Basics of Information Retrieval

Term-Document Matrix, Inverted Index and Boolean Retrieval

Debapriyo Majumdar Indian Statistical Institute debapriyo@isical.ac.in

Information Retrieval



User needs some information.



An information retrieval system tries to bridge this gap.



Assumption: the required information is present somewhere.

The goal of an information retrieval system is to satisfy user's information need.

Query

Google

Google Search I'm Feeling Lucky

Society > Honor

Google results 1-10 of about 23,000 for april fool's. Search took 0.07 seconds

10 results

April Fool's Day

Quick response

Basic example

User expresses the information need in the form of a query.

The system returns a (ranked) list of results.

AprilFools.com

...BUILD-A-DAD Foolishness · April Fool's History · ...

Description: AprilFools.com is Gag Central for April Fools 1999, the cool Category: Society > Holidays > April Fool's Day aprilfools.infospace.com/ - Cached - 10k - GoogleScout

April Fool's Day 1999 - Practical Jokes, Pranks, Tips and

...Bangkok.com April Fool's Day Page April...
... Fool's Day is here! Do you have a favorite April...
bangkok.com/fools/ - Cached - 42k - GoogleScout

Ranked results

Google in 2000

Courtesy: http://googlesystem.blogspot.com/

2007/12/google-in-2000.html

Many matches

Meh Images Video Audio Di

arch Results

- Also try: 2004 election results, election results 2006, 2000 election results More...
- Looking for Election Result Products? www.ebay.com - Find exactly what you want today.

TAHOO! SEARCH election results

- The 2006 Elections
 www.nytimes.com Continuous coverage of the 2006 elections and results from The New York Times.
- Y News Results for election results

Election results from the South - AP via Yahoo! News - 45 minutes ago
Election results from the Midwest - AP via Yahoo! News - 1 hour, 37 minutes ago
Election results from the Northeast - AP via Yahoo! News - 2 hours, 0 minute ago
Yahoo! Shortcut - About

1. Elections 2006 - CNN

Ongoing analysis and coverage of the 2006 midterm elections. Includes an issue spotlight and the latest headlines.

Category: 2006 Midterm Elections www.cnn.com/ELECTION/2006 - 57k - Cached - More from this site

2. You Decide 2006 - Fox News

Features coverage and news analysis of the 2006 midterm elections.

Category: 2006 Midterm Elections
www.foxnews.com/politics/youdecide2006 - 68k - Cached - More from this site

Midterm Elections 2006 Full Coverage on Yahoo! News
 Ongoing coverage of the key races for the 2006 midterm elections.
 Category: 2006 Midterm Elections
 news.yahoo.com/fc/US/Midterm_Elections_2006 - 50k - Cached - More from this site

Yahoo search in 2006

Courtesy: https://www.searchenginewatch.com/

2006/11/08/in-the-election-results-race-yahoos-the-winner/

Ranked results

Collection and Documents

The curse of the black pearl

Ship Captain Jack
Sparrow Caribbean
Elizabeth Gun Fight

The Dark Knight

Bruce Wayne Batman
Joker Harvey Gordon
Gun Fight Crime

Finding Nemo

Ocean Fish Nemo Reef Animation

Skyfall

007 James Bond MI6 Gun Fight

Tintin

Ocean Animation
Ship Captain
Haddock Tintin

Silence of the Lambs

Hannibal Lector FBI Crime Gun Cannibal

<u>Titanic</u>

Ship Rose Jack
Atlantic Ocean
England Sink
Captain

The Ghost Ship

Ship Ghost Ocean Death Horror

- Document: unit of retrieval
- Collection: the group of documents from which we retrieve
 - Also called the corpus (a body of text)

Boolean retrieval

The curse of the black pearl

Ship Captain Jack
Sparrow Caribbean
Elizabeth Gun Fight

The Dark Knight

Bruce Wayne Batman
Joker Harvey Gordon
Gun Fight Crime

Finding Nemo

Ocean Fish Nemo Reef Animation

Skyfall

007 James Bond MI6 Gun Fight

Tintin

Ocean Animation
Ship Captain
Haddock Tintin

Silence of the Lambs

Hannibal Lector FBI Crime Gun Cannibal

Titanic

Ship Rose Jack
Atlantic Ocean
England Sink
Captain

The Ghost Ship

Ship Ghost Ocean Death Horror

- Find all documents containing a word w
- Find all documents containing a word w_1 but not containing the word w_2
- Queries in the form of any Boolean expression
- Query: Jack

Boolean retrieval

The curse of the black pearl

Ship Captain Jack
Sparrow Caribbean
Elizabeth Gun Fight

The Dark Knight

Bruce Wayne Batman
Joker Harvey Gordon
Gun Fight Crime

<u>Finding Nemo</u>

Ocean Fish Nemo Reef Animation

Skyfall

007 James Bond MI6 Gun Fight

Tintin

Ocean Animation
Ship Captain
Haddock Tintin

Silence of the Lambs

Hannibal Lector FBI Crime Gun Cannibal

Titanic

Ship Rose Jack
Atlantic Ocean
England Sink
Captain

The Ghost Ship

Ship Ghost Ocean Death Horror

- Find all documents containing a word w
- Find all documents containing a word w_1 but not containing the word w_2
- Queries in the form of any Boolean expression
- Query: Jack

Term – document matrix

	Black	Finding	Tintin	Titanic	Dark	Skyfall	Silence of	Ghost
	pearl	Nemo			Knight		lambs	ship
Ship	1	0	1	1	0	0	0	1
Jack	1	0	0	1	0	0	0	0
Bond	0	0	0	0	0	1	0	0
Gun	1	0	0	0	1	1	1	0
Ocean	1	1	1	1	0	0	0	1
Captain	1	0	1	1	0	0	0	0
Batman	0	0	0	0	1	0	0	0
Crime	0	0	0	0	1	0	1	0

- The entry (w, d) = 1 if and only if the word w is present in document d
- Terms are dimensions of this matrix (units of index; we will discuss later)
- Commonly called the term document matrix
- Term and word are not same, though often words are used as terms

Boolean retrieval

	Black	Finding	Tintin	Titanic	Dark	Skyfall	Silence of	Ghost
	pearl	Nemo			Knight		lambs	ship
Ship	1	0	1	1	0	0	0	1
Jack	1	0	0	1	0	0	0	0
Bond	0	0	0	0	0	1	0	0
Gun	1	0	0	0	1	1	1	0
Ocean	1	1	1	1	0	0	0	1
Captain	1	0	1	1	0	0	0	0
Batman	0	0	0	0	1	0	0	0
Crime	0	0	0	0	1	0	1	0

Query: Jack

Results: 10010000

9

Boolean retrieval

	Black	Finding	Tintin	Titanic	Dark	Skyfall	Silence of	Ghost
	pearl	Nemo			Knight		lambs	ship
Ship	1	0	1	1	0	0	0	1
Jack	1	0	0	1	0	0	0	0
Bond	0	0	0	0	0	1	0	0
Gun	1	0	0	0	1	1	1	0
Ocean	1	1	1	1	0	0	0	1
Captain	1	0	1	1	0	0	0	0
Batman	0	0	0	0	1	0	0	0
Crime	0	0	0	0	1	0	1	0

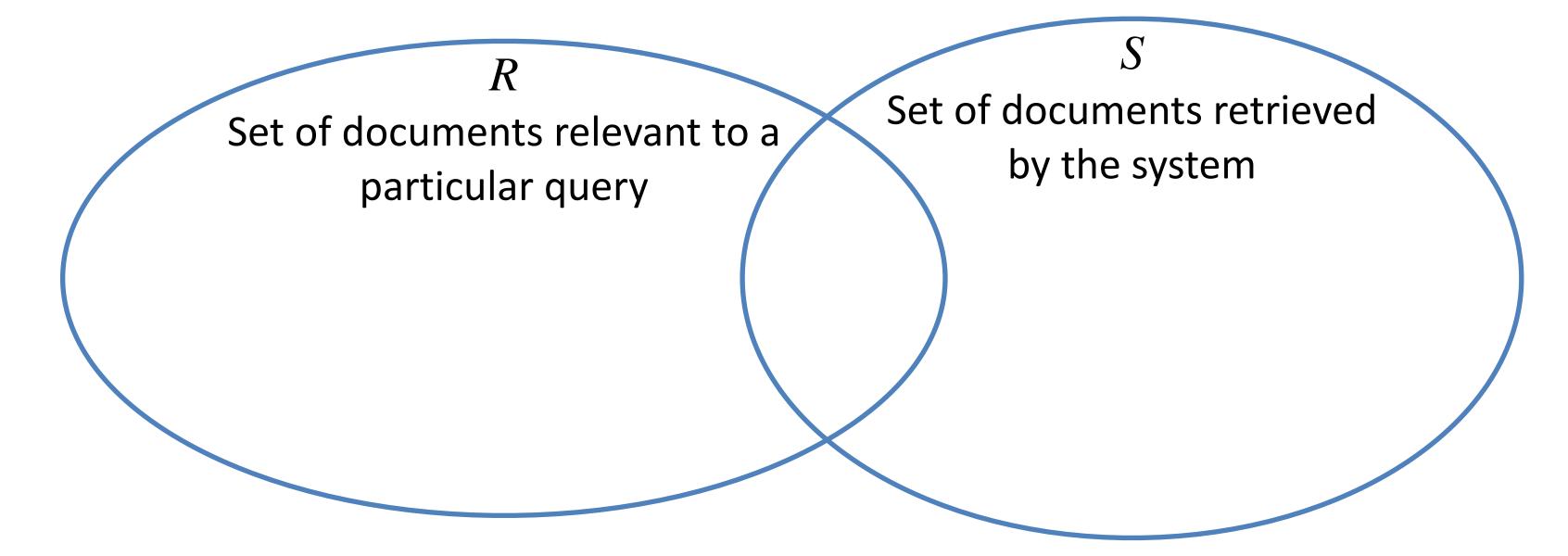
Query: Captain AND Gun

Results: 10110000 && 10001110 = 10000000

Query and relevant documents

- Query: given by user, represents the information need
 - Information need is the topic, conceptually what the user wants to know
 - Query is the representation of information need that the user conveys to the retrieval system
- Relevant document: a document that satisfies the information need, as perceived by the user
 - Merely matching the query terms does not mean a document is relevant
 - A relevant document must satisfy the actual information need

Precision and recall



What fraction of the returned results are relevant?

$$Precision = \frac{|R \cap S|}{|S|}$$

What fraction of the relevant documents in the collection were returned by the system?

$$Recall = \frac{|R \cap S|}{|R|}$$

What if the collection is "large"?

	Black	Finding	Tintin	Titanic	Dark	Skyfall	Silence of	Ghost
	pearl	Nemo			Knight		lambs	ship
Ship	1	0	1	1	0	0	0	1
Jack	1	0	0	1	0	0	0	0
Bond	0	0	0	0	0	1	0	0
Gun	1	0	0	0	1	1	1	0
Ocean	1	1	1	1	0	0	0	1
Captain	1	0	1	1	0	0	0	0
Batman	0	0	0	0	1	0	0	0
Crime	0	0	0	0	1	0	1	0

- About 1 million documents (still not so large)
- About 500,000 distinct terms
- A term document matrix of 500,000 × 1 million Boolean entries ~ 500GB

What if the collection is "large"?

	Black	Finding	Tintin	Titanic	Dark	Skyfall	Silence of	Ghost
	pearl	Nemo			Knight		lambs	ship
Ship	1	0	1	1	0	0	0	1
Jack	1	0	0	1	0	0	0	0
Bond	0	0	0	0	0	1	0	0
Gun	1	0	0	0	1	1	1	0
Ocean	1	1	1	1	0	0	0	1
Captain	1	0	1	1	0	0	0	0
Batman	0	0	0	0	1	0	0	0
Crime	0	0	0	0	1	0	1	0

Sparse matrix → inverted index

	Black pearl	Finding Nemo	Tintin	Titanic	Dark Knight	Skyfall	Silence of lambs	Ghost ship
Ship	1		1	1				1
Jack	1			1				
Bond						1		
Gun	1				1	1	1	
Ocean	1	1	1	1				1
Captain	1		1	1				
Batman					1			
Crime					1		1	

- In reality term document matrices are very sparse
- Most terms are NOT present in most documents
- For every term, store only the documents where the term is present

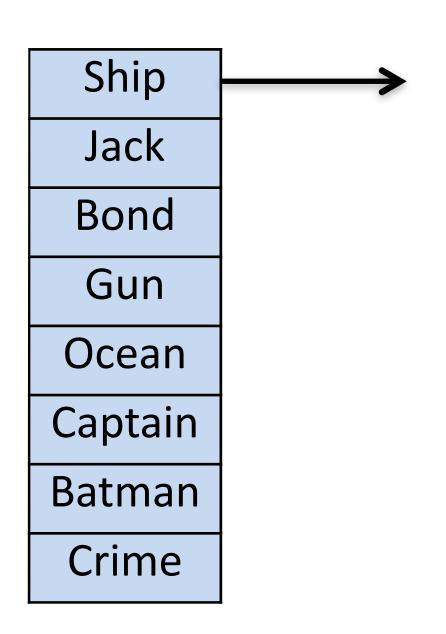
Sparse matrix — inverted index

	1	2	3	4	5	6	Cilo 7	8
	Brack	Firming	Timen	Titarric	Dark	Skyrall	Silence of	
	pearl	Nemo			Knight		lambs	ship
Ship	1		1	1				1
Jack	1			1				
Bond						1		
Gun	1				1	1	1	
Ocean	1	1	1	1				1
Captain	1		1	1				
Batman					1			
Crime					1		1	

Represent documents by document IDs

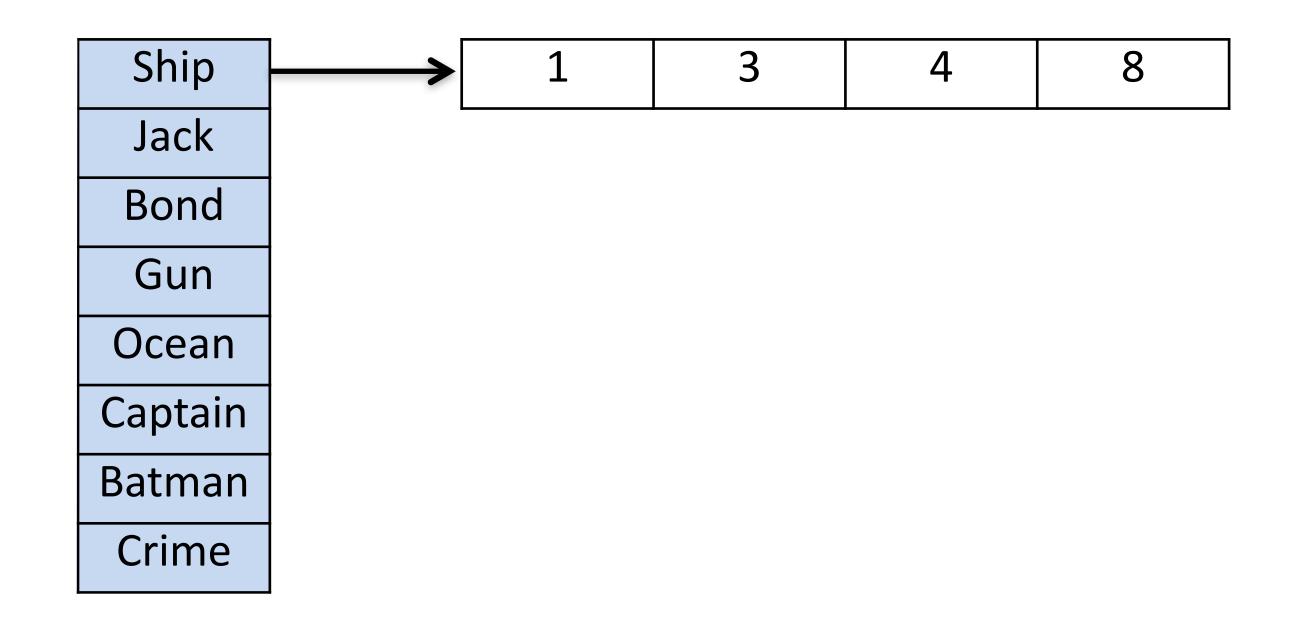
Sparse matrix —→ inverted index

1 Brack Firming Timm Titanic Dark Skyrall Silence of Gnost pearl Nemo Knight lambs ship



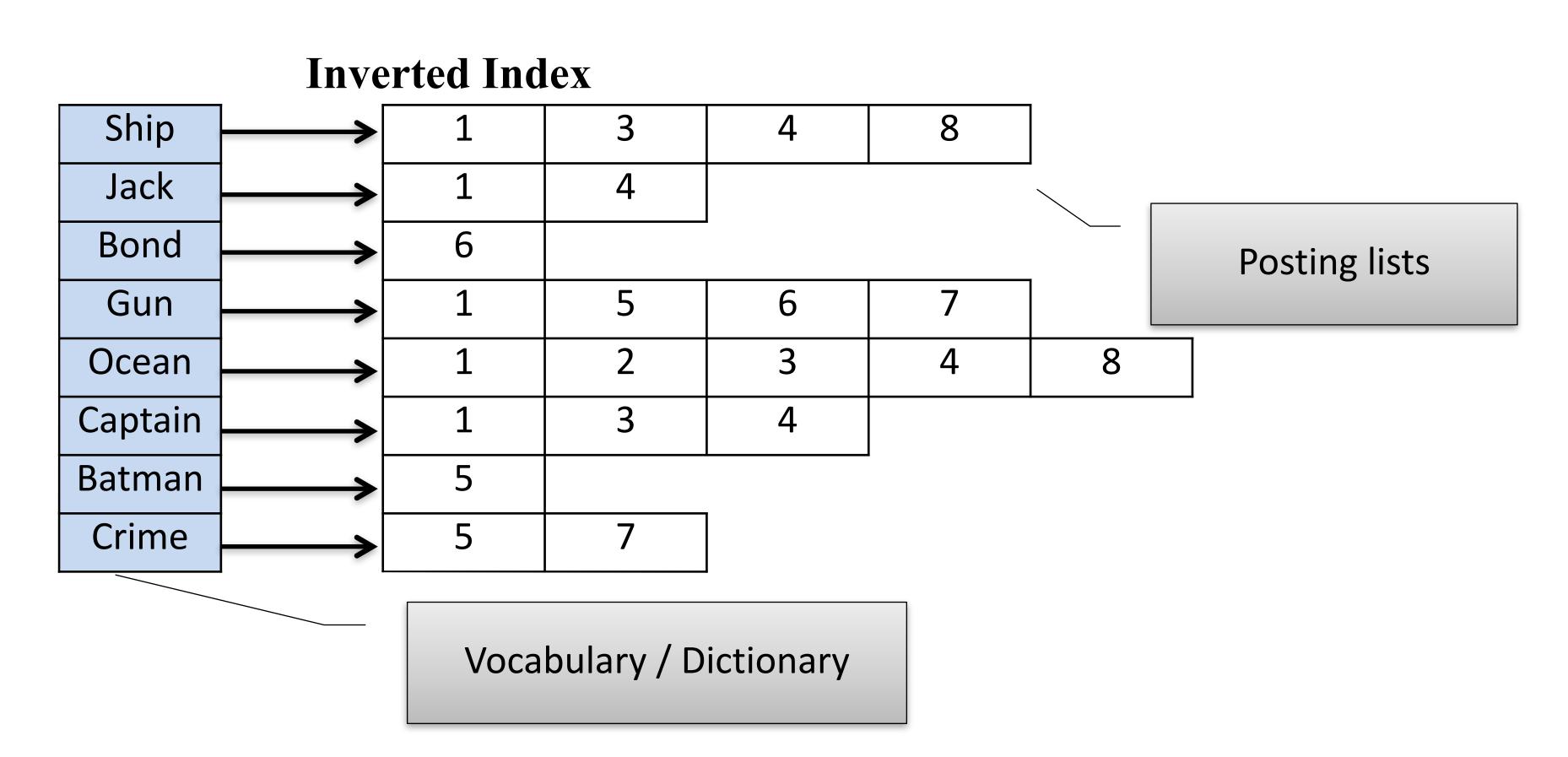
Sparse matrix —→ inverted index

1 2 3 4 5 Skyrall Silence of Grost pearl Nemo Knight lambs ship



Sparse matrix → inverted index





Creating an inverted index: basic idea

- curse of the black pearl mip Captain Jack Sparrow Caribbean Elizabeth Gun
 - Fight
- nding Nemo Ocean Fish Nemo Reef Animation
 - <u>Tintin</u> cean Animation Ship Captain Haddock Tintin
 - **Titanic** onip Rose Jack Atlantic Ocean England Sink Captain

- **The Dark Knight** Bruce Wayne Batman Joker Harvey Gordon Gun Fight Crime
- **Skyfall** 007 James Bond MI6 Gun Fight
- nce of the Lambs ⊓annibal Lector FBI Crime Gun Cannibal
- **The Ghost Ship** Ship Ghost Ocean Death Horror

- For each document, write out pairs (term, docid)
- Sort by term, then group

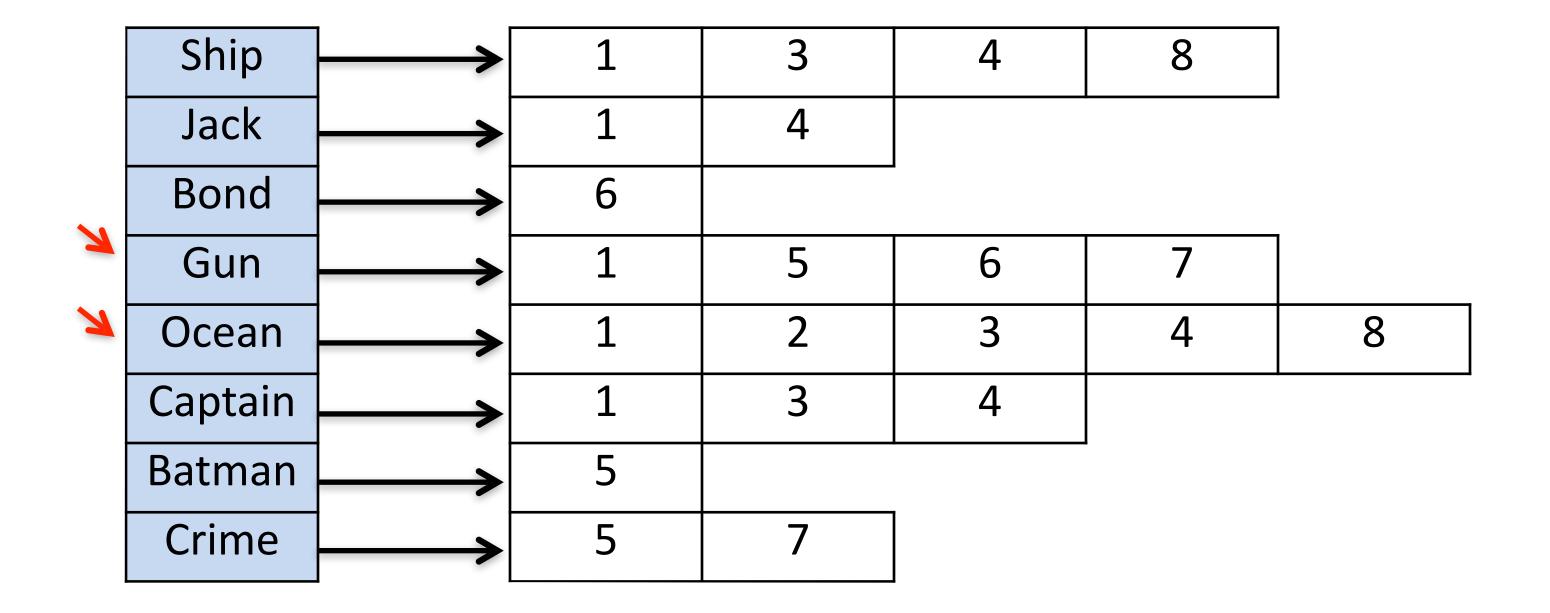
Term	docId
Ship	1
Captain	1
Jack	1
• • •	•••
Ship	3
Tintin	3
• • •	• • •
Jack	4
•••	•••

Group by term

Term	docId
• • •	• • •
Captain	1
• • •	•••
Jack	1
Jack	4
•••	•••
Ship	1
Ship	3
• • •	• • •

Term	docId	docId	docId
Captain	1	• • •	
Jack	1	4	• • •
Ship	1	3	• • •
• • •	• • •		

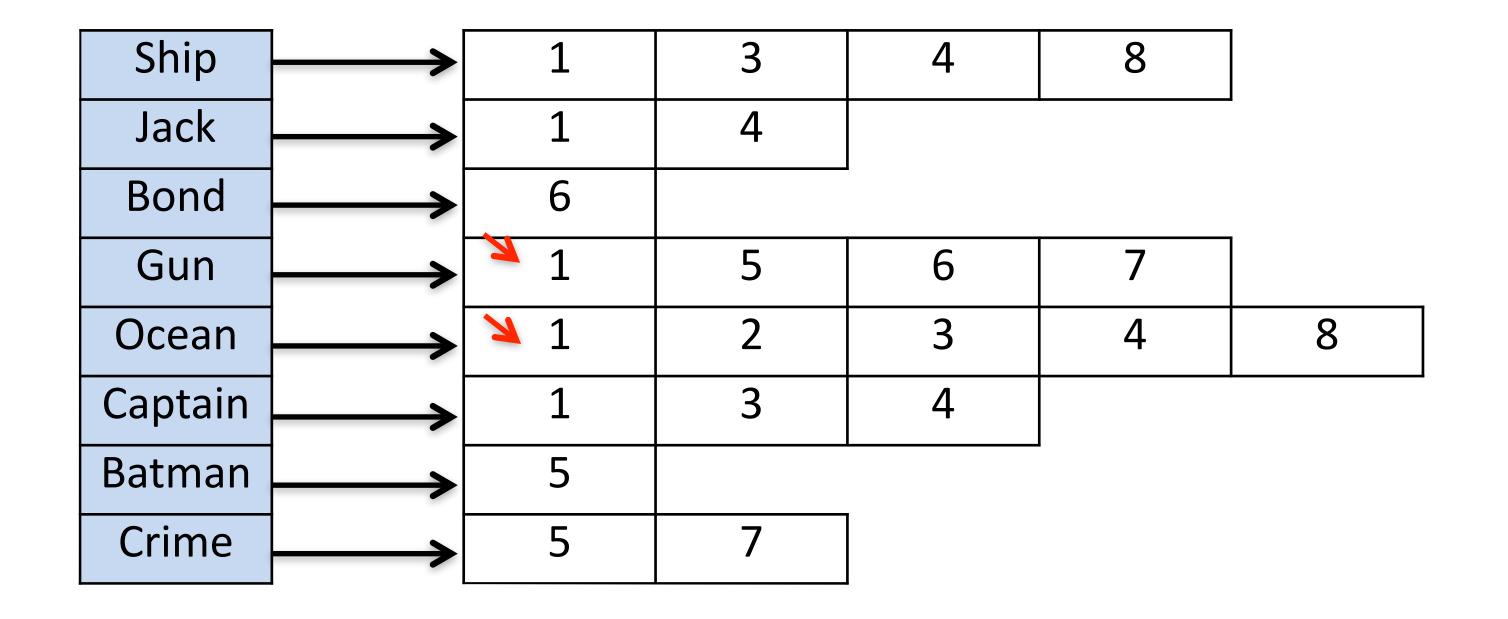
Query: Gun OR Ocean



Need to perform merge union of the two lists sorted by document ID

Start with a pointer at the beginning of each list

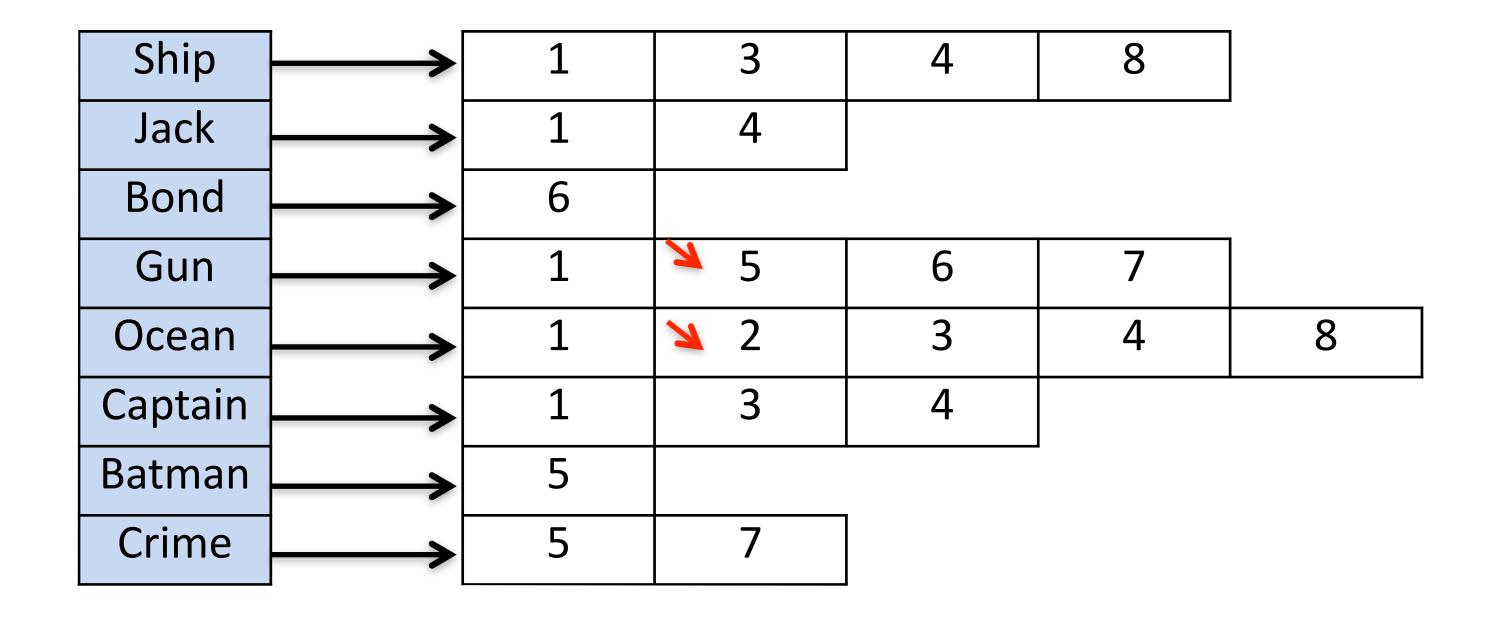
Query: Gun OR Ocean



Both doc IDs are same ⇒ add to result list and advance pointers in both lists

Results: 1

Query: Gun OR Ocean



Doc IDs are not same

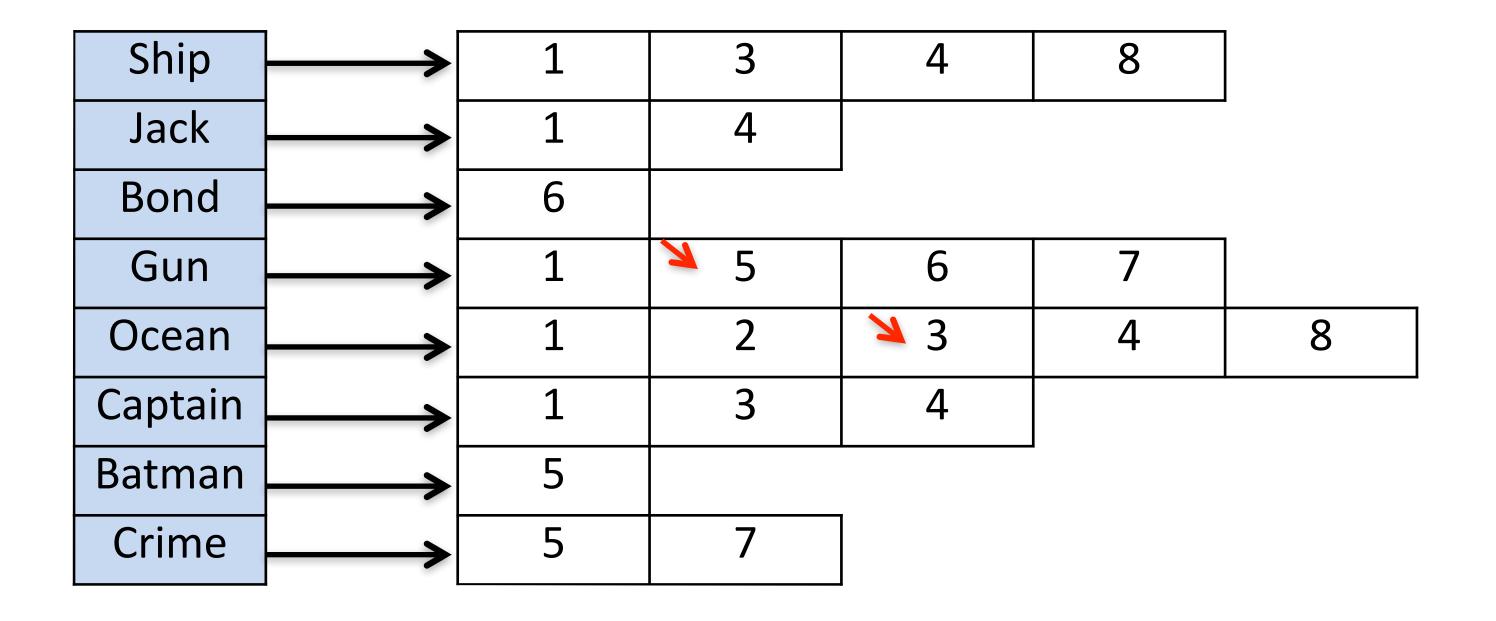
⇒ add the smaller ID

to result list and

advance only in that list

Results: 1 2

Query: Gun OR Ocean



Doc IDs are not same

⇒ add the smaller ID

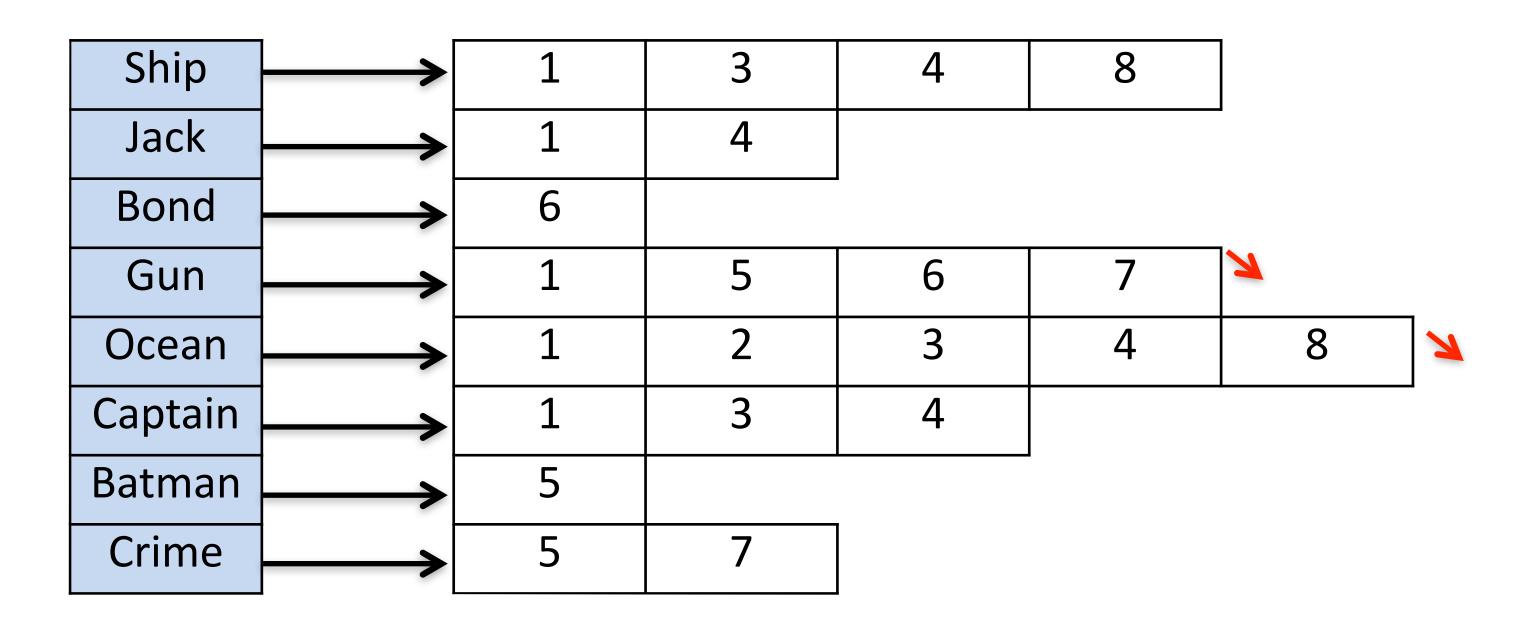
to result list and

advance only in that list

Results: 1 2 3

Results:

Query: Gun OR Ocean



Final result

O(n) algorithm if lists are of length O(n)

Merge intersection works in a similar way

Boolean retrieval use cases

Westlaw (<u>www.westlaw.com</u>)

- Largest commercial legal document search
- Tens of TB of text data
- Half a million users, million queries a day

Examples of queries:

- Information need: cases about a host's responsibility for drunk guests
- Example query: host! /p (responsib! liab!) /p (intoxicat! drunk!) /p guest

References

<u>Christopher D. Manning</u>, <u>Prabhakar Raghavan</u> and <u>Hinrich Schütze</u>. "<u>Introduction to Information</u>
 <u>Retrieval</u>", Cambridge University Press. 2008.