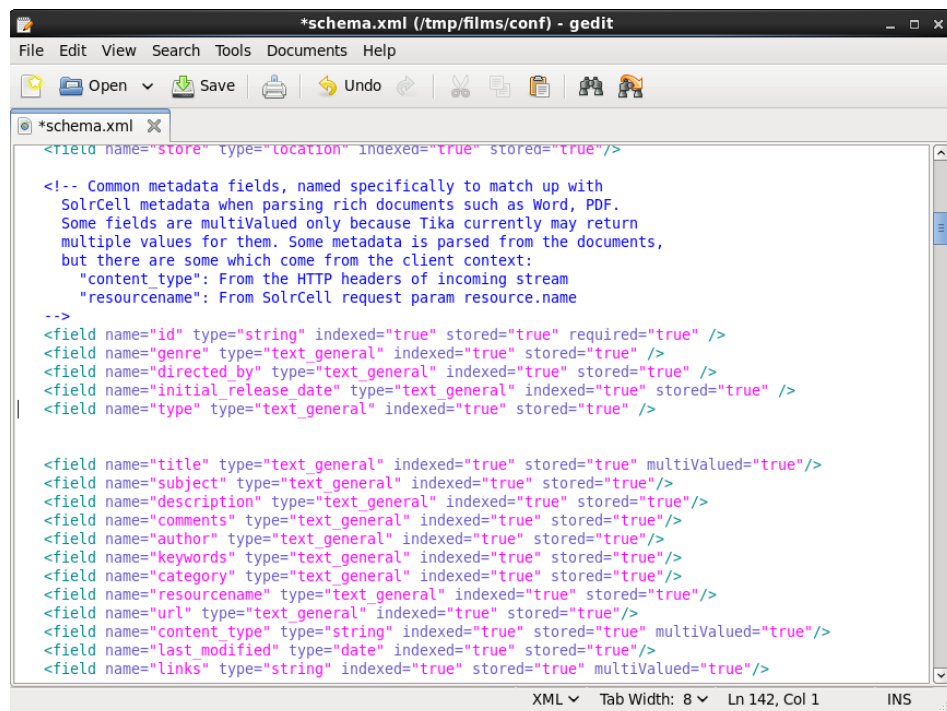Uyen Dang

CS 5590 Summer 2020

PART 1: FILMS DATASET

Step 1: Create instancedir and edit the schema.xml

```
[cloudera@quickstart ~]$ solrctl instancedir --generate /tmp/films
Error: subdirectory /tmp/films already exists
[cloudera@quickstart ~]$ ls /tmp/films/
conf
[cloudera@quickstart ~]$ ls /tmp/films/conf/
admin-extra.html                        _schema_analysis_synonyms_english.json
admin-extra.menu-bottom.html            schema.xml
admin-extra.menu-top.html               schema.xml~
clustering                              scripts.conf
currency.xml                            solrconfig.xml
elevate.xml                             solrconfig.xml.secure
lang                                    spellings.txt
mapping-FoldToASCII.txt                 stopwords.txt
mapping-ISOLatin1Accent.txt             synonyms.txt
protwords.txt                           update-script.js
_rest_managed.json                      velocity
_schema_analysis_stopwords_english.json xslt
[cloudera@quickstart ~]$ gedit /tmp/films/conf/schema.xml&
[1] 11758
[cloudera@quickstart ~]$
```

```
*schema.xml (/tmp/films/conf) - gedit                          _ □ ×
File  Edit  View  Search  Tools  Documents  Help

  Open ∨   Save          Undo              ✂            🔍 🔍

 *schema.xml ✕

    <field name="store" type="location" indexed="true" stored="true"/>

    <!-- Common metadata fields, named specifically to match up with
      SolrCell metadata when parsing rich documents such as Word, PDF.
      Some fields are multiValued only because Tika currently may return
      multiple values for them. Some metadata is parsed from the documents,
      but there are some which come from the client context:
        "content_type": From the HTTP headers of incoming stream
        "resourcename": From SolrCell request param resource.name
    -->
    <field name="id" type="string" indexed="true" stored="true" required="true" />
    <field name="genre" type="text_general" indexed="true" stored="true" />
    <field name="directed_by" type="text_general" indexed="true" stored="true" />
    <field name="initial_release_date" type="text_general" indexed="true" stored="true" />
    <field name="type" type="text_general" indexed="true" stored="true" />


    <field name="title" type="text_general" indexed="true" stored="true" multiValued="true"/>
    <field name="subject" type="text_general" indexed="true" stored="true"/>
    <field name="description" type="text_general" indexed="true" stored="true"/>
    <field name="comments" type="text_general" indexed="true" stored="true"/>
    <field name="author" type="text_general" indexed="true" stored="true"/>
    <field name="keywords" type="text_general" indexed="true" stored="true"/>
    <field name="category" type="text_general" indexed="true" stored="true"/>
    <field name="resourcename" type="text_general" indexed="true" stored="true"/>
    <field name="url" type="text_general" indexed="true" stored="true"/>
    <field name="content_type" type="string" indexed="true" stored="true" multiValued="true"/>
    <field name="last_modified" type="date" indexed="true" stored="true"/>
    <field name="links" type="string" indexed="true" stored="true" multiValued="true"/>

                              XML ∨   Tab Width: 8 ∨   Ln 142, Col 1        INS
```

Step 2: Create a new instancedir and its collection

```
[cloudera@quickstart ~]$ solrctl instancedir --create films /tmp/films
Uploading configs from /tmp/films/conf to quickstart.cloudera:2181/solr. This ma
y take up to a minute.
[1]+  Done                    gedit /tmp/films/conf/schema.xml
[cloudera@quickstart ~]$ solrctl collection --create films
 <?xml version="1.0" encoding="UTF-8"?> <response> <lst name="responseHeader"> <
int name="status"> 0</int> <int name="QTime"> 13406</int> </lst> <lst name="fail
ure"> <str> org.apache.solr.client.solrj.impl.HttpSolrServer$RemoteSolrException
:Error CREATEing SolrCore 'films_shard1_replica1': Unable to create core [films_
shard1_replica1] Caused by: [schema.xml] Duplicate field definition for 'id' [[[
id{type=string,properties=indexed,stored,omitNorms,omitTermFreqAndPositions,sort
MissingLast,required, required=true}]]] and [[[id{type=string,properties=indexed
,stored,omitNorms,omitTermFreqAndPositions,sortMissingLast,required, required=tr
ue}]]]</str> </lst> </response>
```

Step 3: Open Solr web interface and start working on the films dataset

QUERY 1 - Keyword matching: Print movies with genres = drama

QUERY 2 - Wildcard matching: The search string "A?" would match both test and text.

QUERRY 3 - Proximity matching: print movie name "air" within 3 words of each other in a document.

QUERY 4 - Range searches: movie release before 2006-12-10

QUERRY 5 - Fuzzy logic: Select movie contain word "air" 50%

Apache **Solr**

- Dashboard
- Logging
- Cloud
- Core Admin
- Java Properties
- Thread Dump

films_shard1_r...

- Overview
- Analysis
- Dataimport
- Documents
- Files
- Ping
- Plugins / Stats
- Query
- Replication
- Schema Browser

Request-Handler (qt)
`/select`

— common

q
`name:air ~0.5`

fq
`[ ]` ⊖ ⊕

sort
`name asc`

start, rows
`0` `1001`

fl
` `

df
` `

Raw Query Parameters
`key1=val1&key2=val2`

wt
`json`

☑ indent
☐ debugQuery

☐ dismax
☐ **e**dismax
☐ hl
☐ facet
☐ spatial
☐ spellcheck

Execute Query

http://quickstart.cloudera:8983/solr/films_shard1_replica1/select?q=name%3Aair+~0.5&sort=name+asc&r
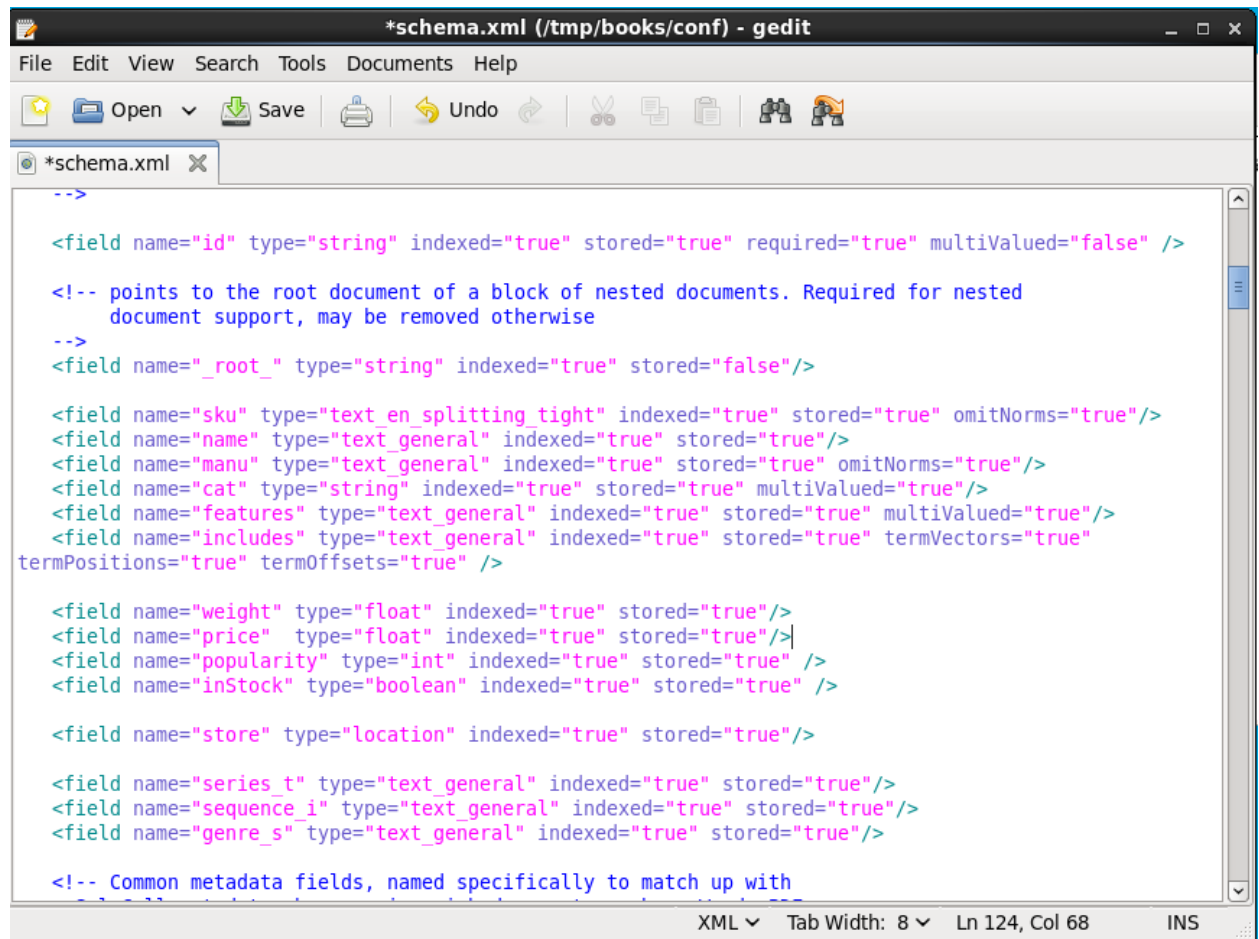
```
{
  "responseHeader": {
    "status": 0,
    "QTime": 247,
    "params": {
      "sort": "name asc",
      "indent": "true",
      "q": "name:air ~0.5",
      "_": "1593042817447",
      "wt": "json",
      "rows": "1001"
    }
  },
  "response": {
    "numFound": 5,
    "start": 0,
    "docs": [
      {
        "name": "Air",
        "directed_by": "Osamu Dezaki",
        "genre": "Fantasy|Anime|Animation|Japanese Movies|Drama",
        "id": "/en/air_2005",
        "initial_release_date": "2005-02-05",
        "_version_": 1670425374119428000
      },
      {
        "name": "Air Bud: Spikes Back",
        "directed_by": "Mike Southon",
        "genre": "Family|Sports|Comedy",
        "id": "/en/air_bud_spikes_back",
        "initial_release_date": "2003-06-24",
        "_version_": 1670425374121525200
      },
      {
        "name": "Air Buddies",
        "directed_by": "Robert Vince",
        "genre": "Family|Animal Picture|Children's/Family|Family-Oriented Adventure|Comedy",
        "id": "/en/air_buddies",
```

PART 2: BOOKS DATASET

Step 1: Create instancedir and edit the schema.xml

```
[cloudera@quickstart ~]$ solrctl instancedir --generate /tmp/books
[cloudera@quickstart ~]$ gedit /tmp/books/conf/schema.xml&
[1] 13923
```


Step 2: Create a new books instancedir.

```
[cloudera@quickstart ~]$ solrctl instancedir --create books /tmp/books
Uploading configs from /tmp/books/conf to quickstart.cloudera:2181/solr. This ma
y take up to a minute.
[cloudera@quickstart ~]$ solrctl collection --create books
```

Step 3: Go to web browser and open solr

QUERRY 1 - Keyword matching: find books in which its name contain "the"

QUERRY 2 - Wildcard matching: name:th*

QUERRY 3 - Proximity matching: with name contain "the"

QUERRY 4 - Range searches: search books with price between 0 and 7.99

QUERRY 5 - Fuzzy logic: select book name contain "Book" with 80% similarity.