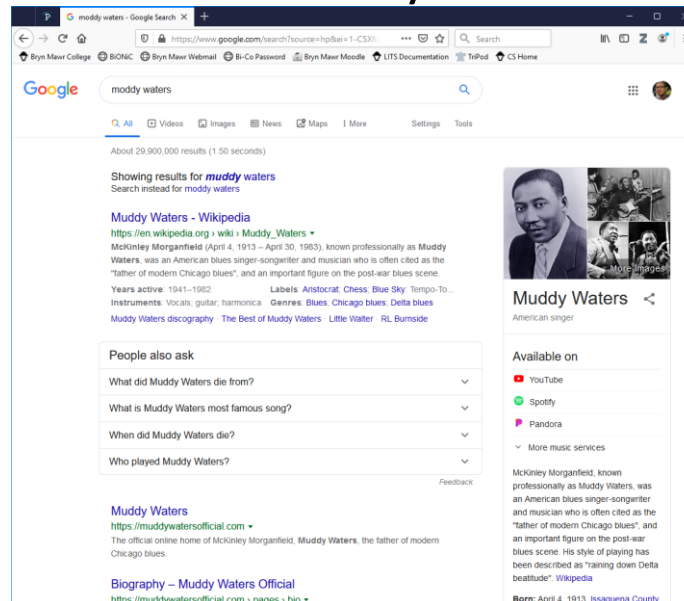


Query



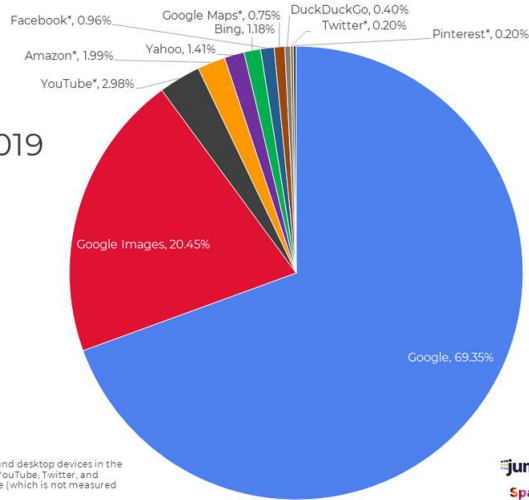
Search Engines...

Altavista	Entireweb	Leapfish	Spotify
Ask	Excite	Lycos	Stinky Teddy
Baidu	Faroo	Maktoob	Stumpedia
Bing	Info.com	Miner.hu	Swisscows
Blekkio	Fireball	Monster Crawler	Teoma
ChaCha	Gigablast	Naver	Walla
Dogpile	Google	Omgili	WebCrawler
Daum	Go	Rediff	Yahoo!
Dmoz	Goo	Scrub The Web	Yandex
Du	Hakia	Seznam	Yippy
Egerin	HotBot	Sogou	Youdao
ckDuckGo		Soso	

Search Engine Marketshare 2019

Search Engine Market Share Q2 2019

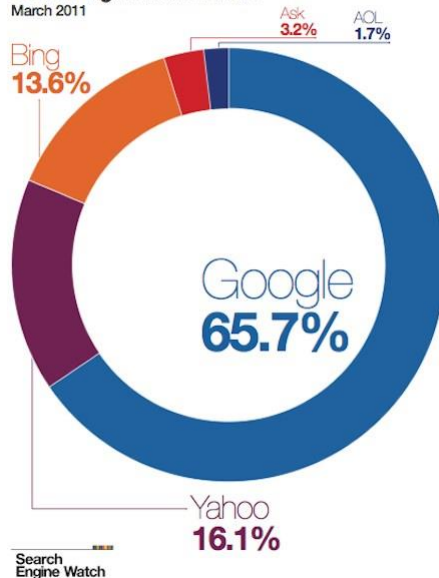
94%
of all searches happen on a Google property



* Data from 230B+ browser-based searches on millions of mobile and desktop devices in the United States. Search share on Google Maps, Facebook, Amazon, YouTube, Twitter, and Pinterest are likely underrepresented due to heavy mobile app use (which is not measured by Jumpshot's browser-based panel)

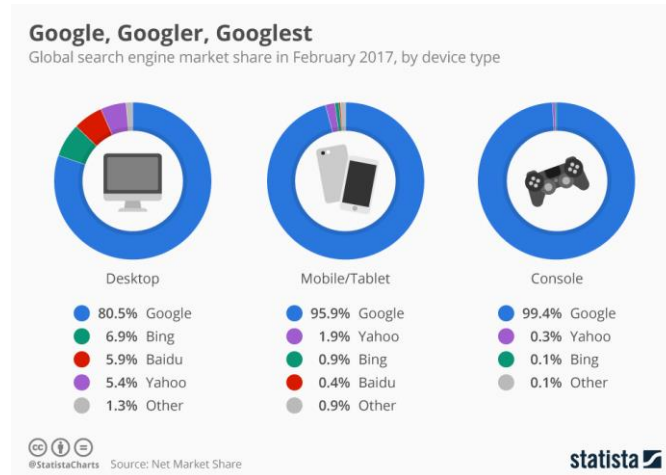


Search Engine Market Share March 2011

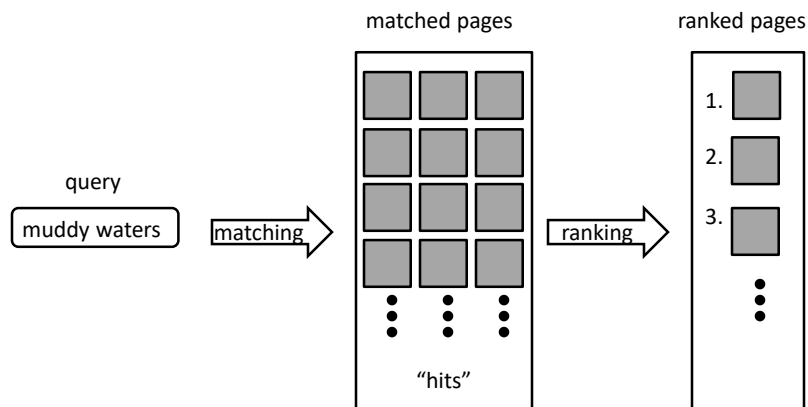


Search Engine Watch

Search Engine Marketshare 2017



Matching & Ranking



Index

Gregorian Calendar 242
Grey Poupon 38, 94

H

Hallway Cruiser 120
Hektor robot 261
Hertz (Hz) 169
hi-fidelity 170
HiLo game 154
Hoare, C. A. R. 227
Hogg, David 127
Hugs & Kisses 210
Human-robot interaction 262-63

I

iCat robot 263
IDLE 8, 22, 23, 29, 38
Idle, Eric 23
if-statement 100, **103**, 118, **128**. 270-71
image 182
Image 168, **176**, 280
image processing 190
image understanding 195
Imitation Game 206
import 138-39, 275
in 92, **103**, 270
Indecisive 117
init **13**, 275
initialization 9, 10, **13**, 275

Jones, Crispin 259
Jones, Mick 107
JPEG 183
Julia Sets 178

K

Kasparov, Gary 209
Kismet robot 263
Kitaoka, Akiyoshi 181
Koch Snowflakes 178
Konane 209

L

Ladybug 107
Larson, Doug 227
LavenderBlush 159
Law, Jude 205
Leap Frog 260
leap year 241-
LED 73
LEGO Mindstorms 4
len 92, **104**, 270
Lenhi, Jurg 261
Light following 121-22
Line 160, **176**, 280
linear time algorithms 252
List comprehensions 214, **224**
lists 49, 91-93, 270
Loan calculator 139-47

Inverted Index

- A mapping from content (words) to location.
- Example:

1	the cat sat on the mat		2	the dog stood on the mat		3	the cat stood while a dog sat
---	---------------------------	--	---	-----------------------------	--	---	----------------------------------

Inverted Index

1 the cat sat on
the mat

2 the dog stood on
the mat

3 the cat stood
while a dog sat

a	3
cat	1 3
dog	2 3
mat	1 2
on	1 2
sat	1 3
stood	2 3
the	1 2 3
while	3

Inverted Index

1 the cat sat on
the mat

2 the dog stood on
the mat

3 the cat stood
while a dog sat

a	3
cat	1 3
dog	2 3
mat	1 2
on	1 2
sat	1 3
stood	2 3
the	1 2 3
while	3

Every word in every
web page is indexed!

Searching

1 the cat sat on
the mat

2 the dog stood on
the mat

3 the cat stood
while a dog sat

query

cat

a	3
cat	1 3
dog	2 3
mat	1 2
on	1 2
sat	1 3
stood	2 3
the	1 2 3
while	3

Searching

1 the cat sat on
the mat

2 the dog stood on
the mat

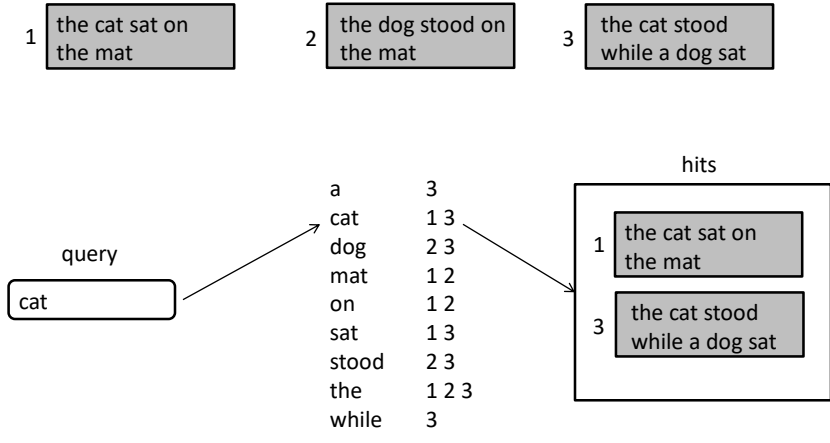
3 the cat stood
while a dog sat

query

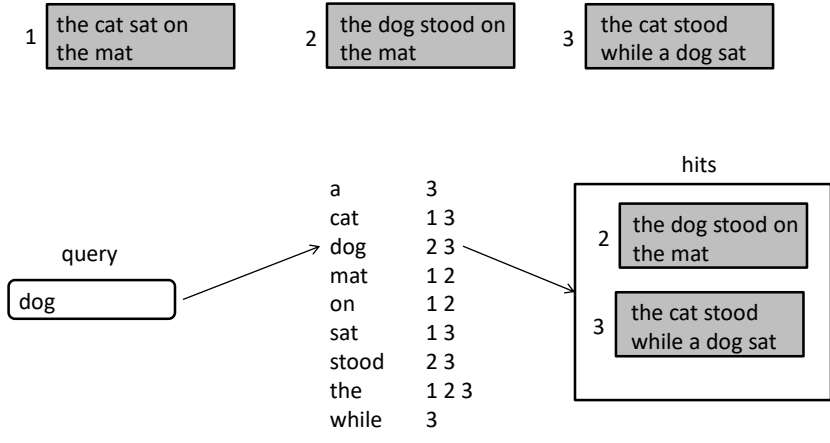
cat

a	3
cat	1 3
dog	2 3
mat	1 2
on	1 2
sat	1 3
stood	2 3
the	1 2 3
while	3

Searching



Searching



Searching

1 the cat sat on
the mat

2 the dog stood on
the mat

3 the cat stood
while a dog sat

query
cat dog

a	3
cat	1 3
dog	2 3
mat	1 2
on	1 2
sat	1 3
stood	2 3
the	1 2 3
while	3

Searching

1 the cat sat on
the mat

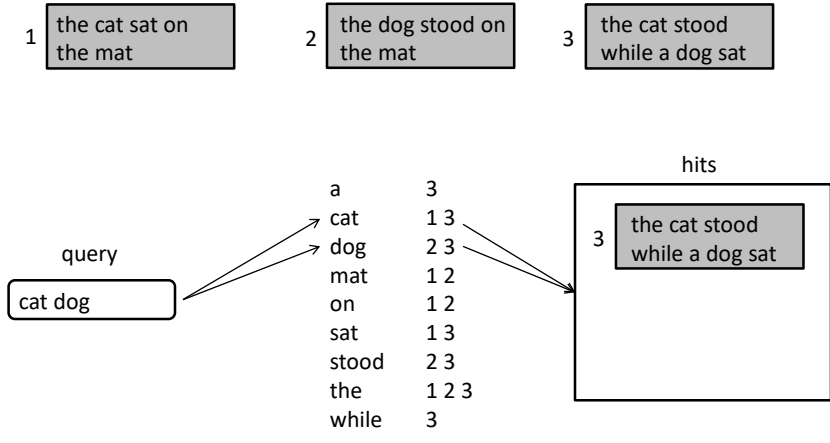
2 the dog stood on
the mat

3 the cat stood
while a dog sat

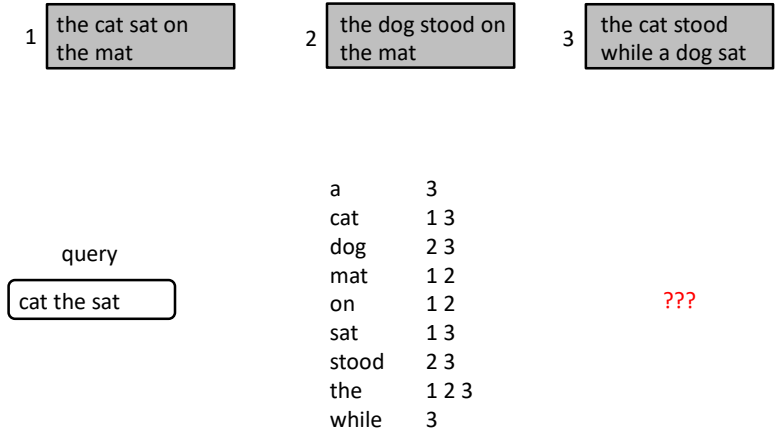
query
cat dog

a	3
cat	1 3
dog	2 3
mat	1 2
on	1 2
sat	1 3
stood	2 3
the	1 2 3
while	3

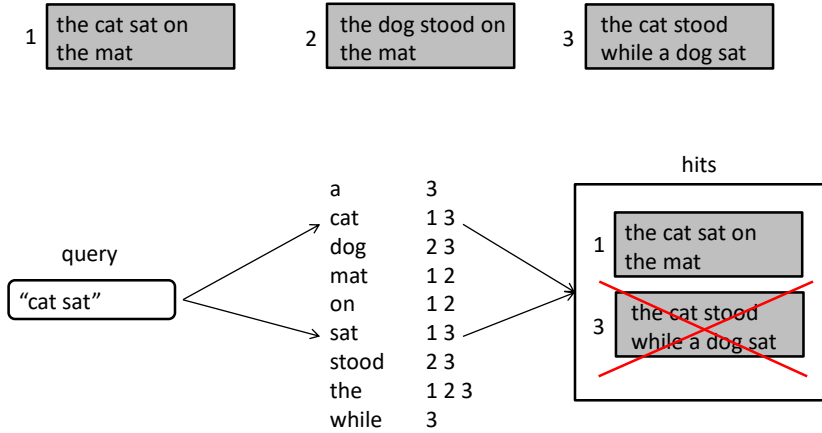
Searching



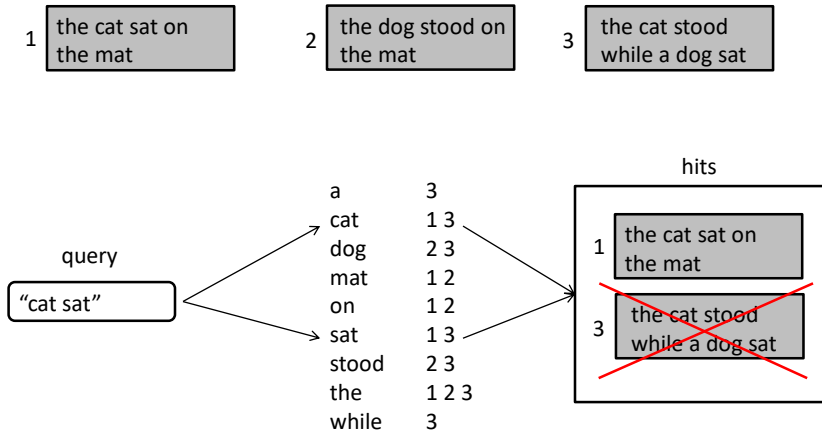
Searching



Phrase Queries

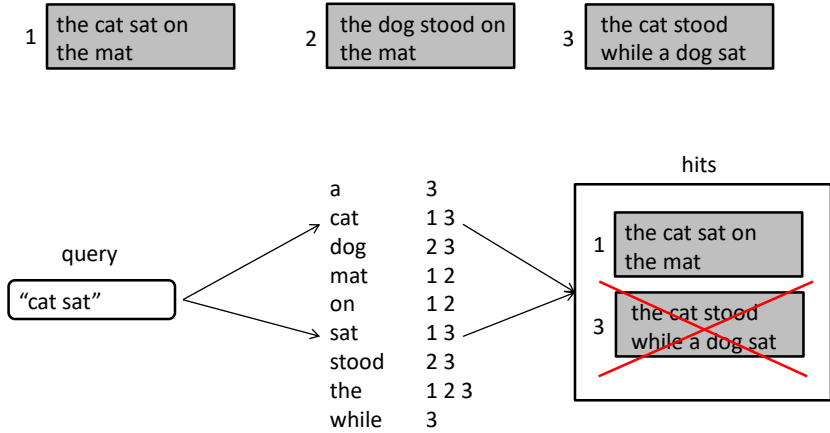


Phrase Queries



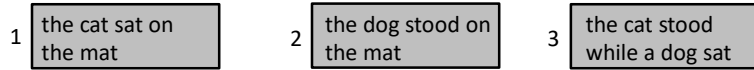
How to tell if two words occur next to each other?

Phrase Queries



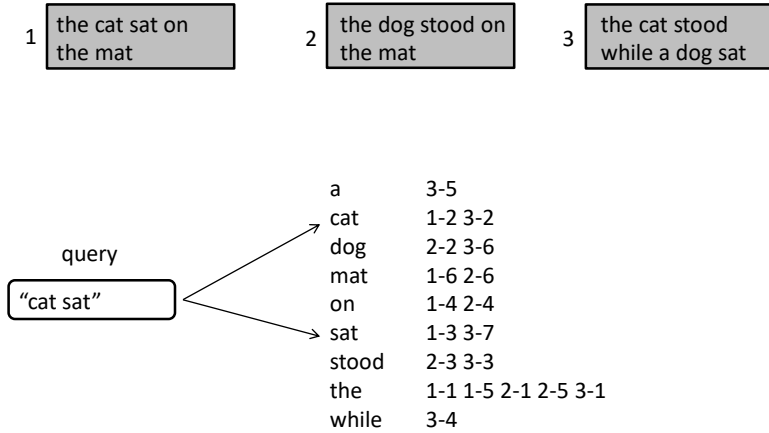
How to tell if two words occur next to each other? **EFFICIENTLY???**

Inverted Index **with Location**

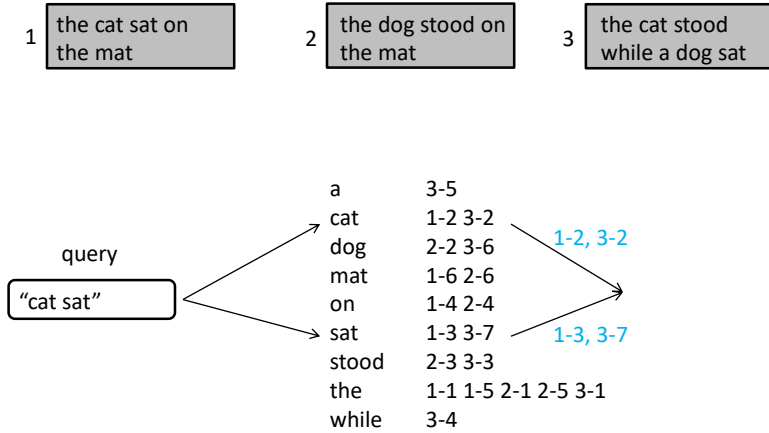


a	3-5
cat	1-2 3-2
dog	2-2 3-6
mat	1-6 2-6
on	1-4 2-4
sat	1-3 3-7
stood	2-3 3-3
the	1-1 1-5 2-1 2-5 3-1
while	3-4

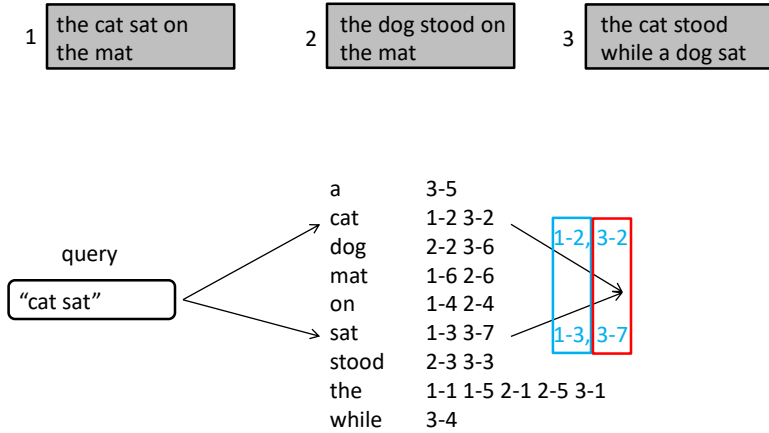
Inverted Index with Location



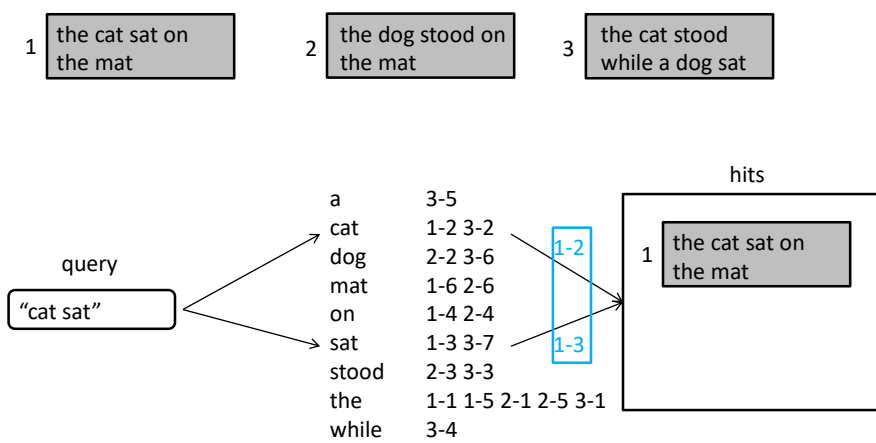
Inverted Index with Location



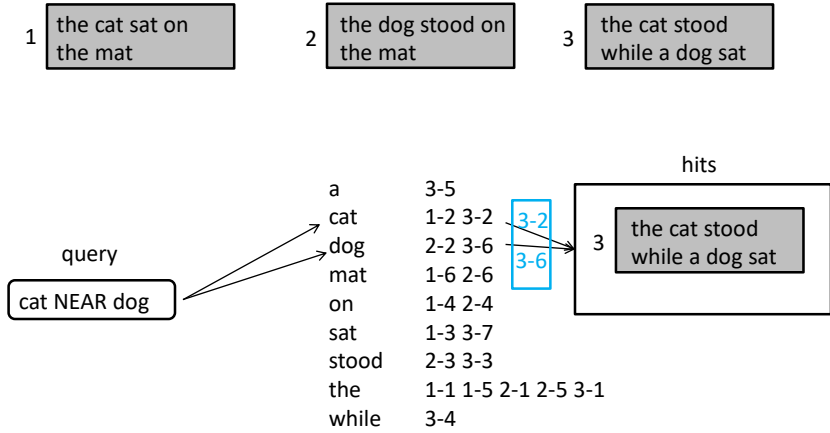
Inverted Index with Location



Inverted Index with Location

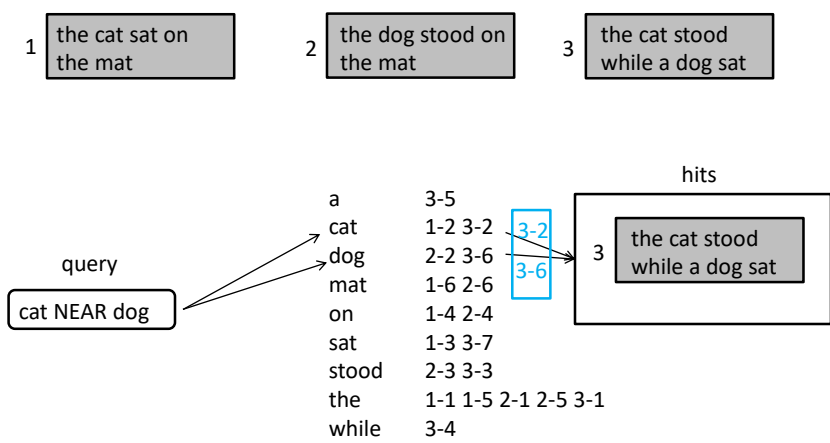


NEAR* Queries



*NEAR: distance <= 5

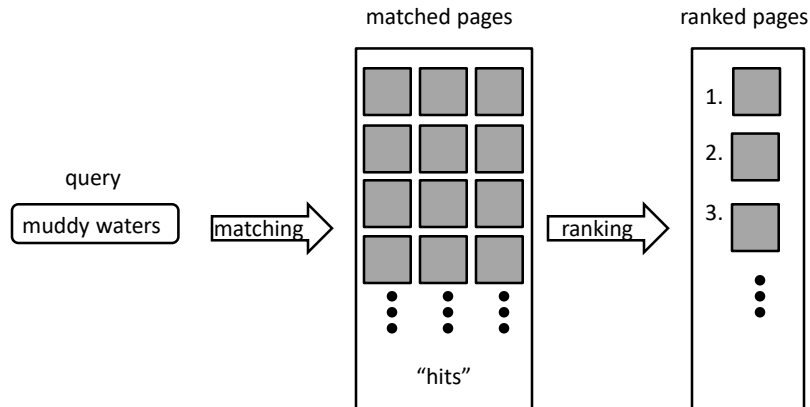
NEAR* Queries



Useful in ranking!

*NEAR: distance <= 5

Matching & Ranking



Ranking & Relevance

1 By far the most common cause of malaria is being bitten by an infected mosquito, but there are also other ways to contract the disease.

2 Our cause was not helped by the poor health of the troops, many of whom were suffering from malaria and other tropical diseases.

Ranking & Relevance

1 By far the most common cause of malaria is being bitten by an infected mosquito, but there are also other ways to contract the disease.

2 Our cause was not helped by the poor health of the troops, many of whom were suffering from malaria and other tropical diseases.

also 1-19
 ...
 cause 1-6 2-2
 ...
 malaria 1-8 2-19
 ...
 whom 2-15

Ranking & Relevance

1 By far the most common cause of malaria is being bitten by an infected mosquito, but there are also other ways to contract the disease.

2 Our cause was not helped by the poor health of the troops, many of whom were suffering from malaria and other tropical diseases.

query
 malaria cause

also 1-19
 ...
 cause 1-6 2-2
 ...
 malaria 1-8 2-19
 ...
 whom 2-15

Ranking & Relevance

1 By far the most common cause of malaria is being bitten by an infected mosquito, but there are also other ways to contract the disease.

2 Our cause was not helped by the poor health of the troops, many of whom were suffering from malaria and other tropical diseases.

query
malaria cause

also 1-19
...
cause 1-6 2-2
...
malaria 1-8 2-19
...
whom 2-15

Nearness can resolve the ranking!

Using Metadata

The screenshot shows a web browser window with the following content:

- Browser Tab:** CS380: Science of Information
- Address Bar:** https://cs.brynmawr.edu/Courses/c...
- Page Header:** Bryn Mawr College
- Course Title:** CS 380: Recent Advances in Computer Science
- Topic:** Science of Information
- Term:** Fall 2019
- Class Number:** 2283
- Section:** Course Materials
- Navigation Menu:** Information, Texts, Important Dates, Assignments, Lectures, Grading, Links
- General Information:**
 - Instructor:** Deepak Kumar, 202 Park Hall, 526-7485
 - E-Mail:** dkumar at cs.brynmawr dot edu
 - Twitter:** @bmcdeepak
 - WWW:** http://cs.brynmawr.edu/~dkumar
 - Lecture Hours:** Mondays & Wednesdays 11:40a to 1:00p
 - Office Hours:** Wednesdays 2:00 to 3:30p
 - Room:** Park Science Building, Room 159
 - Laboratories:**
 - Computer Science Lab Room 231 (Science Building), self scheduled.
- Course Description:**

Claude Shannon's foundations of *information theory* have paved the way for data storage, compression, encoding, and transmission for the Internet, CDs, DVDs, MP3 players, JPEGs, WiFi, iPods, mobile phones, and a whole host of applications underlying today's information technologies. The past six decades have brought information theory to the crossroads of several traditional disciplines: mathematics, statistics, computer science, physics, neurobiology, and electrical engineering. This course introduces students to the fundamentals of Information Theory and leads them to a broader understanding of the concept of "information" that transcends boundaries between disciplines, especially between physical and life sciences, communication, and knowledge extraction from massive datasets. Students in several disciplines will be able to draw upon the latest discoveries from multiple disciplines, replicate and discuss recent research, and learn to apply the techniques and tools of information-based inquiry in their lives.

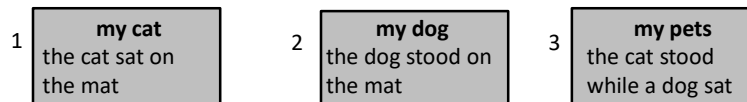
Using Metadata

```

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"
"http://www.w3.org/TR/html4/loose.dtd">
<html>
  <head> <meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
    <title>CS380: Science of Information (Course Page)</title> ...
  </head>
  <body>
    <P>
      <CENTER>
        <h3>Bryn Mawr College<BR>
        <B><FONT SIZE="+2">CS 380: Recent Advances in Computer Science
        <br> Topic: Science of Information Fall 2019</FONT>
        </B>
        <br> BMC Class Number: 2283
        <BR>
        <B><FONT SIZE="+2">Course Materials</FONT></B>
      ...

```

Metadata



Metadata

1	<p>my cat the cat sat on the mat</p>	2	<p>my dog the dog stood on the mat</p>	3	<p>my pets the cat stood while a dog sat</p>
1	<p><title>my cat </title> <body> the cat sat on the mat </body></p>	2	<p><title>my dog </title><body> the dog stood on the mat</body></p>	3	<p><title>my pets </title><body>th e cat stood while a dog sat</p>

Metadata

1	<p><title>my cat </title> <body> the cat sat on the mat </body></p>	a	3-10
		cat	1-3 1-7 3-7
		dog	2-3 2-7 3-11
		mat	1-11 2-11
		my	1-2 2-2 3-2
		on	1-9 2-9
		pets	3-3
		sat	1-8 3-12
		stood	2-8 3-8
		the	1-6 1-10 2-6 2-10 3-6
		while	3-9
		<body>	1-5 2-5 3-5
		</body>	1-12 2-12 3-13
		<title>	1-1 2-1 3-1
		</title>	1-4 2-4 3-4
2	<p><title>my dog </title><body> the dog stood on the mat</body></p>		
3	<p><title>my pets </title><body>th e cat stood while a dog sat</p>		

Structure Queries

query

intitle: dog

```

a          3-10
cat       1-3 1-7 3-7
dog       2-3 2-7 3-11
mat       1-11 2-11
my        1-2 2-2 3-2
on        1-9 2-9
pets      3-3
sat       1-8 3-12
stood     2-8 3-8
the       1-6 1-10 2-6 2-10 3-6
while     3-9
<body>    1-5 2-5 3-5
</body>   1-12 2-12 3-13
<title>   1-1 2-1 3-1
</title>  1-4 2-4 3-4

```

Structure Queries

query

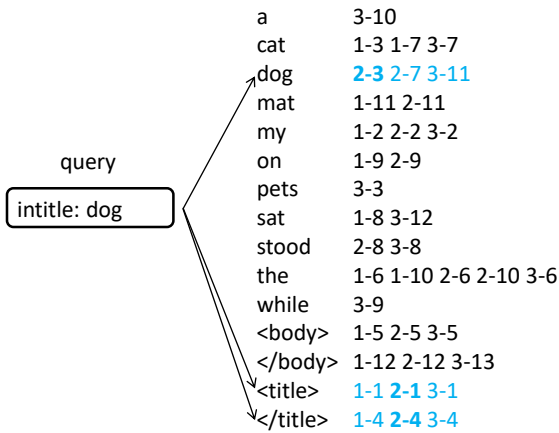
intitle: dog

```

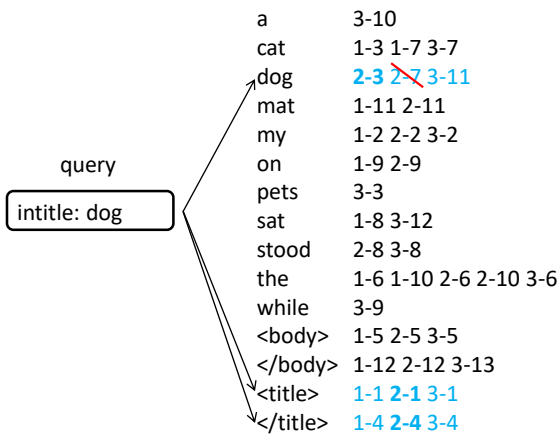
a          3-10
cat       1-3 1-7 3-7
dog       2-3 2-7 3-11
mat       1-11 2-11
my        1-2 2-2 3-2
on        1-9 2-9
pets      3-3
sat       1-8 3-12
stood     2-8 3-8
the       1-6 1-10 2-6 2-10 3-6
while     3-9
<body>    1-5 2-5 3-5
</body>   1-12 2-12 3-13
<title>   1-1 2-1 3-1
</title>  1-4 2-4 3-4

```

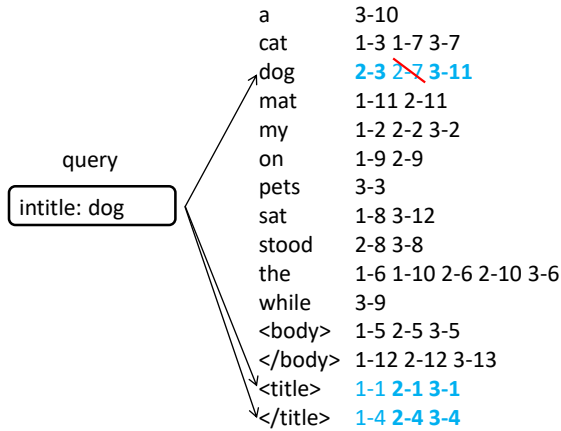
Structure Queries



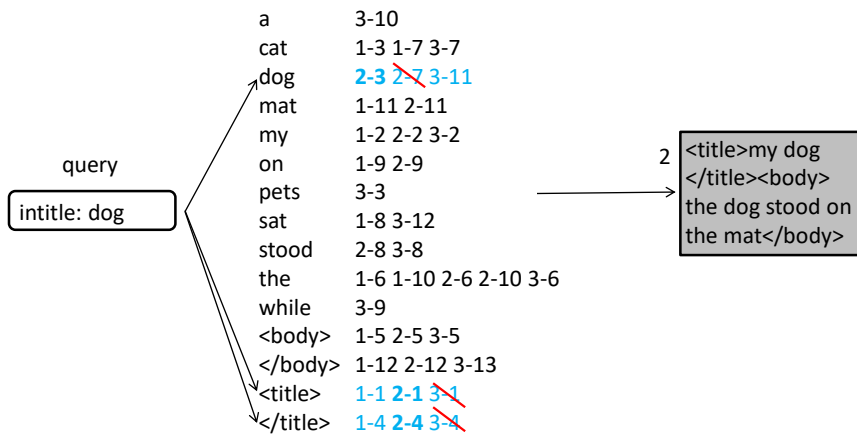
Structure Queries



Structure Queries



Structure Queries



Web Information Retrieval

- Search Engines
- Queries
 - phrase queries
 - structure queries (NEAR, intitle:, ...)
- Matching
- Inverted Index
 - page number
 - location
- Ranking & Relevance
- Metadata

Web Information Retrieval

- Search Engines
- Queries
 - phrase queries
 - structure queries
- Matching
- Inverted Index
 - page number
 - location
- Ranking & Relevance
- Metadata

**Efficient matching
is only one half the story.**

**The other grand challenge
is how to rank the
matching pages**

References

- *Google's PageRank and Beyond*, Amy N. Langville and Carl D. Meyer, Princeton University Press, 2006.
- *Nine Algorithms That Changed The Future*, John MacCormick, Princeton University Press, 2012.
- *Learning Computing with Robots*, Deepak Kumar, IPRE 2011.