Full Text Search

By Fred Katona

https://github.com/bigkat73/ru-findit

https://github.com/bigkat73/indexer_demo

Full Text Search

- One Solution...iterate over collection searching each document using regex
- Or write sql to wildcard search

Select * from documents where body like '%fred%';

Full Text Search

- Drawbacks?
- When to use this solution?

Full Text Search

- Problem Overview
  - Lots of document data
  - Want to find relevant documents that contain word or set of words

Full Text Search

- Solution

Full Text Search

- Solution

Full Text Search

- The first thing we need is an index
- What kind?

{ }

Full Text Search

- In our index what we need is a key...
- Use the word as the key
  - Leave out things like stemming the word
  - Synonyms
  - Etc...

Full Text Search

Final Index Structure

{ “word”: total_frequency,

{ “document_id”: “term_frequency”}

}

Full Text Search

Process

1) Have a collection of documents
2) For each document, split into tokens
3) Add each token to an hash index

Full Text Search

- Catalog Overview
- Indexer Overview
- Tokenizer Overview
- Stopper Overview

Full Text Search

- ActiveRecord Integration Overview
- Simple DSL

```
class Document < ActiveRecord::Base
        include RuFindit::Model::Searcher
                indexable do
                        indexes :field_name
                        ....
                end
```

Full Text Search

- Demo

Full Text Search

Open Source Options

- Both built on Lucene

Solr

- advantages – maturity, query language is fairly simple
- disadvantages – more complex setup, need for pipline

Elastic Search

- advantages – restful interface, distributed setup out of box, backups are easy
- disadvantages – one key developer

Full Text Search

Both indexers require adapters from rails

For either one, the dsl's are fairly robust. Sunspot, however, is currently missing the wildcard search....

search 'straw*'

- Needs a patch

Other Uses

- Text Mining
- Web based-document search
- Clustering of documents(finding patterns)