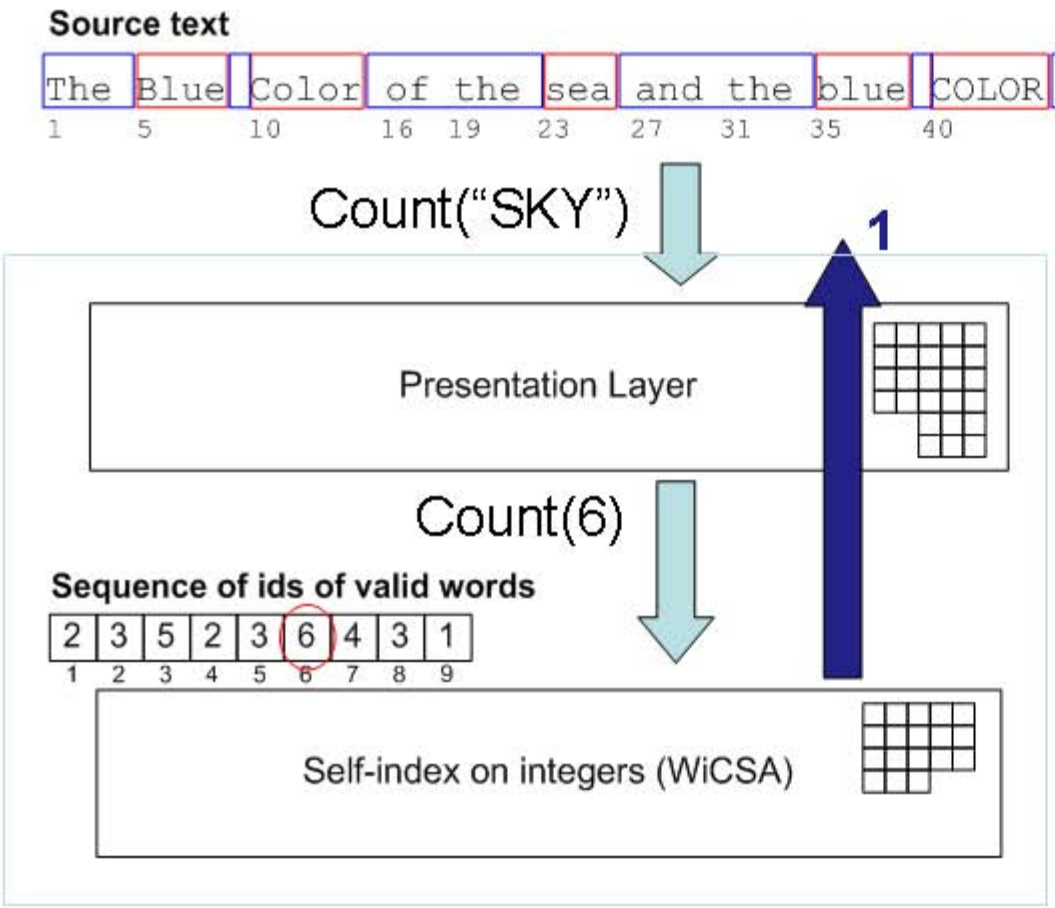
2	3	5	2	3	6	4	3	1
1	2	3	4	5	6	7	8	9



- Presentation Layer replaces the source text
- WiCSA replaces the sequence of integers of the “valid” words

Presentation Layer + WiCSA

General Structure: Count

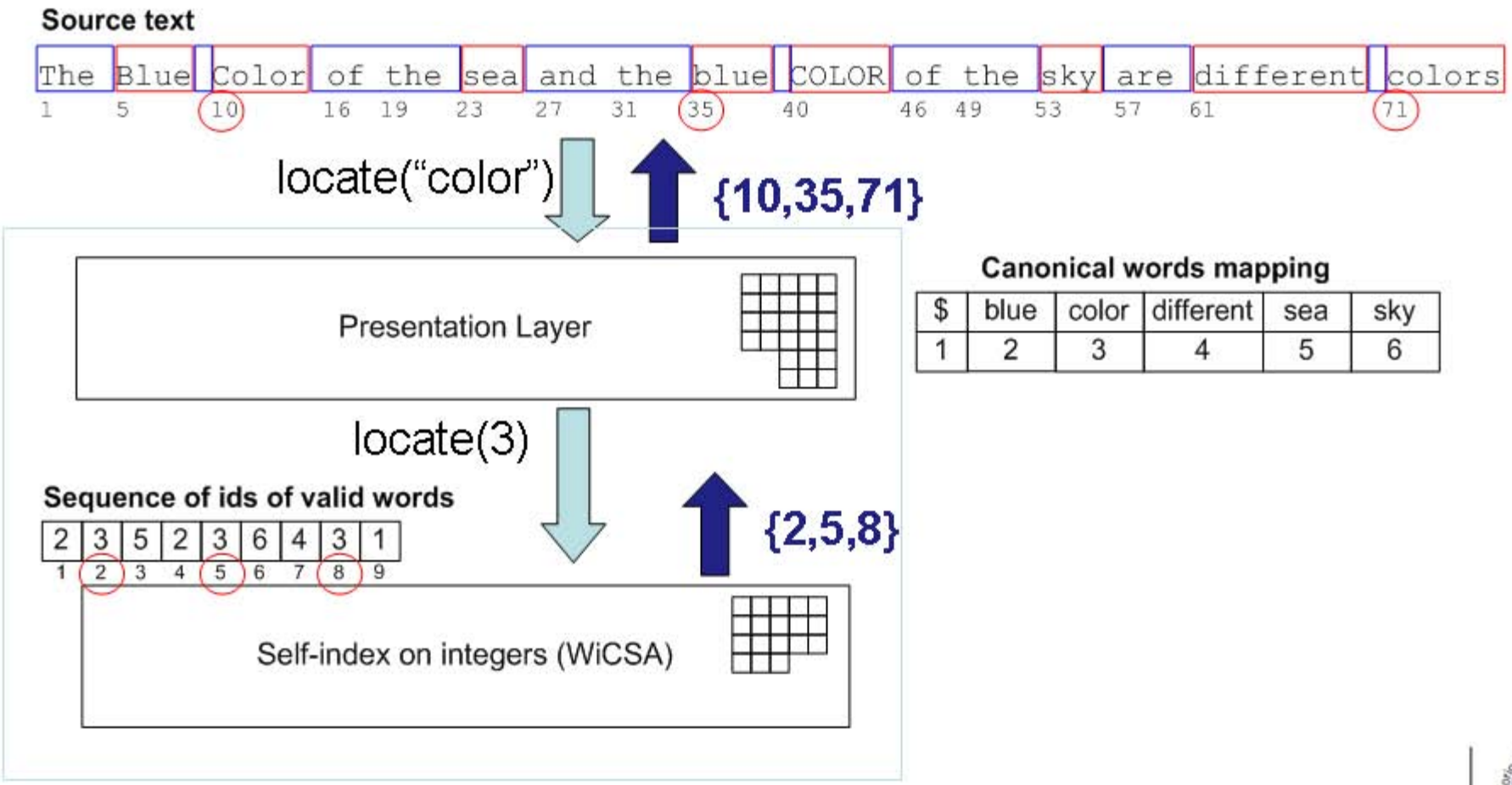


Canonical words mapping

$	blue	color	different	sea	sky
1	2	3	4	5	6

Presentation Layer + WiCSA

General Structure: locate

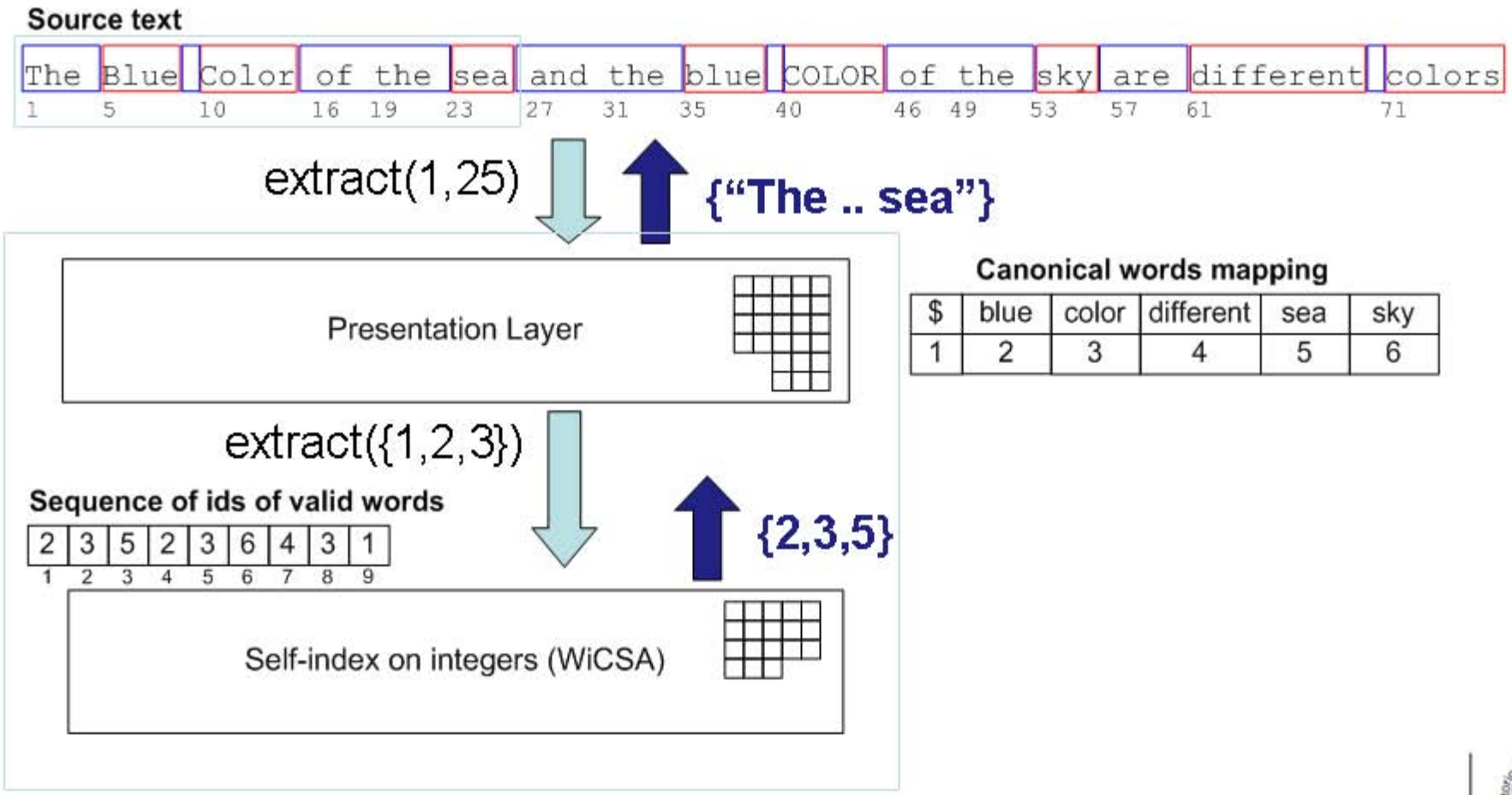


Canonical words mapping

$	blue	color	different	sea	sky
1	2	3	4	5	6

Presentation Layer + WiCSA

General Structure: extract / display

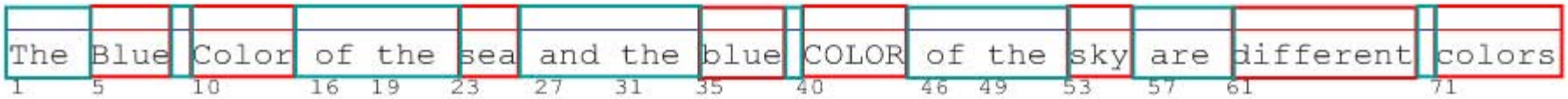


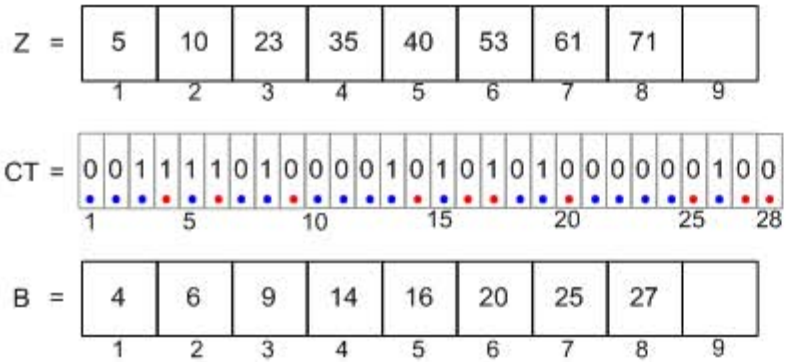
Canonical words mapping

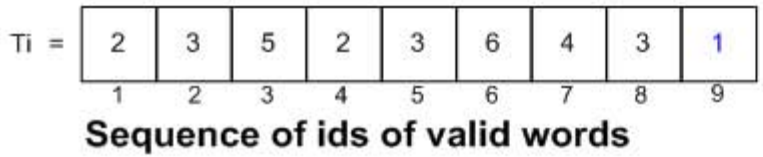
$	blue	color	different	sea	sky
1	2	3	4	5	6

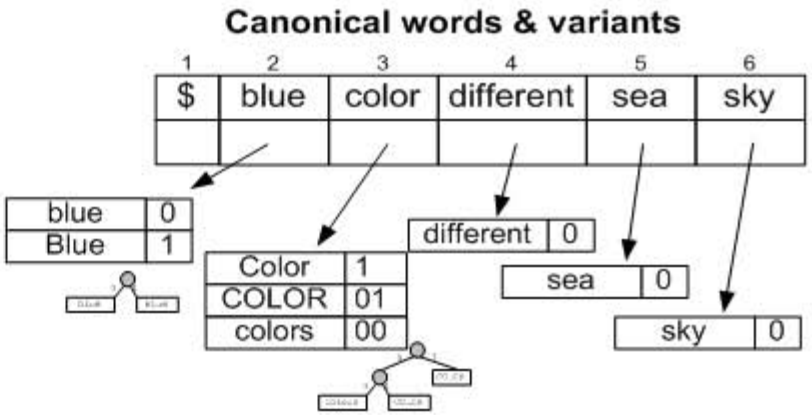
Presentation Layer + WiCSA

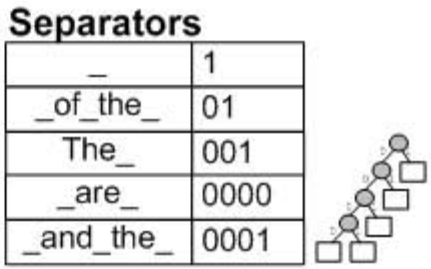
Presentation Layer





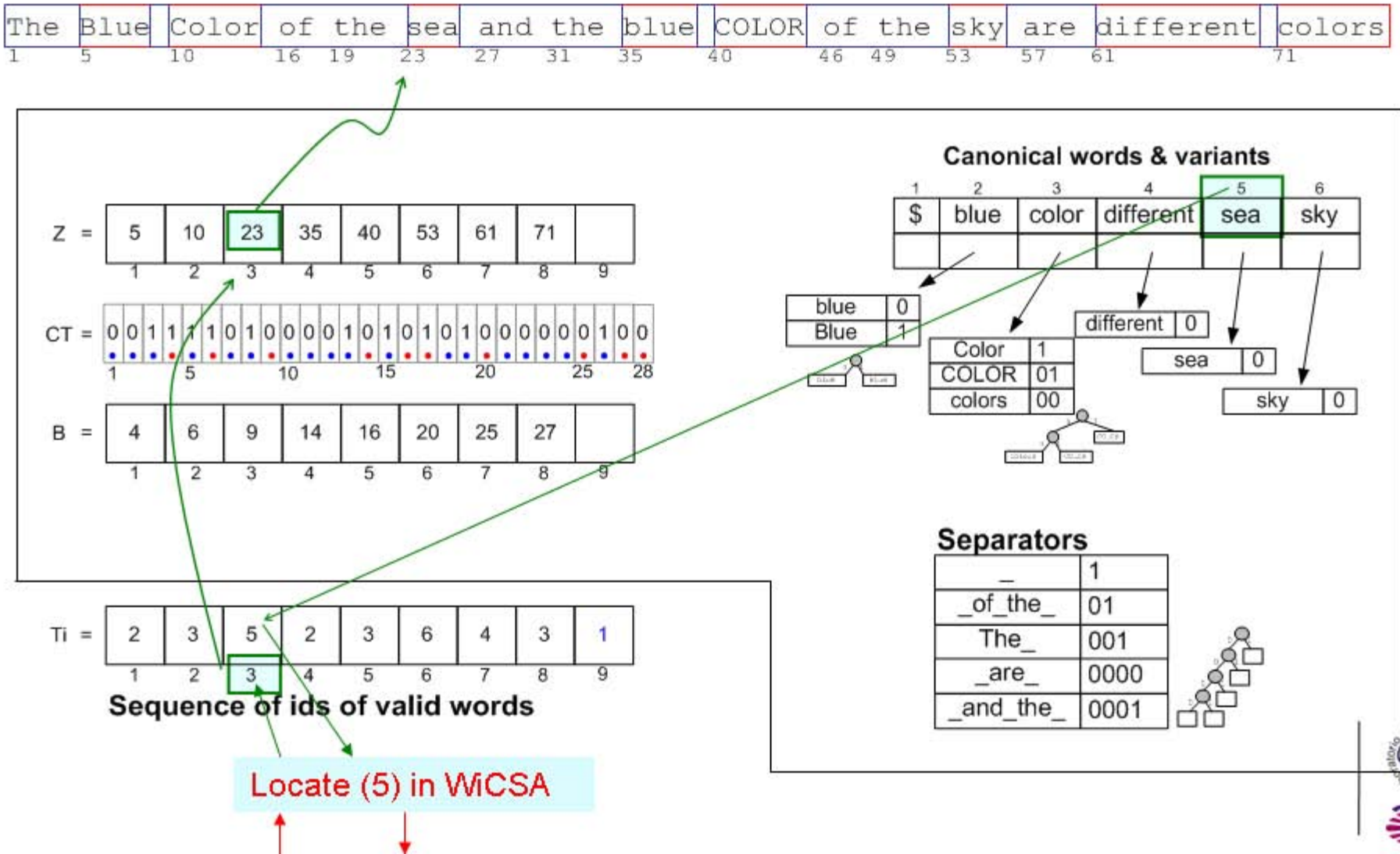






Presentation Layer + WiCSA

Presentation Layer: locate("SEA")



Presentation Layer + WiCSA

Presentation Layer: display("sea",13)

The	Blue		Color	of	the	sea	and	the	blue		COLOR	of	the	sky	are	different		colors
1	5		10		16	19	23	27	31	35	40		46	49	53	57	61		71

Z =

5	10	23	35	40	53	61	71	
1	2	3	4	5	6	7	8	9

CT =

0	0	1	1	1	1	0	1	0	0	0	0	1	0	1	0	1	0	1	0	0	0	0	0	0	1	0	0
1				5					10					15					20					25				28

B =

4	6	9	14	16	20	25	27	
1	2	3	4	5	6	7	8	9

Ti =

2	3	5	2	3	6	4	3	1
1	2	3	4	5	6	7	8	9

Sequence of ids of valid words

Canonical words & variants

1	2	3	4	5	6
$	blue	color	different	sea	sky

blue	0
Blue	1

Color	1
COLOR	01
colors	00

different	0

sea	0

sky	0

Separators

_	1
_of_the_	01
The_	001
are	0000
_and_the_	0001

Presentation Layer + WiCSA

Presentation Layer: extract(9,25)

The	Blue	Color	of	the	sea	and	the	blue	COLOR	of	the	sky	are	different	colors
1	5	10	16	19	23	27	31	35	40	46	49	53	57	61	71

Z =

5	10	23	35	40	53	61	71	
1	2	3	4	5	6	7	8	9

CT =

0	0	1	1	1	1	0	1	0	0	0	0	1	0	1	0	1	0	1	0	0	0	0	0	0	1	0	0
1				5					10					15					20					25			28

B =

4	6	9	14	16	20	25	27	
1	2	3	4	5	6	7	8	9

Ti =

2	3	5	2	3	6	4	3	1
1	2	3	4	5	6	7	8	9

Sequence of ids of valid words

Canonical words & variants

1	2	3	4	5	6
$	blue	color	different	sea	sky

blue	0
Blue	1

Color	1
COLOR	01
colors	00

different	0

sea	0

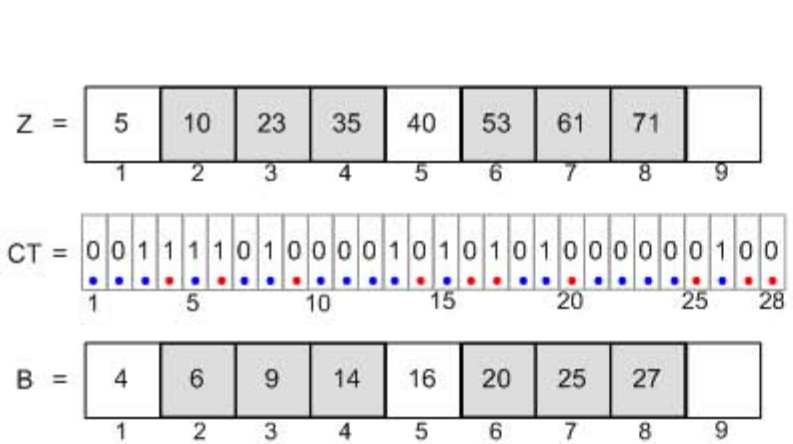
sky	0

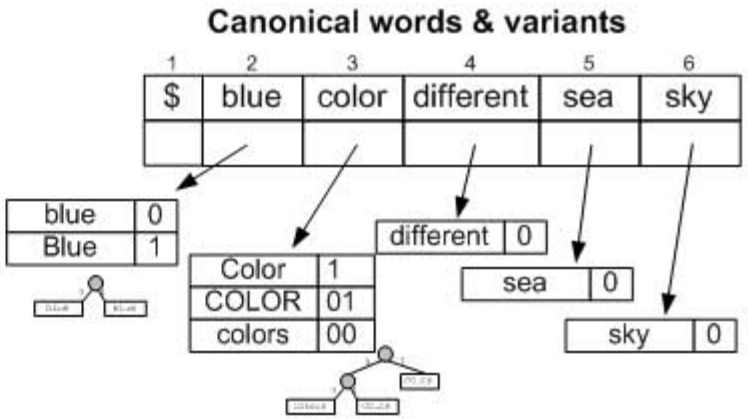
Separators

_	1
_of_the_	01
The_	001
are	0000
_and_the_	0001

Presentation Layer + WiCSA

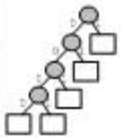
Presentation Layer: Sampling to reduce space





Separators

_	1
_of_the_	01
The_	001
are	0000
_and_the_	0001



- Sampled values → Ok !
- Non-sampled values → sincronize CT in previous sample + decoding

WiCSA: CSA on Integers

Source text

```
The Blue Color of the sea and the blue COLOR of the sky are different colors
1   5    10    16 19  23  27  31  35   40    46 49  53  57  61        71
```

Sequence of ids of valid words

Ti = | 2 | 3 | 5 | 2 | 3 | 6 | 4 | 3 | 1 |

Canonical words mapping

$	blue	color	different	sea	sky
1	2	3	4	5	6

	1	2	3	4	5	6	7	8	9
A =	9	1	4	8	2	5	7	3	6
A^{-1} =	2	5	8	3	6	9	7	4	1
Ψ =	2	5	6	1	8	9	4	3	7
D =	1	1	0	1	0	0	1	1	1

S = | ~~1~~ | ~~2~~ | ~~3~~ | ~~4~~ | ~~5~~ | ~~6~~ | **Implicit**

Presentation Layer + WiCSA

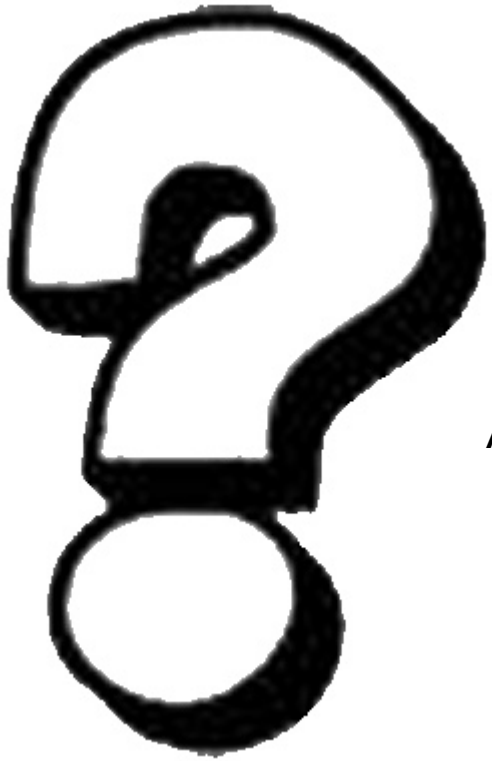
Some preliminary results on space (disk)

- *Ap Newswire 1998*, from *TREC-2 (250Mb English text)*

	t = sample period	
	64	32
WiCSA	39,543,689	44,916,341
Pres. Layer	43,838,071	47,008,551
Total (bytes)	**83,381,760**	**91,924,892**
WiCSA %	15.78%	17.92%
Pres. Layer %	17.49%	18.76%
C.R (%)	**33.27%**	**36.68%**

AP: 250,634,187bytes

A Compressed Self-Index on Words



Antonio Fariña
fari@udc.es