Search Text

- Web search
- Desktop search
- Applications
  - Search posts in a bulletin board
  - Search product descriptions at an online retailer
  - ...

Database Query

- Find the posts regarding "SSHD login errors".
  
  ```sql
  select * from posts
  where content like '%%SSHD login errors%%';
  ```

  Here are the steps to take to fix the SSHD login errors:
  ...

  Please help! I got SSHD login errors!

Problems with Database Queries

- Please help! I got an error when I tried to login through SSHD!
- There a problem recently discovered regarding SSHD and login. The error message is usually ...
- The solution for sshd/login errors: ...

- And how about performance??
Accuracy of FTS

Precision = \frac{\# \text{ of relevant documents retrieved}}{\# \text{ of documents retrieved}}

Recall = \frac{\# \text{ of relevant documents retrieved}}{\# \text{ of relevant documents}}

Journey of a Document

Document

- **Original**
  
  ```html
  <html>
  <body>
  <p>
  The solution for sshd/login errors: ...
  </p>
  </body>
  </html>
  ```

- **Text-only**
  
  The solution for sshd/login errors: ...

Tokenizing

- [the] [solution] [for] [sshd] [login] [errors]
- ...

Stop Words

- **Words that do not help in search and retrieval**
  - Function words: a, an, and, the, of, for ...
  - Domain specific: “to be or not to be”
- **After stop words removal:**
  
  ![Stop Words Example](image)

Stemming

- **Reduce a word to its stem or root form.**
- **Examples:**
  
  - connection, connections \rightarrow connect
  - [solution] [sshd] [login] [errors] \rightarrow [solve] [sshd] [login] [error]
Inverted Index

Query Processing

FTS Implementations

Lucene Overview

Example 1: Index Text Files

Directory
Document

A document consists of a number of user-defined fields

<table>
<thead>
<tr>
<th>Field</th>
<th>Stored</th>
<th>Indexed</th>
<th>Analyzed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title: FTS with Lucene</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Author: Chengyu Sun</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content: lots of words ... lots of words ...</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

Analyzer

Pre-processing the document or query text – tokenization, stop words removal, stemming ...

Lucene built-in analyzers
- WhitespaceAnalyzer, SimpleAnalyzer, StopAnalyzer
- StandardAnalyzer
  - Grammar-based
  - Recognize special tokens such as email addresses
  - Handle CJK text

Analyzer Chinese Text

Unigram
- Lucene StandardAnalyzer
- MySQL, PostgreSQL

Bigram
- Lucene CJKAnalyzer

Grammar-based
- Usually in commercial products

Chinese Text Example

Text: 今天天气不错。

Unigram: [今] [天] [气] [不] [错]

Bigram: [今天] [天天] [天气] [气不] [不错]

Grammar-based: [今天] [天气] [不错]

IndexWriter

- addDocument( Document )
- close()
- optimize()
Example 2: Search

◆ Query and QueryParser
◆ IndexSearcher
◆ Hits
◆ Document (again)

Query and QueryParser

Query ::= ( Clause )*
Clause ::= ["+", "]" [ <TERM> : ] ( <TERM> | "" Query "" )

Sample Queries

full text search
+full +text search
+full +text –search
+title:"text search"
+(title:full title:text) -author:"bob dole"

IndexSearcher

◆ search( Query )
◆ close()

Hits

◆ A ranked list of documents used to hold search results
◆ Methods
  ○ Document doc( int n )
  ○ int id( int n )
  ○ int length()
  ○ float score( int n )

Document (again)

◆ Methods to retrieve data stored in the document
  ○ String get( String name )
  ○ Field getField( String name )
## Handle Rich Text Documents

- **HTML**
  - NekoHTML, JTidy, TagSoup
- **PDF**
  - PDFBox
- **MS Word**
  - TextMining

## Example: FTS in Evelyn

- **Indexer and Searcher interface**
- **FileHandler interface**
- **File handler implementations**
  - DefaultFileHandler
  - TextFileHandler
  - HtmlFileHandler
  - PdfFileHandler
- **Spring beans configuration**

## Further Readings

- *Lucene in Action* by Otis Gospodnetic and Erik Hatcher