String Matching

A Lecture in CE Freshman Seminar Series:
Ten Puzzling Problems in Computer Engineering

About This Presentation

Edition	Released	Revised	Revised
First	May 2007		

Word Search Puzzles

Type 1, With Word List Supplied

S	E	A	L	P	V	C	A	T	P	A
W	A	S	A	B	A	C	G	U	G	L
I	D	S	P	W	L	F	L	I	V	I
T	A	E	A	C	V	L	T	M	O	N
C	S	M	W	T	E	A	L	E	A	T
H	X	B	W	Y	T	J	P	U	T	S
P	T	L	Q	O	H	C	T	U	L	C
X	R	Y	R	R	E	D	I	L	G	R
Z	S	L	O	R	T	N	O	C	T	E
G	N	I	L	P	U	O	C	Y	K	E
C	O	N	N	E	C	T	O	R	S	N

AGITATOR
ASSEMBLY
CLUTCH
CONNECTORS
CONTROL
COUPLING
GLIDE
LINT SCREEN
PULLEY
SEAL
SWITCH
VALVE

The puzzle below is a little harder than the normal word search: one of the 36 first/last names has been left out (which one?)

Y	O	U	R	M	O	J	D	Z	E	R	N	A	S	O
S	U	S	A	N	A	R	A	B	R	A	B	H	T	R
R	T	R	V	B	E	H	N	N	Y	W	D	L	O	G
U	K	C	A	L	R	I	A	L	B	K	R	N	L	J
F	S	L	H	I	G	Y	L	M	E	P	A	U	L	A
M	P	A	S	A	Y	O	M	V	M	A	H	G	I	V
A	M	C	V	N	R	U	I	A	Y	P	C	U	L	S
H	K	A	O	A	B	N	E	T	S	R	I	K	A	Z
G	L	T	C	F	G	G	O	S	R	T	R	V	P	E
N	U	E	A	U	V	E	O	E	S	A	E	A	A	R
I	F	L	E	R	B	R	Z	B	R	U	C	E	L	N
B	P	L	B	E	K	I	M	M	E	L	L	I	L	E
N	O	S	L	E	N	O	J	E	R	B	A	K	E	R

AMY STEEL
KEVIN BLAIR
RON PALILLO
BARBARA BINGHAM
KIRSTEN BAKER
SHAVAR ROSS
BRUCE MAHLER
LARRY ZERNER
STU CHARNO
CAROL LACATELL
MARK NELSON
SUSAN BLU
DANA KIMMELL
PAUL KRATKA
TONY GOLDWYN
JOHN FUREY
RICHARD YOUNG
TRACIE SAVAGE

“Ten Puzzling Problems in Computer Engineering” Word Search

WORD LIST:

BINARY SEARCH
BYZANTINE GENERALS
CRYPTOGRAPHY
EASY HARD IMPOSSIBLE
MALFUNCTION DIAGNOSIS
PLACEMENT AND ROUTING
SATISFIABILITY
SORTING NETWORKS
STRING MATCHING
TASK SCHEDULING

V	K	T	A	S	K	S	C	H	E	D	U	L	I	N	G	J	Y	Y	M
L	Y	O	W	F	H	L	X	E	Z	P	X	F	L	O	N	T	D	A	G
L	E	A	S	Y	H	A	R	D	I	M	P	O	S	S	I	B	L	E	Y
O	O	F	B	H	Q	R	J	W	H	S	B	F	N	L	T	F	T	H	B
X	M	G	Y	L	D	E	R	S	K	R	A	C	I	N	U	D	P	T	B
E	O	N	N	P	G	N	P	S	J	F	J	B	Z	N	O	A	E	B	Y
C	O	I	A	A	B	E	H	E	U	K	A	Q	C	B	R	T	M	H	H
Q	R	H	Z	K	A	G	T	D	A	I	H	T	Y	G	D	Y	C	R	D
B	S	C	F	K	U	E	I	G	F	Q	I	M	O	K	N	R	U	Y	T
Y	I	T	H	Z	S	N	Z	S	T	O	A	T	X	V	A	C	J	S	W
X	P	A	L	S	M	I	I	S	N	T	P	Z	J	E	T	E	E	Q	Z
W	W	M	S	L	K	T	Q	D	J	Y	U	W	S	G	N	G	K	W	J
B	O	G	W	F	A	N	I	F	R	J	W	Y	F	Q	E	C	X	S	Q
Z	A	N	C	S	J	A	J	C	D	J	R	B	Z	U	M	I	T	N	Q
O	Y	I	N	X	G	Z	Y	F	L	A	E	M	B	X	E	G	C	Y	J
R	W	R	Y	N	W	Y	A	N	N	C	E	U	D	N	C	D	K	C	L
D	K	T	O	H	P	B	B	I	B	M	V	F	B	C	A	D	W	P	G
Q	Q	S	O	U	C	L	B	C	V	L	L	J	L	R	L	N	Q	F	W
C	I	N	Z	P	J	V	E	A	E	L	H	X	Z	P	P	B	D	E	Y
S	O	R	T	I	N	G	N	E	T	W	O	R	K	S	X	J	I	Y	Q

Puzzle generated at:
http://puzzlemaker.school.discovery.com/WordSearchSetupForm.html

Word Search Puzzle

Type 2, With Clues Supplied for the Words to be Found

Seven birds

☐ ☐ ☐ ☐ ☐ ☐ ☐

Five units of length

☐ ☐ ☐ ☐ ☐

Four currencies

☐ ☐ ☐ ☐

Two things football players wear

☐ ☐

Large gland in the neck

☐

L	E	A	G	L	E	U	R	O	K	R	D
P	O	X	W	Y	A	R	D	R	X	E	O
I	E	O	T	H	Y	R	O	I	D	T	L
H	N	S	N	T	E	T	P	B	N	E	L
A	J	C	O	Z	S	L	M	O	I	M	A
W	Z	O	H	C	J	N	M	I	U	N	R
K	F	J	E	R	S	E	Y	E	L	N	B
V	E	G	R	E	T	X	Z	J	T	E	D

USA Today's "Word Roundup" for May 16, 2007: http://puzzles.usatoday.com/

```
00                  12                  24                  36
LEAGLEUROKRDPOXWYARDRXEOIEOTHYROIDTLHNSNTETPBNEL
AJCOZSLMOIMAWZOHCJNMIUNRKFJERSEYELNBVEGRETXZJTED
48                  60                  72                  84
```

String Matching: Problem Definition

Given a data string with *n* symbols and a pattern string with *m* symbols:

1. Does the pattern string appear in the data string?

2. What are the locations of all occurrences of the pattern in the data?

The brute-force, or sliding window, algorithm
Consider all possible positions where the pattern might begin ($n - m + 1$)
For each start position, do up to *m* comparison to see if there is a match

Worst-case complexity = $O(mn)$; e.g., pattern "aaaaa", data "aaaaaaaaaa"

Pattern string of length $m = 5$ symbols: **EAGLE**

Data string of length $n = 96$ symbols

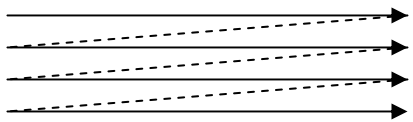
```
EAGLE          12                  24                  36
LEAGLEUROKRDPOXWYARDRXEOIEOTHYROIDTLHNSNTETPBNEL
AJCOZSLMOIMAWZOHCJNMIUNRKFJERSEYELNBVEGRETXZJTED
48             60                  72                  84
```

Converting 2D Search Puzzles to 1D Searches

A 2D word search puzzle looks more exotic but it can be readily converted to a 1D string search puzzle

L	E	A	G	L	E	U	R	O	E	X	T
R	A	X	W	Y	A	R	D	R	X	E	O
I	E	O	T	H	Y	R	O	I	D	T	L
H	N	S	N	T	E	T	P	B	N	E	L
A	J	C	O	Z	S	L	M	O	I	M	A
W	Z	O	H	C	J	N	M	I	U	N	R
K	F	J	E	R	S	E	Y	E	L	N	B
V	E	G	R	E	T	X	Z	J	T	E	D

Row-major order



LEAGLEUROEXTRAXWYARDRXEOIEOTHYROIDTLHNSNTETPBNEL
AJCOZSLMOIMAWZOHCJNMIUNRKFJERSEYELNBVEGRETXZJTED

Insert a special symbol (#) between rows to ensure that new words or patterns are not created by the expansion

LEAGLEUROEXT#RAXWYARDRXEO#IEOTHYROIDTL#HNSNTETPBNEL#
AJCOZSLMOIMA#WZOHCJNMIUNR#KFJERSEYELNB#VEGRETXZJTED#

Column-major order

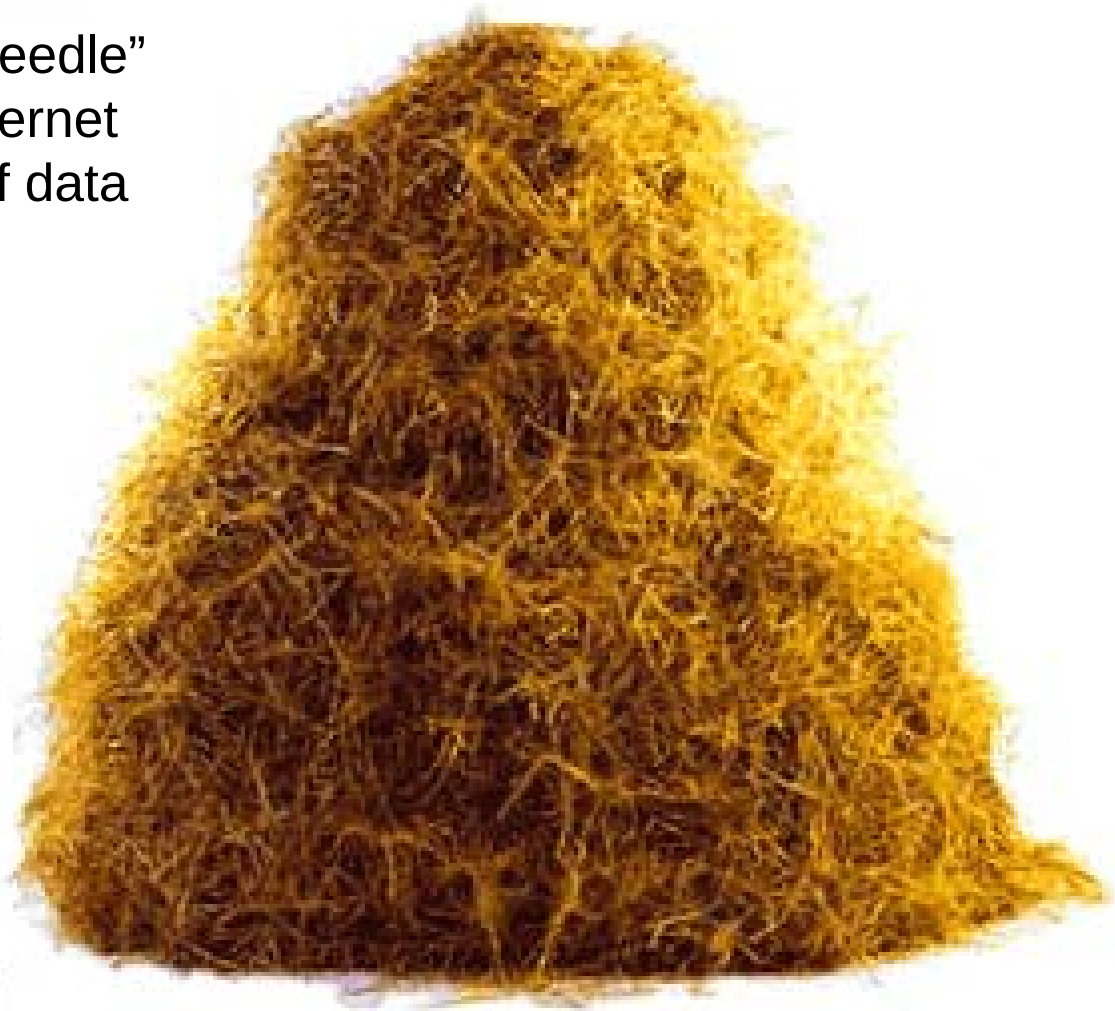
Similarly for (anti)diagonal

Finding a Needle in a Haystack

Search for the 10-symbol “needle” h-e-l-e-n- -h-u-n-t in the Internet “haystack” with many TBs of data

The brute force algorithm amounts to the following: “Look in this corner, now in this other corner, then over there, and so on.”

The Internet holds some 20B pages, each page on average containing in excess of 100 KB



$m \cong 10$

$n \cong 20\times10^9\times10^5 = 2\times10^{12}$ B

$O(mn) \cong 2\times10^{13}$ comparisons $\Rightarrow$ 20,000 s (>5 hr), with 10^9 comparisons/s

Needle in a Haystack: Internet Search

"helen hunt" - Google Search - Windows Internet Explorer

http://www.google.com/search?num=50&hl=en&newwindow=1&safe=off&q=%22helen+hunt%22&btnG=Search

File Edit View Favorites Tools Help

Favorites Cen... "helen hunt" - Google Search

Page Tools

3 Weeks Ago
2 Weeks Ago
Last Week
Today

Web Images Video News Maps Gmail more ▼ Sign in

Google "helen hunt" Search

Search for the 10-symbol string h-e-l-e-n- -h-u-n-t

Web Images Results 1 - 50 of about 1,330,000 for "helen hunt". (0.21 seconds)

Helen Hunt (I)
Helen Hunt (I) on IMDb: Movies, TV, Celebs, and more...
www.imdb.com/name/nm0000166/ - 49k - May 19, 2007 - Cached - Similar pages

IMDb Name Search
A search for "Helen Hunt" found the following results: ... Helen Hunt (II) (Make-Up Department, Guess Who's Coming to Dinner (1967)); Helen Hunt (III) ...
www.imdb.com/Name?Hunt,+Helen - 16k - Cached - Similar pages
[More results from www.imdb.com]

Helen Hunt - Wikipedia, the free encyclopedia
Helen Hunt. Birth name, Helen Elizabeth Hunt. Born, June 15, 1963 (1963-06-15) (age 43)

Sponsored Links
Helen Hunt
Find Pics, News, Movies, Interviews Filmography and More at Moviefone
Moviefone.com

"hellen hunt" - Google Search - Windows Internet Explorer

http://www.google.com/search?num=50&hl=en&newwindow=1&safe=off&q=%22hellen+hunt%22&btnG=Search

File Edit View Favorites Tools Help

Favorites Cen... "hellen hunt" - Google Search

Page Tools

3 Weeks Ago
2 Weeks Ago
Last Week
Today

Web Images Video News Maps Gmail more ▼ Sign in

Google "hellen hunt" Search Advanced Search Preferences

Web Results 1 - 50 of about 735 for "hellen hunt". (0.06 seconds)

Did you mean: "*helen* hunt"

Helen Hunt (I)

Sponsored Links
Helen Hunt

Needle in a Haystack: Doing Less Work

For a particular pattern and unpredictable data strings, preprocess the pattern so that searching for it in various data strings becomes faster

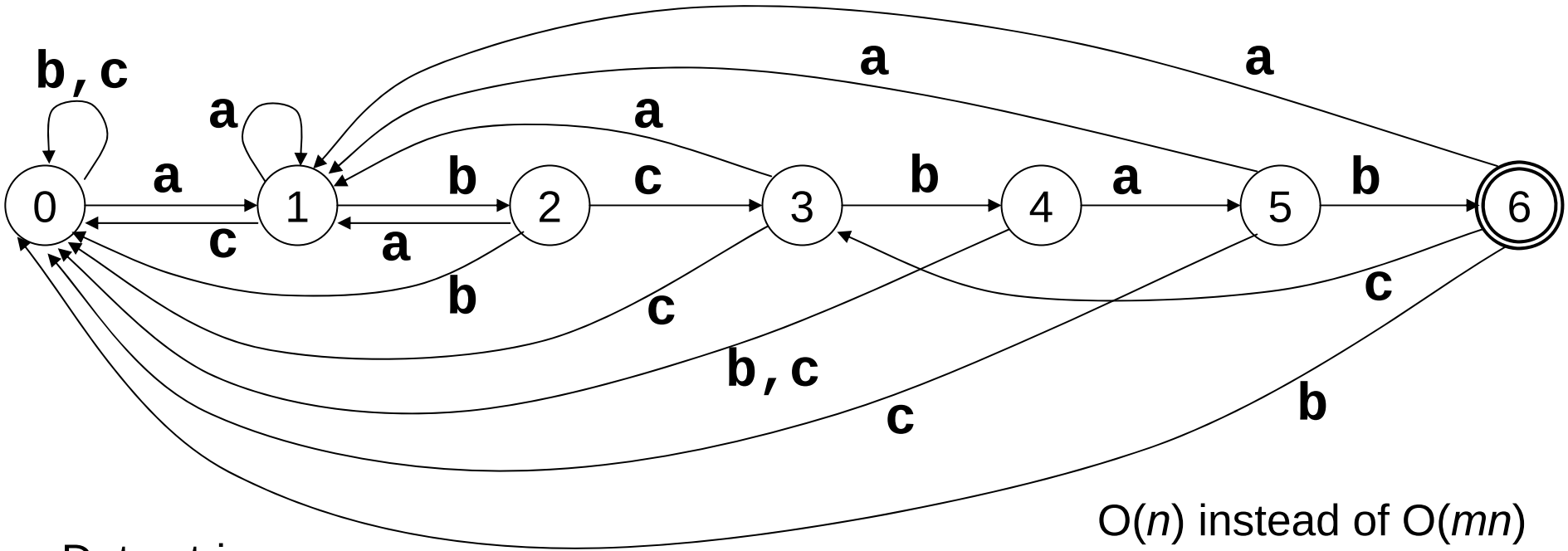
Analogy: Magnetize the needle

For a particular data string and unpredictable patterns, preprocess the data string so that when a pattern is supplied, we can readily find it with much less work

Analogy: Do a thorough search of the haystack for different types of needles and place markers to guide future searches

Example of Preprocessing the Pattern String

Devise an efficient method for finding the pattern "**abcbab**" in various data strings formed from the symbols **a**, **b** and **c**



O(*n*) instead of O(*mn*)

Data string:

a	**b**	**c**	**b**	**b**	**b**	**a**	**b**	**c**	**b**	**a**	**b**	**b**	**c**	**a**	**a**	**b**	**c**	**b**	**a**	**b**	**c**	**b**	**a**	**b**	**c**	**b**	**b**
0	1	2	3	4	0	0	1	2	3	4	5	**6**	0	0	1	1	2	3	4	5	**6**	3	4	5	**6**	3	4	0

Example of Preprocessing the Data String

Devise an efficient method for finding various patterns in the data string:

a	b	c	b	b	b	a	b	c	b	a	b	b	c	a	a	b	c	b	a	b	c	b	a	b	c	b	b
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27

a a b	14
a b b	10
a b c	0, 6, 15, 19, 23
b a b	5, 9, 18, 22
b b a	4
b b b	3
b b c	11
b c a	12
b c b	1, 7, 16, 20, 24
c a a	13
c b a	8, 17, 21
c b b	2, 25

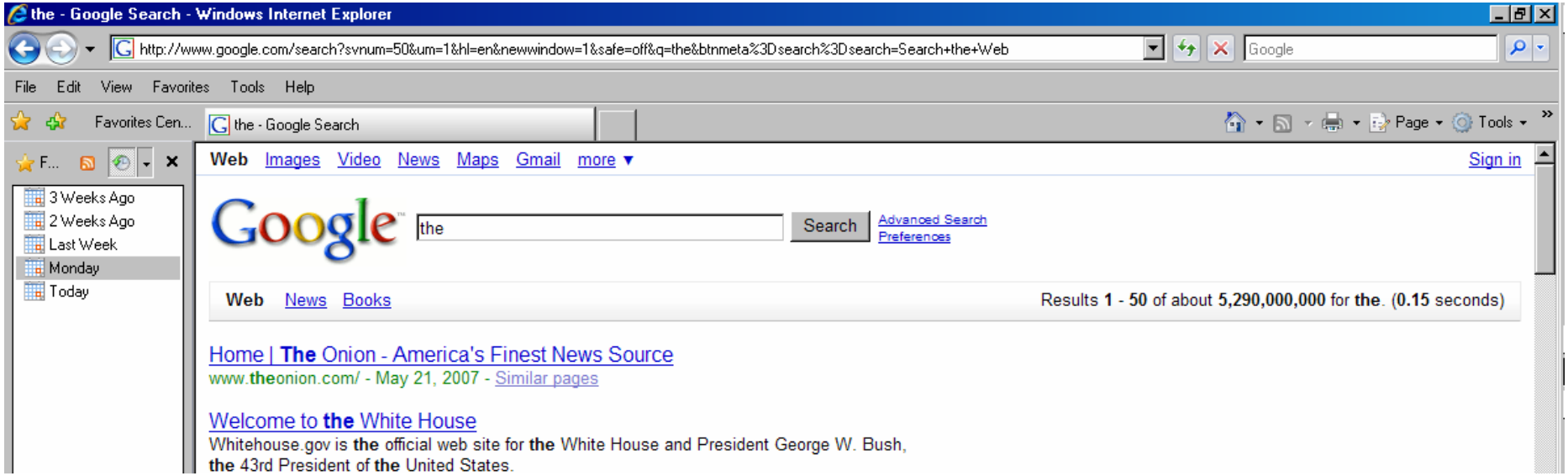
Find all occurrences of the pattern "**abcbab**"

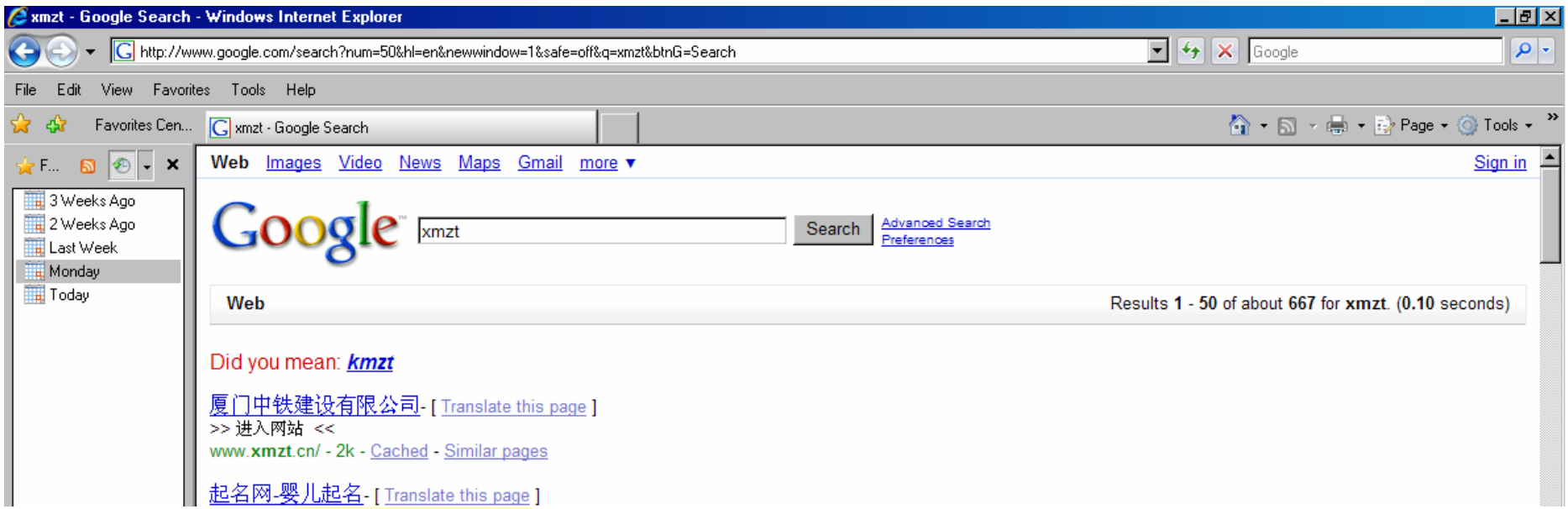
a b c	0, 6, 15, 19, 23
b c b	1, 7, 16, 20, 24
c b a	8, 17, 21
b a b	5, 9, 18, 22

a b c	0, 6, 15, 19, 23
b c b	1, 7, 16, 20, 24
c b a	8, 17, 21
b a b	5, 9, 18, 22

Alternate strategy: Focus on the locations of **a b c** and **b a b**

Search Engine Indexes





Approximate String Matching

Notion of string distance

Each of the following transformations in a string creates a distance of 1

1. Deletion of a symbol
2. Insertion of an extra symbol
3. Transposition of two adjacent symbols

Example distance-1 strings for **`helen hunt:`**

`hellen hunt`	Insertion
`elen hunt`	Deletion
`helen hnut`	Transposition

Example distance-2 strings for **`helen hunt:`**

`hellen hnut`	Insertion + Transposition
`elen huntt`	Deletion + Insertion
`lheen hunt`	2 Transpositions

Wildcard symbols can help in formulating approximate string searches

`h* hunt` means any string that begins with an "**h**", ends with "**`hunt`**", and has an arbitrary set of symbols between the two

Melvyl (UC library catalog) allows such searches, e.g., author: **`hunt h*`**