Querying
Introduction to Information Retrieval
INF 141
Donald J. Patterson

Content adapted from Hinrich Schütze
http://www.informationretrieval.org
Querying

Overview

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

From the bottom
Querying

From the bottom

- "Grep"
From the bottom

• “Grep”

• Querying without an index or a crawl
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.
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:
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
Querying

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
Querying

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"
Querying
Querying

- “Grep”
• “Grep”

• “Brutus AND Caesar” is the query.
“Grep”

“Brutus AND Caesar” is the query.

This is a boolean query. Why?
“Grep”

“Brutus AND Caesar” is the query.

This is a boolean query. Why?

What other operators could be used?
“Grep”

“Brutus AND Caesar” is the query.

This is a boolean query. Why?

What other operators could be used?

The grep solution:
• “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
Querying
“Grep”
Querying

- "Grep"
- Slow for large corpora
Querying

- "Grep"
  - Slow for large corpora
  - Calculating "NOT" requires exhaustive scanning
Querying

- “Grep”
  - Slow for large corpora
  - Calculating “NOT” requires exhaustive scanning
  - Some operations not feasible
Querying

- "Grep"
  - Slow for large corpora
  - Calculating "NOT" requires exhaustive scanning
  - Some operations not feasible
    - Query: "Romans NEAR Countrymen"
• “Grep”
  • Slow for large corpora
  • Calculating “NOT” requires exhaustive scanning
  • Some operations not feasible
    • Query: “Romans NEAR Countrymen”
  • Doesn’t support ranked retrieval
Querying

• “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.