

Querying

Introduction to Information Retrieval

INF 141/ CS 121

Donald J. Patterson

Content adapted from Hinrich Schütze

<http://www.informationretrieval.org>



Overview

- Boolean Retrieval
- Weighted Boolean Retrieval
- Zone Indices
- Term Frequency Metrics
- The full vector space model



From the bottom

- “Grep”
 - Querying without an index or a crawl
 - Whenever you want to find something you look through the entire document for it.
 - Example:
 - You have the collected works of Shakespeare on disk
 - You want to know which play contains the words
 - “Brutus AND Caesar”



- “Grep”
 - “Brutus AND Caesar” is the **query**.
 - This is a **boolean query**. Why?
 - What other operators could be used?
 - The grep solution:
 - Read all the files and all the text and output the intersection of the files



- “Grep”
 - Slow for large corpora
 - Calculating “NOT” requires exhaustive scanning
 - Some operations not feasible
 - Query: “Romans NEAR Countrymen”
 - Doesn’t support ranked retrieval
- Moving beyond grep is the motivation for the **inverted index**.



Our **inverted index** is a 2-D array or Matrix

A Column For Each Document

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Anthony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0
...						

A Row for Each Word (or "Term")

