



## Configuração de um sistema de recuperação de informação com Solr

Marcelo Hiroshi Noguti – 6793181

Walter Perez Urcia – 9410313

Armazenamento e Recuperação de Informação

Instituto de Matemática e Estatística da Universidade de São Paulo

Dezembro de 2015

## Organização do documento

O presente documento tem como objetivo reportar o que foi feito durante o projeto final da disciplina de Armazenamento e Recuperação de Informação.

A primeira seção contém os relatos a respeito do momento durante a familiarização com o *Solr*, suas funcionalidades e componentes.

A segunda seção contém os relatos a respeito do desenvolvimento do projeto propriamente dito. Aqui são mostradas as diversas configurações feitas nos arquivos de configuração do *Solr* para que o sistema de recuperação de informação desenvolvido pudesse funcionar. Nessa seção também são relatados os testes realizados sobre sistema de recuperação configurado, bem como os resultados obtidos.

# Conhecendo o Solr

## Configurações utilizadas para instalação

Para o projeto foi utilizado o pacote *solr-5.3.1.tgz*, disponível em

<http://ftp.unicamp.br/pub/apache/lucene/solr/5.3.1/>

A máquina foi configurada com *Java 1.7.0\_80*, que é a versão mínima recomendada para executar essa versão do *Solr*. Uma lista completa de versões de *JRE* pode ser encontrada em

<http://www.oracle.com/technetwork/java/javase/downloads/java-archive-downloads-javase7-521261.html>.

## Instalação

Após fazer o download do arquivo *solr-5.3.1.tgz*, foi preciso descompactar o mesmo. A descompactação foi feita no diretório raiz. Considerando que o arquivo *solr-5.3.1.tgz* já se encontrava no diretório raiz, executar o comando

```
tar -xvzf solr-5.3.1.tgz
```

foi o suficiente para realizar a descompactação.

Em seguida, bastou iniciar o *Solr* pelos comandos

```
cd solr-5.3.1
```

```
bin/solr start
```

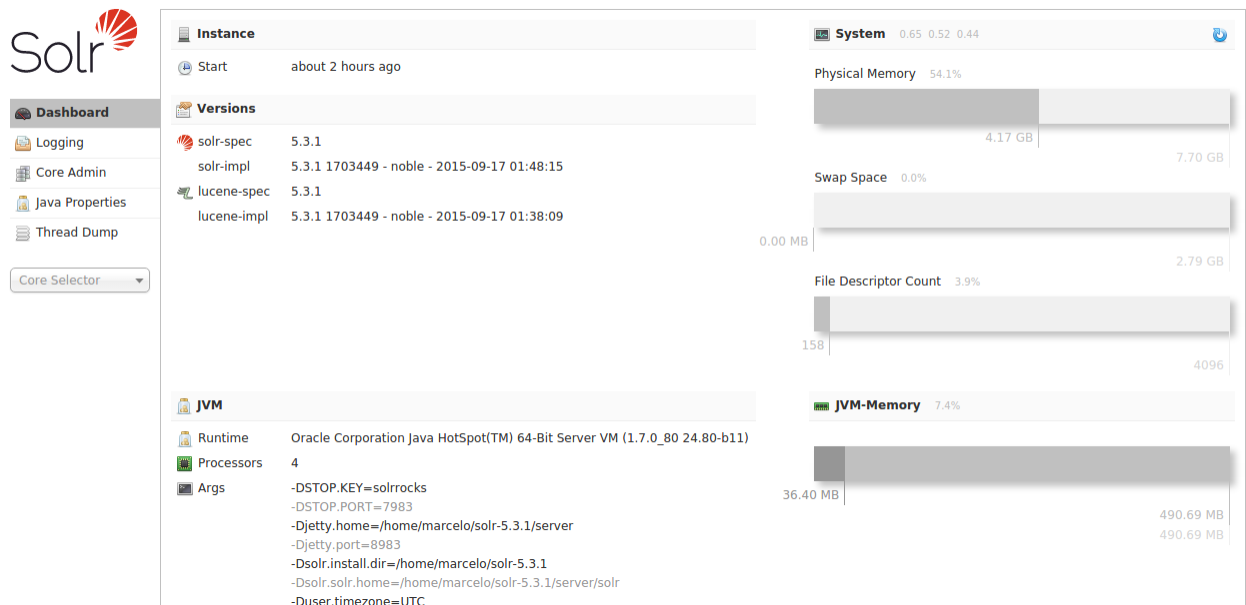
Por padrão, o *Solr* é iniciado e passa a ouvir em duas portas, sendo uma delas a 8983.

Em seguida, foi verificado se o servidor do *Solr* realmente foi iniciado e estava em execução.

```
bin/solr status
```

Como a resposta foi positiva, foi possível acessar o painel administrativo do *Solr* com interface gráfica pelo navegador

<http://localhost:8983/solr/>



## Desligando o Solr

Durante o desenvolvimento do projeto, várias vezes foi necessário recomeçar com novas coleções de documentos e configurações. Para isso, os comandos a seguir foram bastante úteis

```
bin/solr stop -all ; rm -Rf <diretorio_da_colecao>
```

## Indexando documentos

Como visto em aula, um sistema de recuperação precisa indexar os documentos de uma coleção para que posteriormente consultas possam ser executadas sobre essa coleção.

No *Solr*, é possível adicionar arquivos a uma coleção pelo comando

```
bin/post -c gettingstarted docs/
```

onde *gettingstarted* é a coleção onde os novos documentos devem ser adicionados e indexados; e *docs/* é o local onde estão os diversos documentos (arquivos a serem adicionados à coleção).

Para um teste inicial, foram indexados alguns arquivos numa nova coleção chamada *teste*. Tais arquivos são os providos pelo próprio *Solr* para caráter de testes.

```
bin/post -c teste docs/
```

É possível verificar pela interface administrativa do *Solr* que os arquivos foram adicionado à coleção e indexados.

**Solr**

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

teste

**Overview**

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

**Statistics**

Last Modified: 9 minutes ago  
 Num Docs: 3774  
 Max Doc: 3774  
 Heap Memory: -1  
 Usage:  
 Deleted Docs: 0  
 Version: 26  
 Segment Count: 7  
 Optimized:   
 Current:

**Replication (Master)**

	Version	Gen	Size
Master (Searching)	1447511797722	6	15.34 MB
Master (Replicable)	1447511797722	6	-

**Admin Extra**

We found no "admin-extra.html" file.

**Instance**

CWD: /home/marcelo/solr-5.3.1/server  
 Instance: /home/marcelo/solr-5.3.1/server/solr/teste  
 Data: /home/marcelo/solr-5.3.1/server/solr/teste/data  
 Index: /home/marcelo/solr-5.3.1/server/solr/teste/data/index  
 Impl: org.apache.solr.core.NRTCachingDirectoryFactory

**Healthcheck**

Ping request handler is not configured with a healthcheck file.

[Documentation](#) [Issue Tracker](#) [IRC Channel](#) [Community forum](#) [Solr Query Syntax](#)

Assim como visto em aula, cada documento possui um identificador único. No caso da coleção de testes do exemplo, temos um total de 3774 documentos indexados (esse número pode ser diferente do número de arquivos, já que um arquivo pode conter mais de um documento).

Podemos observar que existe outra variável max Doc dizendo que há 3774 documentos. Esse número também pode ser diferente devido ao fato que documentos podem ser removidos da coleção, mas a indexação não ser refeita novamente. Nesse caso, a técnica utilizada é a de listas auxiliares, como vista em aula, indicando quais documentos não pertencem mais à coleção.

Um outro detalhe interessante é o menu Files, que possui arquivos básicos de configuração, como listas de *stopwords* em diferentes idiomas ou mapeamento de palavras sinônimas.

**Solr**

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

teste

**Overview**

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

currency.xml  
 elevate.xml  
 lang  
 contractions\_ca.txt  
 contractions\_fr.txt  
 contractions\_ga.txt  
 contractions\_it.txt  
 hyphenations\_ga.txt  
 stemdict\_nl.txt  
 stoptags\_ja.txt  
 stopwords\_ar.txt  
 stopwords\_bg.txt  
 stopwords\_ca.txt  
 stopwords\_cz.txt  
 stopwords\_da.txt  
 stopwords\_de.txt  
 stopwords\_el.txt  
 stopwords\_en.txt  
 stopwords\_es.txt  
 stopwords\_eu.txt  
 stopwords\_fa.txt  
 stopwords\_fi.txt  
 stopwords\_fr.txt  
 stopwords\_ga.txt  
 stopwords\_gl.txt  
 stopwords\_hi.txt  
 stopwords\_hu.txt  
 stopwords\_hy.txt  
 stopwords\_id.txt

[http://localhost:8983/solr/teste/admin/file?file=lang/stopwords\\_en.txt&contentType=text/plain; charset=utf-8](http://localhost:8983/solr/teste/admin/file?file=lang/stopwords_en.txt&contentType=text/plain; charset=utf-8)

```
# Licensed to the Apache Software Foundation (ASF) under one or more
# contributor license agreements. See the NOTICE file distributed with
# this work for additional information regarding copyright ownership.
# The ASF licenses this file to You under the Apache License, Version 2.0
# (the "License"); you may not use this file except in compliance with
# the License. You may obtain a copy of the License at
#
# http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.
#
# a couple of test stopwords to test that the words are really being
# configured from this file:
stopwords
stopwords

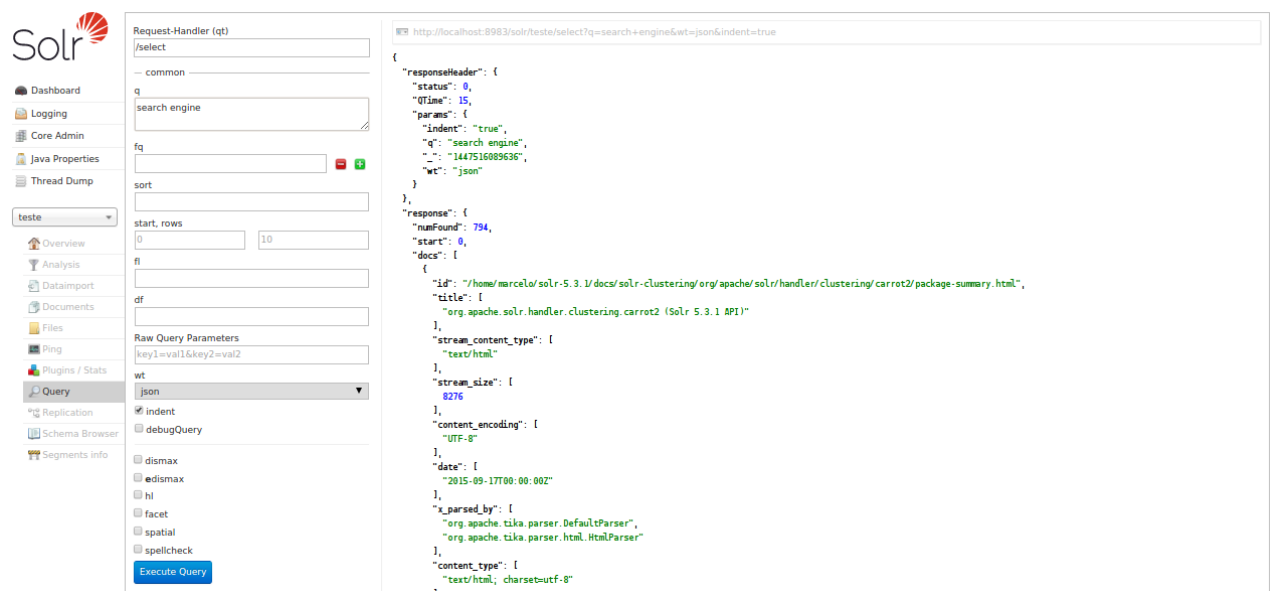
# Standard english stop words taken from Lucene's StopAnalyzer
a
an
and
are
as
at
be
but
by
for
```

Para a lista de *stopwords* em português, por padrão são listadas as principais preposições e artigos, assim como alguns pronomes e conjunções. Há também alguns verbos e suas principais conjugações.

A lista de sinônimos por padrão é inicializada quase vazia, mas ela é importante para que, por exemplo, consultas com termos como "carro" também levem em consideração documentos que possuem o termo "veículo", como discutido em sala.

## Consultas

O menu Query oferece uma interface para realizar consultas e configurar os diversos parâmetros que a mesma pode ter, como correção de digitação (*spellcheck*), número de resultados por página entre outros.



The screenshot displays the Solr Query interface. On the left, a sidebar contains navigation links: Dashboard, Logging, Core Admin, Java Properties, Thread Dump, and a dropdown menu with 'teste' selected. Below these are links for Overview, Analysis, Dataimport, Documents, Files, Ping, Plugins / Stats, Query (highlighted), Replication, Schema Browser, and Segments info. The main panel is titled 'Request-Handler (qt) /select' and includes fields for 'common' (q: search engine), 'fq' (empty), 'sort' (empty), 'start, rows' (0, 10), 'fl' (empty), 'df' (empty), and 'Raw Query Parameters' (key1=val1&key2=val2). A 'wt' dropdown is set to 'json', and checkboxes for 'indent' and 'debugQuery' are present. A blue 'Execute Query' button is at the bottom. The right panel shows the resulting JSON response from the URL `http://localhost:8983/solr/teste/select?q=search+engine&wt=json&indent=true`. The response includes a 'responseHeader' with status, QTime, and params, and a 'response' object containing numFound (734), start (0), and a list of documents. The first document is a summary for 'org.apache.solr.handler.clustering.carrot2 (Solr 5.3.1 API)'.

É interessante observar que a chamada para o motor de busca é por uma *URL*. No exemplo,

`http://localhost:8983/solr/teste/select?q=search+engine&wt=json&indent=true`

é a chamada para a consulta `search engine` na coleção `teste` que foi criada anteriormente, com retorno em formato `json`. Assim sendo, é bastante intuitivo imaginar como integrar o serviço do *Solr* em aplicações próprias: é somente preciso que a aplicação construa a *URL* e faça a chamada, e então aguarde a resposta no formato desejado.

Uma maneira bastante simples de tornar isso real é criar uma página web em *PHP*, por exemplo, na qual há um campo de texto e um botão "pesquisar". Quando um usuário clicar no botão, o texto inserido no campo de entrada é utilizado para construir a *URL* de chamada para o *Solr*, que por sua vez devolve o resultado da consulta. Basta que a própria página faça a leitura do resultado e o exiba ao usuário.

## Caractere reservado

Quando uma consulta possui mais de um termo, o espaço em branco é substituído pelo caractere `+`. No entanto, em algumas ocasiões, tal caractere é necessário para a consulta, seja para montar a *string* final de busca, ou porque os termos buscados realmente possuem o caractere `+`. Para evitar conflitos, o caractere `+` é substituído por `%2B` para diferenciar do `+` que denota espaço.

## Snippets

*Snippets* são pequenas janelas de texto extraídas dos documentos que possam justificar o porquê de um documento ter sido retornado no resultado. Para que o *Solr* devolva tais *snippets*, basta que na *URL* de chamada sejam adicionados alguns campos, como

- `hl` configurado como *true*;
- `hl.fields` configurando quais campos devem ser levados em consideração para a geração de *snippets*.

`http://localhost:8983/solr/teste/select?q=engine&wt=json&indent=true&hl=true&hl.fl=title`

A chamada anterior busca pelo termo `engine` e possui a configuração para que os *snippets* sejam mostrados caso eles estejam no campo `title` de um documento.

## Paginação

Durante as aulas a respeito de *ranqueamento* do resultado retornado para uma consulta, ficou claro que o comportamento de um usuário comum era o de dar mais importância àqueles documentos que retornavam no topo da lista. Raramente usuários olhavam para documentos que estavam em posições mais baixas ou em páginas seguintes.

Por padrão, o *Solr* devolve os resultados paginados, mostrando até 10 documentos em cada página. Esse comportamento é configurável, bastando adicionar à *URL* de chamada os parâmetros `start` e `rows`, que indicam respectivamente a posição do primeiro documento e a quantidade de documentos a partir dele que devem ser mostrados.

`http://localhost:8983/solr/teste/select?q=computation&start=3&rows=2&wt=json&indent=true`

O exemplo anterior consulta por `computation` e requer dois resultados a partir da posição 3.

## Campos de documentos

Como visto logo no início do capítulo 6, os documentos possuem campos de *metadados* como nome, data etc.. É possível realizar consultas direcionando a busca para determinados campos de um documento. Basta explicitar o nome do campo e o valor desejado

`http://localhost:8983/solr/teste/select?wt=json&indent=true&q=content_encoding:UTF-8`

No caso da coleção "teste" e dos documentos indexados a ela, somente documentos que possuem o valor UTF-8 no campo `content_encoding` serão retornados no resultado.

## Zoneamento com pesos

Além de direcionar a busca a determinados campos de um documento, existe o conceito de pontuação de zonas por peso (*weighted zone scoring*). Na prática, isso significa dar mais relevância a determinados campos de um documento para um melhor *rankeamento*, como visto no capítulo 6.

```
http://localhost:8983/solr/teste/select?q=ISO-8859-1&wt=json&indent=true&defType=edismax&qf=title^2+content_encoding^1.5
```

O exemplo anterior configura a consulta para que o campo `title` de um documento seja mais relevante que o campo `content_encoding` quando procura-se por `ISO-8859-1`.

## Bigramas e Trigramas

Logo nos capítulos iniciais vimos os conceitos de bigramas e trigramas, que, na prática, é tratar dois termos (ou três termos) como sendo somente um. O *Solr* permite que os documentos sejam melhores *rankeados* com base em bigramas e trigramas presentes em determinados campos de um documento.

```
http://localhost:8983/solr/teste/select?q=apache+solr+documentation&wt=json&indent=true&defType=edismax&pf2=title%5E2
```

O exemplo anterior consulta por `apache solr documentation` e dá uma maior importância para documentos que possuem os bigramas `apache solr` e `solr documentation` no campo título.

## Spellcheck

Uma funcionalidade importante para um sistema de recuperação é o corretor. É comum que usuários digitem termos errados distraidamente ou intencionalmente (quando não sabem a grafia correta de um termo, por exemplo). Nesse caso, como visto em aula, um sistema de recuperação não é capaz de recuperar os documentos de maneira a satisfazer a consulta. Em aula vimos que uma das maneiras de corrigir erros em termos é pela distância de palavras, como a distância de Levenshtein.

O *Solr* provê mecanismos para algumas técnicas de correção. A principal delas é a correção baseada nos termos indexados dos documentos. Outra técnica é pela quebra ou combinação de palavras: quando um termo é digitado erroneamente com um espaço, dividindo-o em dois; ou quando um termo é escrito separadamente mas foi digitado sem espaços. É possível, também, prover um dicionário separadamente para que os termos nele contidos sejam utilizados para a correção de termos digitados incorretamente.



# Colocando em prática

## Criando a coleção do projeto

Para criar a coleção do projeto, primeiro foi iniciado o *Solr* e então criado um novo core chamado projeto.

```
bin/solr start
```

```
bin/solr create -c projeto
```

## Populando a coleção

Para a população da coleção do projeto foram utilizadas diversas fontes de documentos disponíveis publicamente. No caso, foi utilizada uma coleção de títulos de livros que estão disponíveis para venda por *ScienceDirect*. O arquivo no formato *.xlsx* com as informações de livros está disponível em

<https://www.elsevier.com/solutions/sciencedirect/content/book-title-lists>

Tais arquivos *.xlsx* foram convertidos para o formato *.csv* contendo somente algumas das informações de livros.

Além disso, foi usado o conjunto de documentos Reuters-21578 disponível em

<https://archive.ics.uci.edu/ml/machine-learning-databases/reuters21578-mld/>

De maneira similar, os documentos foram convertidos do formato *.sgm* para *.xml*. A conversão se deu graças a um *script* desenvolvido em *Python*.

## Configurando o projeto

Inicialmente foram criados os arquivos de configuração *schema.xml* e *solrconfig.xml* para a configuração do core. Ambos os arquivos se encontram no diretório */conf*.

Nesse mesmo diretório se encontram os diversos arquivos de configuração do core atual (sinônimos, *spellcheck*, *stopwords* etc.). Pelo painel administrativo é possível visualizar todos eles no menu Files.

Sempre que algum parâmetro nos arquivos de configuração são alterados, é preciso atualizar o core para que as novas configurações entrem em vigor. Na versão do *Solr* sendo utilizada, basta carregar a seguinte *URL* no navegador

<http://localhost:8983/solr/cidades/update?commit=true>

Nesse caso, as alterações feitas no core projeto estão sendo aplicadas. Caso as alterações não sejam aplicadas, é só tentar finalizar o *Solr* e iniciá-lo novamente.

No arquivo *schema.xml* são configurados os campos que os documentos da coleção têm, assim como os tipos dos mesmos.

É interessante notar que é possível explicitar a cópia de valores de um campo para outro, de maneira que funcionalidades do sistema de busca sejam mais eficientes. É possível explicitar que os valores dos campos *titulo* e *autor* de uma coleção de livros, por exemplo, sejam copiados para um

terceiro campo chamado *briefing*, que por sua vez é o campo com maior peso utilizado para a busca e para o corretor.

Já no arquivo *solrconfig.xml* é onde a mágica acontece. É nesse arquivo de configuração no qual são explicitados os parâmetros a serem utilizados nas diversas funcionalidades do *Solr*.

São basicamente dois conceitos os mais importantes: *searchComponent* e *requestHandler*.

- *SearchComponent*: na prática, são pequenos trechos reutilizáveis de funcionalidades;
- *RequestHandler*: reúne diversos trechos de funcionalidades (*searchComponent*), definindo a lógica para qualquer requisição.

É possível ter diversos *searchComponents* e *requestHandlers* declarados no arquivo de configuração, desde que eles sejam identificados por nomes diferentes. A combinação das configurações é o que determinará a robustez do sistema de recuperação de informação.

### Arquivo *schema.xml*

A configuração do arquivo *schema.xml* se deu da seguinte maneira:

Para os campos dos documentos, foram definidos os seguintes:

```
<field name="ano" type="text_general" indexed="true" stored="true"/>
<field name="issn" type="text_general" indexed="true" stored="true"/>
<field name="titulo_serie" type="text_general" indexed="true" stored="true"/>
<field name="editora" type="text_general" indexed="true" stored="true"/>
<field name="autor" type="text_general" indexed="true" stored="true" />
<field name="titulo" type="text_general" indexed="true" stored="true" multiValued="true"/>
<field name="url" type="text_general" indexed="true" stored="true" multiValued="true"/>
<field name="DATE" type="text_general" indexed="true" stored="true"/>
<field name="TOPICS" type="text_general" indexed="true" stored="true" multiValued="true"/>
<field name="PLACES" type="text_general" indexed="true" stored="true" multiValued="true"/>
<field name="ORGS" type="text_general" indexed="true" stored="true" multiValued="true"/>
<field name="PEOPLE" type="text_general" indexed="true" stored="true" multiValued="true"/>
<field name="EXCHANGES" type="text_general" indexed="true" stored="true"
multiValued="true"/>

<field name="content" type="text_general" indexed="false" stored="true" multiValued="true"/>
```

Todos os campos definidos anteriormente, exceto o campo *url*, são copiados para um novo campo chamado *text*. Tal campo possui a informação de todos os outros juntos e é utilizado em diversos recursos da busca, como *spellcheck* e sugestão de completar palavras, por exemplo:

```
<copyField source="ano" dest="text"/>
<copyField source="issn" dest="text"/>
```

```
<copyField source="content" dest="text"/>
<copyField source="titulo_serie" dest="text"/>
<copyField source="editora" dest="text"/>
<copyField source="autor" dest="text"/>
<copyField source="titulo" dest="text"/>
<copyField source="DATE" dest="text"/>
<copyField source="TOPICS" dest="text"/>
<copyField source="PLACES" dest="text"/>
<copyField source="ORGS" dest="text"/>
<copyField source="PEOPLE" dest="text"/>
<copyField source="EXCHANGES" dest="text"/>
```

As demais configurações permaneceram inalteradas e refletem o padrão.

### Arquivo solrconfig.xml

A configuração do arquivo solrconfig.xml se deu da seguinte maneira:

#### Tag requestHandler

O arquivo de configuração solrconfig.xml possui uma *tag* requestHandler personalizada para o projeto. A mesma foi nomeada handler\_projeto:

```
<requestHandler name="/handler_projeto" class="solr.SearchHandler">
```

Dentro da *tag requesthandler* estão duas *tags* principais: a `<lst name="defaults">` e `<arr name="last-components">`. A primeira delas define várias configurações a serem utilizadas na busca. A segunda referencia componentes do tipo *searchComponent* que serão utilizados na busca.

#### Indentação e score

Para indentação do resultado da busca. No caso, pedimos para que a resposta seja devolvida no formato *json* indentado, já que a maioria dos serviços de webservice disponíveis também utilizam esse padrão:

```
<str name="wt">json</str>
<str name="indent">>true</str>
<str name="df">text</str>
```

Optou-se por retornar o *score* de cada documento. Tal valor pode ser relevante para casos de depuração, por exemplo:

```
<str name="fl">*,score</str>
```

## Spellcheck

Para o *spellcheck* utilizamos as seguintes tags:

```
<str name="spellcheck">on</str>
<str name="spellcheck.extendedResults">false</str>
<str name="spellcheck.count">1</str>
<str name="spellcheck.alternativeTermCount">1</str>
<str name="spellcheck.maxResultsForSuggest">15</str>
<str name="spellcheck.collate">true</str>
<str name="spellcheck.collateExtendedResults">false</str>
<str name="spellcheck.maxCollationTries">5</str>
<str name="spellcheck.maxCollations">3</str>
```

A princípio podemos resumir a configuração do *spellcheck* da seguinte maneira: as sugestões são geradas somente quando há menos de 15 documentos retornados para uma dada consulta. É pedido somente mais uma sugestão de consulta com os termos corrigidos. Não é requerido explicações a respeito das correções, já que grande parte dos usuários simplesmente ignora as explicações e aceita a correção.

`<str name="spellcheck">on</str>` é utilizado para habilitar o recurso.

`<str name="spellcheck.extendedResults">false</str>` é utilizado para ignorar informações sobre as sugestões, como por exemplo a frequência no índice.

`<str name="spellcheck.count">1</str>` é utilizado para limitar o número de sugestões em somente um.

`<str name="spellcheck.alternativeTermCount">1</str>` é utilizado para limitar o número de sugestão de cada termo em um.

`<str name="spellcheck.maxResultsForSuggest">15</str>` é utilizado para que as sugestões sejam geradas somente quando uma consulta retorne menos que 15 documentos.

`<str name="spellcheck.collate">true</str>` é utilizado para que o *Solr* gere, a partir da melhor sugestão (no caso pedimos somente uma sugestão) uma nova consulta com as devidas correções.

`<str name="spellcheck.collateExtendedResults">false</str>` é utilizado para ignorar qualquer explicação adicional que o *Solr* venha dar para as sugestões de correção.

`<str name="spellcheck.maxCollationTries">5</str>` é utilizado para limitar em cinco o número de tentativas que o *Solr* deve fazer para gerar uma nova consulta com os termos corrigidos. Após essas cinco tentativas, não há uma sugestão de nova consulta com os termos corrigidos.

`<str name="spellcheck.maxCollations">1</str>` é utilizado para que somente uma nova sugestão de consulta seja gerada.

Uma segunda parte é necessária para o funcionamento do *spellcheck*. Ela é definida pela tag *searchComponent*. Tal componente foi nomeado como `spellcheck_projeto`:

```

<searchComponent name="spellcheck_projeto" class="solr.SpellCheckComponent">
  <str name="queryAnalyzerFieldType">text_general</str>
  <lst name="spellchecker">
    <str name="name">default</str>
    <str name="field">text</str>
    <str name="classname">solr.DirectSolrSpellChecker</str>
    <str name="distanceMeasure">internal</str>
    <float name="accuracy">0.5</float>
    <int name="maxEdits">2</int>
    <int name="minPrefix">1</int>
    <int name="maxInspections">5</int>
    <int name="minQueryLength">4</int>
    <float name="maxQueryFrequency">0.01</float>
  </lst>
</searchComponent>

```

As *tags* anteriores configuram o *spellcheck* de maneira que o campo *text* seja levado em consideração para a correção. Além disso, é utilizada a distância de palavras padrão: de Levenhstein:

```

<str name="field">text</str>
<str name="distanceMeasure">internal</str>

```

Por fim, existe também uma configuração do *spellcheck* leva em consideração palavras definidas em um arquivo externo. No caso, tal funcionalidade não é implementada. Outra funcionalidade do *spellcheck* que não é utilizada é o *wordbreak*. *Wordbreak* é capaz de detectar que palavras que são escritas separadamente estão digitadas incorretamente quando juntas. Tal funcionalidade foi desabilitada por atrapalhar o *spellcheck* convencional por distância de Levenhstein.

Exemplo para teste: uma consulta foi gerada com erros propositais: *dairr*. Após o *spellcheck*, o *Solr* sugeriu uma nova consulta: *dairy*.

Por fim, podemos verificar que a funcionalidade de correção está funcionando de maneira satisfatória.

## Suggester

A funcionalidade *suggester* vem a complementar o *spellcheck*. Quando um usuário realiza uma consulta mas não digita um termo por completo, o *suggester* sugere um termo completo.

```
<str name="suggest">true</str>
<str name="suggest.count">1</str>
```

Para o projeto, o *suggester* é habilitado e configurado para retornar uma sugestão de termo caso o mesmo não seja digitado por completo.

Para a configuração completa foi preciso criar mais um `searchComponent` que foi chamado de `suggest_projeto`:

```
<searchComponent name="suggest_projeto" class="solr.SuggestComponent">
  <lst name="suggester">
    <str name="name">default</str>
    <str name="lookupImpl">FuzzyLookupFactory</str>
    <str name="dictionaryImpl">HighFrequencyDictionaryFactory</str>
    <str name="field">text</str>
    <str name="suggestAnalyzerFieldType">text_general</str>
    <str name="buildOnStartup">true</str>
  </lst>
</searchComponent>
```

O componente atua no campo `text` e é construído sempre que o *Solr* é iniciado.

```
<str name="field">text</str>
<str name="suggestAnalyzerFieldType">text_general</str>
```

Os dicionários e a implementação são configurados por

```
<str name="lookupImpl">FuzzyLookupFactory</str>
<str name="dictionaryImpl">HighFrequencyDictionaryFactory</str>
```

Exemplo para teste: foi gerada a consulta com o termo `prof`. O *Solr* retornou uma sugestão como `profissional`. A mesma sugestão, por sua vez, possuía um peso bastante grande, indicando que a confiança na sugestão era grande.

## Snippets

*Snippets* são importantes para gerar contexto ao usuário. *Snippets* são trechos de documentos que contém os termos buscados, dando, assim, uma explicação do porquê um documento foi retornado pela busca. Para a configuração de *snippets* no projeto foram utilizadas as seguintes *tags*: `<str name="hl">on</str>`

```
<str name="hl.fl">titulo autor content TOPICS PEOPLE</str>
```

```
<str name="hl.preserveMulti">true</str>

<str name="hl.encoder">html</str>

<str name="hl.simple.pre">&lt;b&gt;</str>

<str name="hl.simple.post">&lt;/b&gt;</str>

<str name="f.titulo.hl.fragsize">0</str>

<str name="f.autor.hl.fragsize">0</str>

<str name="f.autor.hl.alternateField">autor</str>

<str name="f.content.hl.fragsize">50</str>

<str name="f.content.hl.alternateField">content</str>

<str name="f.TOPICS.hl.fragsize">0</str>

<str name="f.TOPICS.hl.alternateField">TOPICS</str>

<str name="f.PEOPLE.hl.fragsize">0</str>

<str name="f.PEOPLE.hl.alternateField">PEOPLE</str>
```

`<str name="hl">on</str>` é utilizado para habilitar a geração de *snippets*.

`<str name="hl.fl">titulo autor content TOPICS PEOPLE</str>` é utilizado para configurar qual campo dos documentos (*fields* - zonas) devem ser utilizados para a geração de *snippets*. No caso, os campos titulo, autor, content, TOPICS e PEOPLE são os utilizados.

`<str name="hl.preserveMulti">true</str>` é utilizado para que, em campos do tipo multivalorados, a ordenação dos termos seja mantida.

`<str name="hl.encoder">html</str>`, `<str name="hl.simple.pre">&lt;b&gt;</str>` e `<str name="hl.simple.post">&lt;/b&gt;</str>` são utilizados para que, os *snippets* sejam formatados em *html* com negrito.

`<str name="f.titulo.hl.fragsize">0</str>` é utilizado para que o *snippet* gerado a partir do campo titulo seja mostrado inteiro. Valores diferentes de 0 indicam o tamanho da janela do *snippet*. Como títulos geralmente são curtos, não há razão para criar janelas, mas simplesmente mostrar o título inteiro.

`<str name="f.autor.hl.alternateField">autor</str>` é utilizado como campo alternativo para a geração de *snippet*. Caso não haja ocorrências no campo titulo, então são buscadas ocorrências no campo autor.

De maneira similar, os campos alternativos são configurados para content, TOPICS e PEOPLE. É interessante notar que para o campo content, por geralmente conter textos mais longos, a janela possui um comprimento de até 50 caracteres.

Exemplo para teste: uma consulta foi gerada com os termos profissional. Ao final da resposta havia uma *tag* no *json* chamada highlighting. Nessa *tag* estavam listadas todas as janelas (*snippets*) geradas e a qual documento elas se referiam. No caso, a palavra *professional* estava cercada por `<b>` e `</b>`, como requerido na configuração. Havia somente um *snippet* para cada campo, como solicitado no configuração.

## Facets

*Facets*, na prática, retornam um sumário do resultado da busca, enumerando ou contando quais valores existentes nos documentos estão presentes nos documentos retornados.

```
<str name="facet">on</str>
<str name="facet.missing">false</str>
<str name="facet.field">ano</str>
<str name="facet.field">TOPICS</str>
<str name="facet.mincount">1</str>
```

No caso do projeto, os *facets* foram habilitados e colocados no campo ano e TOPICS. Assim, o resultado da busca informa quais são anos e os tópicos dos documentos que foram retornados na consulta, além de fornecer uma contagem para cada ano e tópico. Pela configuração, é necessário que pelo menos um documento contendo o ano seja retornado pela consulta para que o ano apareça na listagem com o contador. Além disso, documentos que não possuem um ano/tópico atribuído não são levados em consideração.

Tal funcionalidade é bastante importante principalmente no momento de oferecer filtros para o usuário pós consultas. Não é conveniente, por exemplo, após uma consulta retornar centenas de livros, oferecer um filtro de livros para o ano 1990 quando não há livros que se encaixam nesse filtro.

## Boost

Como visto na disciplina, alguns campos possuem mais relevância do que outros. No caso do projeto, o título do livro é mais importante, por exemplo, do que a *url*. Nesse caso, é preciso definir peso aos campos dos documentos.

```
<str name="defType">edismax</str>
<str name="qf">
    titulo^150.0 TOPICS^50.0 autor^5.0 PEOPLE^3.0 editora^0.5 url^0.1 text^0.5 ano^0.5
    issn^0.5 titulo_serie^0.5
</str>
```

Para o projeto foi definido que o título de um livro é o campo de maior relevância, seguido do valor no campo TOPICS, autor e PEOPLE. O campo url, por sua vez, possui uma relevância ínfima em relação aos demais, como é possível observar.

## Stopwords e sinônimos

Para a configuração de *stopwords* e termos sinônimos (um dicionário *thesaurus*) foi utilizada a configuração padrão. Na prática, os arquivos contendo *stopwords* em diversos idiomas e o arquivo de sinônimos localizados em /conf não foram alterados. A lista padrão de *stopwords* em diversos idiomas já é bastante completa, inclusive para o idioma português.





## Testes realizados

A seguir são listados alguns testes realizados e os respectivos retornos. Cada teste possui um objetivo, o qual é explicitado logo no início. Ao final, há uma conclusão geral para todos os testes.

### Teste 1

Objetivo: consulta casual para verificação geral dos parâmetros

Consulta: retrieval

Documentos retornados:

```
"docs": [
  {
    "id": "9781843347224",
    "titulo": [
      "Multimedia Information Retrieval"
    ],
    "ano": "2013",
    "titulo_serie": "Chandos Information Professional Series",
    "editora": "Chandos Publishing",
    "url": [
      "http://www.sciencedirect.com/science/book/9781843347224"
    ],
    "_version_": 1519109898000924679,
    "score": 4.5963683
  },
  {
    "id": "9780128033883",
    "titulo": [
      "Satellite Soil Moisture Retrieval"
    ],
    "ano": "2016",
    "_version_": 1519109898087956489,
    "score": 4.5963683
  },
  {
    "id": "9780128024195",
    "titulo": [
      "View-based 3-D Object Retrieval"
    ],
    "ano": "2015",
    "titulo_serie": "CSRT/Computer Science Reviews and Trends",
    "editora": "Morgan Kaufmann",
    "url": [
      "http://www.sciencedirect.com/science/book/9780128024195"
    ],
    "_version_": 1519109898125705251,
    "score": 3.447276
  },
  {
    "id": "9781843340775",
    "titulo": [
      "Effective Information Retrieval from the Internet"
    ],
    "ano": "2015",
    "titulo_serie": "Information Science Reference",
    "editora": "Elsevier",
    "url": [
      "http://www.sciencedirect.com/science/book/9781843340775"
    ],
    "_version_": 1519109898125705251,
    "score": 3.447276
  }
]
```

```

"ano": "2004",
"titulo_serie": "Chandos Information Professional Series",
"editora": "Chandos Publishing",
"autor": "Stacey, Alison",
"url": [
  "http://www.sciencedirect.com/science/book/9781843340775"
],
"_version_": 1519109898464395295,
"score": 3.447276
},
{
  "id": "9780123693877",
  "titulo": [
    "A Unified Framework for Video Summarization, Browsing and Retrieval"
  ],
  "ano": "2006",
  "editora": "Academic Press",
  "autor": "Xiong, Ziyu",
  "url": [
    "http://www.sciencedirect.com/science/book/9780123693877"
  ],
  "_version_": 1519109898330177566,
  "score": 2.8727303
},
{
  "id": "9780121680305",
  "titulo": [
    "Symbolic Projection for Image Information Retrieval and Spatial Reasoning"
  ],
  "ano": "1996",
  "titulo_serie": "Signal Processing and its Applications",
  "editora": "Academic Press",
  "autor": "Chang, Shi-Kuo",
  "url": [
    "http://www.sciencedirect.com/science/book/9780121680305"
  ],
  "_version_": 1519109898334371873,
  "score": 2.8727303
},
{
  "id": "9781843340799",
  "titulo": [
    "Mastering Information Retrieval and Probabilistic Decision Intelligence Technology"
  ],
  "ano": "2004",
  "titulo_serie": "Chandos Information Professional Series",
  "editora": "Chandos Publishing",
  "autor": "Brown, Daniel",
  "url": [
    "http://www.sciencedirect.com/science/book/9781843340799"
  ],
  "_version_": 1519109898465443849,
  "score": 2.8727303
},
{

```

```

"DATE": "2-MAR-1987 14:55:10.03",
"TOPICS": [
  "acq"
],
"titulo": [
  "ROSPATCH CORP REJECTS OFFER FROM DIAGNOSTIC RETRIEVAL SYSTEMS INC"
],
"id": "reuters-00708",
"_version_": 1519391250582601729,
"score": 2.8727303
},
{
  "DATE": "2-MAR-1987 09:25:42.34",
  "TOPICS": [
    "acq"
  ],
  "titulo": [
    "DIAGNOSTIC/RETRIEVAL SYSTEMS INC MAKES 53 MLN DLR BID FOR ROSPATCH CORP"
  ],
  "id": "reuters-00399",
  "_version_": 1519391250561630218,
  "score": 2.2981842
}
]

```

#### **Facets retornados:**

```

"facet_fields": {
  "ano": [
    "2004",
    2,
    "1996",
    1,
    "2006",
    1,
    "2013",
    1,
    "2015",
    1,
    "2016",
    1
  ],
  "TOPICS": [
    "acq",
    2
  ]
}

```

#### **Snippets retornados:**

```

"highlighting": {
  "9781843347224": {
    "titulo": [

```

```

    "Multimedia Information <b>Retrieval</b>"
  ],
},
"9780128033883": {
  "titulo": [
    "Satellite Soil Moisture <b>Retrieval</b>"
  ],
},
"9780128024195": {
  "titulo": [
    "View-based 3-D Object <b>Retrieval</b>"
  ],
},
"9781843340775": {
  "titulo": [
    "Effective Information <b>Retrieval</b> from the Internet"
  ],
  "autor": [
    "Stacey, Alison"
  ],
},
"9780123693877": {
  "titulo": [
    "A Unified Framework for Video Summarization, Browsing and <b>Retrieval</b>"
  ],
  "autor": [
    "Xiong, Ziyu"
  ],
},
"9780121680305": {
  "titulo": [
    "Symbolic Projection for Image Information <b>Retrieval</b> and Spatial Reasoning"
  ],
  "autor": [
    "Chang, Shi-Kuo"
  ],
},
"9781843340799": {
  "titulo": [
    "Mastering Information <b>Retrieval</b> and Probabilistic Decision Intelligence Technology"
  ],
  "autor": [
    "Brown, Daniel"
  ],
},
"reuters-00708": {
  "titulo": [
    "ROSPATCH CORP REJECTS OFFER FROM DIAGNOSTIC <b>RETRIEVAL</b> SYSTEMS INC"
  ],
  "TOPICS": [
    "acq"
  ],
},
"reuters-00399": {
  "titulo": [

```

```

"DIAGNOSTIC&#x2F;<b>RETRIEVAL</b> SYSTEMS INC MAKES 53 MLN DLR BID FOR ROSPATCH
CORP"
],
"TOPICS": [
  "acq"
]
}
}

```

Conclusão do teste:

O *Solr* retornou documentos de maneira satisfatória. Todos os documentos retornados possuíam o termo *retrieval* no título, que é o campo com maior relevância. Além disso, pode-se observar que os *facets* e *snippets* estão sendo gerados corretamente.

## Teste 2

Objetivo: verificar a geração de *snippets* para os campo alternativos.

Consulta: universal earn

Documentos retornados:

```

"docs": [
  {
    "DATE": "8-APR-1987 10:40:25.85",
    "TOPICS": [
      "earn"
    ],
    "PLACES": [
      "usa"
    ],
    "titulo": [
      "UNIVERSAL MEDICAL UMBIZ> DISTRIBUTION SET"
    ],
    "content": [
      "Qtly distribution 7-1/2 cts vs 7-1/2 cts prior (excludingWn2-1/2 cts special)Wn Pay April 30Wn
Record April 22Wn NOTE: Full name is Universal Medical Buildings L.P.Wn Reuter"
    ],
    "id": "reuters-15031",
    "_version_": 1519391250899271689,
    "score": 2.2812364
  },
  {
    "DATE": "26-MAR-1987 14:24:20.75",
    "TOPICS": [
      "earn"
    ],
    "PLACES": [
      "usa"
    ],
    "titulo": [
      "UNIVERSAL HOLDING CORP UHCO> 4TH QTR LOSS"
    ],
  },

```

```

"content": [
  "Shr profit nil vs profit nine ctsWn Net profit 2,000 vs profit 195,000Wn Revs 2,623,000 vs 2,577,000Wn YearWn Shr loss 21 cts vs profit 13 ctsWn Net loss 425,000 vs profit 278,000Wn Revs 15.4 mln vs 8,637,000Wn Note: Net includes capital gains of 63,000 vs 211,000 forWnqtr and 304,000 vs 292,000 for year. Current year net includesWncharge of 716,000 from contract obligation to former chairman.Wn Reuter"
],
"id": "reuters-10088",
"_version_": 1519391249167024133,
"score": 1.9554304
},
{
  "DATE": "27-MAR-1987 10:52:27.96F",
  "TOPICS": [
    "earn"
  ],
  "PLACES": [
    "usa"
  ],
  "titulo": [
    "UNIVERSAL HOLDING CORP UHCO> 4TH QTR NET"
  ],
  "content": [
    "Shr NAWn Net profit 2,000 vs profit 195,000Wn Revs 2,623,000 vs 2,577,000Wn YearWn Shr NAWn Net loss 425,000 vs profit 278,000Wn Revs 15.4 mln vs 8,637,000Wn Reuter"
  ],
  "id": "reuters-10458",
  "_version_": 1519391249224695814,
  "score": 1.9554304
},
{
  "DATE": "19-OCT-1987 12:09:52.75",
  "TOPICS": [
    "earn"
  ],
  "PLACES": [
    "usa"
  ],
  "titulo": [
    "UNIVERSAL FURNITURE LTD UFURF.O> 3RD QTR NET"
  ],
  "content": [
    "Shr 34 cts vs 26 ctsWn Net 6,150,000 vs 4,743,000Wn Revs 61.4 mln vs 49.5 mlnWn Nine monthsWn Shr 89 cts vs 70 ctsWn Net 16 mln vs 11.8 mlnWn Revs 170 mln vs 137.5 mlnWn NOTE: All share and per share data have been adjusted toWnreflect 100 pct stock dividend distriton on April 24, 1987 andWnthe public offier of two mln shares ofthe company on June 4,Wn1986.Wn Reuter"
  ],
  "id": "reuters-21264",
  "_version_": 1519391250495569928,
  "score": 1.9554304
},
{
  "DATE": "9-APR-1987 12:17:19.72",
  "TOPICS": [
    "earn"
  ]
}

```

```

    ],
    "PLACES": [
        "usa"
    ],
    ],
    "titulo": [
        "UNIVERSAL FOODS CORP UFC> VOTES DIVIDEND"
    ],
    ],
    "content": [
        "Qtly div 20 cts vs 20 cts prior qtrWn    Pay 6 MayWn    Record 21 AprilWn Reuter"
    ],
    ],
    "id": "reuters-15706",
    "_version_": 1519391250941214740,
    "score": 1.9554304
},
{
    "DATE": "9-APR-1987 15:34:29.64",
    "TOPICS": [
        "earn"
    ],
    ],
    "PLACES": [
        "usa"
    ],
    ],
    "titulo": [
        "UNIVERSAL HEALTH REALTY UHT> 1ST QTR NET"
    ],
    ],
    "content": [
        "Shr 26 cts vs nilWn    Net 2,244,000 vs nilWn    Rev 3.4 mln vs nilWn    NOTE: Company's full
name is Universal Health Realty IncomeWnTrust. Quarter is company's first full quarter of earnings.Wn
Reuter"
    ],
    ],
    "id": "reuters-15895",
    "_version_": 1519391250952749069,
    "score": 1.9554304
},
{
    "id": "9781843346333",
    "titulo": [
        "Universal Design"
    ],
    ],
    "ano": "2012",
    "titulo_serie": "Chandos Information Professional Series",
    "editora": "Chandos Publishing",
    "url": [
        "http://www.sciencedirect.com/science/book/9781843346333"
    ],
    ],
    "_version_": 1519109897969467420,
    "score": 1.62903
},
{
    "id": "9780444515865",
    "titulo": [
        "Universal Spaces and Mappings"
    ],
    ],
    "ano": "2005",
    "issn": "0304-0208",

```



```

"titulo_serie": "North-Holland Mathematics Studies",
"editora": "North Holland",
"autor": "Iliadis, S.D.",
"url": [
  "http://www.sciencedirect.com/science/book/9780444515865"
],
"_version_": 1519109898415112204,
"score": 1.303224
},
{
  "DATE": "25-MAR-1987 09:07:56.19",
  "TOPICS": [
    "acq"
  ],
  "PLACES": [
    "canada"
  ],
  "titulo": [
    "TIMMINCO ACQUIRES UNIVERSAL ADHESIVES"
  ],
  "content": [
    "Timminco Ltd> said it acquiredWnUniversal Adhesives Inc, of Memphis, for undisclosed terms,
inWna move to expand Timminco's operations into the United States.Wn The company said Universal
Adhesives, with five U.S.Wnplants, has annual sales of 12 mln U.S. dlrs, which will doubleWnTimminco's
presence in the North American adhesives market.Wn Timminco said Universal Adhesives will
complement theWncompany's Canadian-based industrial adhesives division and is aWnkey step in its
long-term goal for expansion in the specialtyWnchemical field.Wn Reuter"
  ],
  "id": "reuters-09258",
  "_version_": 1519391248707747843,
  "score": 1.303224
},
{
  "DATE": "29-JUN-1987 13:18:08.98",
  "TOPICS": [
    "acq"
  ],
  "PLACES": [
    "usa"
  ],
  "titulo": [
    "UNIVERSAL COMMUNICATION UCS> TO SELL ASSETS"
  ],
  "content": [
    "Universal Communication Systems IncWnsaid it has tentatively agreed to sell substantially all
itsWnassets for about 79 mln dlrs in cash and notes plus limitedWnprofit participation.Wn The company
said the terms of the sale have been approvedWnby its board and by Prime Motor Inns Inc PDQ>, owner of
aboutWn84 pct of Universal's outstanding stock.Wn It described the purchaser as a subsidiary of a
company inWnthe communications field which is one of the 100 largest U.S.Wncorporations.Wn The
company said the transaction involves the payment of 20Wnmln dlrs in cash, a non-interest bearing
payment of 11.3 mlnWndlrs in four equal instalments over four years and twoWnpromissory notes
guaranteed by an affiliate of the purchaser.Wn It said a 31.5 mln dlr 14 pct note is payable in four
equalWninstalments over four years. It said a 16.3 mln dlr 8.5 pctWnnote due Dec 31, 1992, includes
participation in the 1992Wnprofits of the acquiring company. Universdal said the profitWnelement can be

```

terminated with payments by the purchaser of neither five mln dlrs in 1988, six mln dlrs in 1989, seven mln dlrs in 1990 or eight mln dlrs in 1991. Wn Reuter"

```
    ],
    "id": "reuters-19766",
    "_version_": 1519391248425680897,
    "score": 0.97741795
  },
  {
    "DATE": "20-MAR-1987 10:03:51.02",
    "TOPICS": [
      "acq"
    ],
    "PLACES": [
      "usa"
    ],
    "titulo": [
      "UNIVERSAL RESOURCES UVR> TO VOTE ON MERGER"
    ],
    "content": [
```

"Universal Resources Corp said it Wnis holding a special shareholders meeting this morning to vote Wnon the previously-proposed merger between it and Questar Corp WnSTR>. Wn Universal, whose stock was delayed this morning on the WnAmerican Stock Exchange, said it will release a statement later Wnin the day on the vote. Wn Reuter"

```
    ],
    "id": "reuters-07669",
    "_version_": 1519391250096062465,
    "score": 0.97741795
  },
  {
    "DATE": "20-MAR-1987 13:15:12.22",
    "TOPICS": [
      "acq"
    ],
    "PLACES": [
      "usa"
    ],
    "titulo": [
      "UNIVERSAL RESOURCES UVR>HOLDERS APPROVE MERGER"
    ],
    "content": [
```

"Universal Resources Corp said its Wnshareholders approved the merger of the company with Questar WnCorp STR>. Wn Separately, Universal said it will redeem its 15.75 pct Wndebentures due December 15, 1996, on April 19, at 104.8 pct of Wnface amount plus accrued interest. Wn Universal said it will operate as a wholly owned unit of WnQuestar under its current name. Under terms of the merger, Wnwhich took effective today, Universal said its shareholders Wnwill receive three dlrs a share in cash. Wn It said its stock will no longer trade on the AMEX. Wn Reuter"

```
    ],
    "id": "reuters-07846",
    "_version_": 1519391250114936839,
    "score": 0.97741795
  },
  {
    "DATE": "18-MAR-1987 11:31:42.35",
    "PLACES": [
      "usa"
    ],

```

```

    ],
    "EXCHANGES": [
        "nasdaq"
    ],
    "titulo": [
        "UNIVERSAL MEDICAL UMBIZ> JOINS NASDAQ SYSTEM"
    ],
    "content": [
        "Universal Medical Buildings LPWnsaid its partnership units were admitted for trading on
theWnNASDAQ National Market System.Wn Reuter"
    ],
    "id": "reuters-06554",
    "_version_": 1519391250209308672,
    "score": 0.97741795
},
{
    "DATE": "2-MAR-1987 12:53:03.20",
    "titulo": [
        "UNIVERSAL HEALTH REALTY UHT> IN INITIAL PAYOUT"
    ],
    "content": [
        "Universal Health RealtyWnIncome Trust, which recently went public, said its board hasWndeclared
an initial quarterly dividednd of 33 cts per share,Wnpayable MArch 31 to holders of record March 16.Wn
Reuter"
    ],
    "id": "reuters-00577",
    "_version_": 1519391250574213121,
    "score": 0.97741795
},
{
    "DATE": "2-APR-1987 14:25:14.89",
    "TOPICS": [
        "acq"
    ],
    "PLACES": [
        "usa"
    ],
    "titulo": [
        "ACME STEEL ACME> TO ACQUIRE UNIVERSAL TOOL"
    ],
    "content": [
        "Acme Steel Co said it has agreedWnin principle to acquire Universal Tool and Stamping Inc>
forWnan undisclosed price.Wn The agreement is subject to approval by Acme's board andWnexecution
of a definitive purchase agreement, the companiesWnsaid.Wn Universal Tool makes and distributes
automotive jacks andWnrelated equipment for the domestic auto industry.Wn Reuter"
    ],
    "id": "reuters-12678",
    "_version_": 1519391250861522951,
    "score": 0.97741795
}
]

```

*Snippets retornados:*

```

"highlighting": {
  "reuters-15031": {
    "titulo": [
      "<b>UNIVERSAL</b> MEDICAL UMBIZ&gt; DISTRIBUTION SET"
    ],
    "content": [
      " April 22Wn  NOTE: Full name is <b>Universal</b> Medical"
    ],
    "TOPICS": [
      "<b>earn</b>"
    ]
  },
  "reuters-10088": {
    "titulo": [
      "<b>UNIVERSAL</b> HOLDING CORP UHCO&gt; 4TH QTR LOSS"
    ],
    "content": [
      "Shr profit nil vs profit nine ctsWn  Net profit"
    ],
    "TOPICS": [
      "<b>earn</b>"
    ]
  },
  "reuters-10458": {
    "titulo": [
      "<b>UNIVERSAL</b> HOLDING CORP UHCO&gt; 4TH QTR NET"
    ],
    "content": [
      "Shr NAWn  Net profit 2,000 vs profit 195,000"
    ],
    "TOPICS": [
      "<b>earn</b>"
    ]
  },
  "reuters-21264": {
    "titulo": [
      "<b>UNIVERSAL</b> FURNITURE LTD UFURF.O&gt; 3RD QTR NET"
    ],
    "content": [
      "Shr 34 cts vs 26 ctsWn  Net 6,150,000 vs"
    ],
    "TOPICS": [
      "<b>earn</b>"
    ]
  },
  "reuters-15706": {
    "titulo": [
      "<b>UNIVERSAL</b> FOODS CORP UFC&gt; VOTES DIVIDEND"
    ],
    "content": [
      "Qtly div 20 cts vs 20 cts prior qtrWn  Pay 6 May"
    ],
    "TOPICS": [
      "<b>earn</b>"
    ]
  }
}

```

```

},
"reuters-15895": {
  "titulo": [
    "<b>UNIVERSAL</b> HEALTH REALTY UHT&gt; 1ST QTR NET"
  ],
  "content": [
    "<b>Universal</b> Health Realty IncomeWnTrust. Quarter is company&#x27;s"
  ],
  "TOPICS": [
    "<b>earn</b>"
  ]
},
"9781843346333": {
  "titulo": [
    "<b>Universal</b> Design"
  ]
},
"9780444515865": {
  "titulo": [
    "<b>Universal</b> Spaces and Mappings"
  ],
  "autor": [
    "Iliadis, S.D."
  ]
},
"reuters-09258": {
  "titulo": [
    "TIMMINCO ACQUIRES <b>UNIVERSAL</b> ADHESIVES"
  ],
  "content": [
    "Timminco Ltd&gt; said it acquiredWn<b>Universal</b>"
  ],
  "TOPICS": [
    "acq"
  ]
},
"reuters-19766": {
  "titulo": [
    "<b>UNIVERSAL</b> COMMUNICATION UCS&gt; TO SELL ASSETS"
  ],
  "content": [
    "<b>Universal</b> Communication Systems IncWnsaid it has"
  ],
  "TOPICS": [
    "acq"
  ]
},
"reuters-07669": {
  "titulo": [
    "<b>UNIVERSAL</b> RESOURCES UVR&gt; TO VOTE ON MERGER"
  ],
  "content": [
    "<b>Universal</b> Resources Corp said itWnis holding a"
  ],
  "TOPICS": [

```

```

    "acq"
  ],
},
"reuters-07846": {
  "titulo": [
    "<b>UNIVERSAL</b> RESOURCES UVR&gt;HOLDERS APPROVE MERGER"
  ],
  "content": [
    "<b>Universal</b> Resources Corp said itsWnshareholders"
  ],
  "TOPICS": [
    "acq"
  ]
},
"reuters-06554": {
  "titulo": [
    "<b>UNIVERSAL</b> MEDICAL UMBIZ&gt; JOINS NASDAQ SYSTEM"
  ],
  "content": [
    "UUniversal Medical Buildings LPWnsaid its"
  ]
},
"reuters-00577": {
  "titulo": [
    "<b>UNIVERSAL</b> HEALTH REALTY UHT&gt; IN INITIAL PAYOUT"
  ],
  "content": [
    "<b>Universal</b> Health RealtyWnIncome Trust, which"
  ]
},
"reuters-12678": {
  "titulo": [
    "ACME STEEL ACME&gt; TO ACQUIRE <b>UNIVERSAL</b> TOOL"
  ],
  "content": [
    " acquire <b>Universal</b> Tool and Stamping Inc&gt; forWnan"
  ],
  "TOPICS": [
    "acq"
  ]
}
}
}

```

Conclusão do teste:

A geração de *snippets* para os campos alternativo está funcionando de maneira satisfatória. Como pode-se verificar, foram gerados *snippets* para os campos content e TOPICS, por exemplo.

### Teste 3

Objetivo: verificar o funcionamento do *spellcheck*.

Consulta: universiti collegi

Documentos retornados: nenhum documento retornado.

*Spellcheck* retornado:

```
"spellcheck": {
  "suggestions": [
    "universiti",
    {
      "numFound": 1,
      "startOffset": 0,
      "endOffset": 10,
      "suggestion": [
        "university"
      ]
    },
    "collegi",
    {
      "numFound": 1,
      "startOffset": 11,
      "endOffset": 18,
      "suggestion": [
        "college"
      ]
    }
  ],
  "collations": [
    "collation",
    "university college"
  ]
}
```

Conclusão do teste:

O *spellcheck* está funcionando de maneira satisfatória. O *Solr* foi capaz de corrigir o erro tanto no termo *universiti* como no termo *collegi*. A sugestão de consulta final (*collation*) foi *university college*, que é a consulta correta.

#### Teste 4

Objetivo: verificar a funcionalidade do *suggest*, que é capaz de corrigir termos digitados de maneira incompleta.

Consulta: universi

Documentos retornados: nenhum documento foi retornado.

*Suggest* retornado:

```
"suggest": {
  "default": {
    "univer": {
      "numFound": 1,
      "suggestions": [
```

```

{
  "term": "university",
  "weight": 25,
  "payload": ""
}
]
}
}
}

```

Conclusão do teste:

A sugestão de auto completar termos digitados de maneira incompleta está funcionando de maneira satisfatória. A sugestão de um termo completo para universi foi university, que é o termo correto.

### Teste 5

Objetivo: verificar o funcionamento do *boost* nos diversos campos dos documentos.

Consulta: professional

Documentos retornados com o *boosting*:

```

"docs": [
  {
    "id": "9780123850799",
    "titulo": [
      "Foundations of Professional Psychology"
    ],
    "ano": "2011",
    "editora": "Elsevier",
    "url": [
      "http://www.sciencedirect.com/science/book/9780123850799"
    ],
    "_version_": 1519109897919135754,
    "score": 3.7954252
  },
  {
    "id": "9781597494250",
    "titulo": [
      "Professional Penetration Testing"
    ],
    "ano": "2009",
    "editora": "Syngress",
    "autor": "Wilhelm, Thomas",
    "url": [
      "http://www.sciencedirect.com/science/book/9781597494250"
    ],
    "_version_": 1519109897932767242,
    "score": 3.7954252
  },
  {
    "id": "9781843346692",

```



```

"titulo": [
  "Academic and Professional Publishing"
],
"ano": "2012",
"titulo_serie": "Chandos Information Professional Series",
"editora": "Chandos Publishing",
"url": [
  "http://www.sciencedirect.com/science/book/9781843346692"
],
"_version_": 1519109897970515981,
"score": 3.7954252
},
{
  "id": "9781843347705",
  "titulo": [
    "The Emerging Information Professional"
  ],
  "ano": "2015",
  "_version_": 1519109898179182607,
  "score": 3.7954252
},
{
  "id": "9781843340812",
  "titulo": [
    "Continuing Professional Development"
  ],
  "ano": "2005",
  "titulo_serie": "Chandos Information Professional Series",
  "editora": "Chandos Publishing",
  "autor": "Brine, Alan",
  "url": [
    "http://www.sciencedirect.com/science/book/9781843340812"
  ],
  "_version_": 1519109898464395293,
  "score": 3.7954252
},
{
  "id": "9781843340874",
  "titulo": [
    "The New Information Professional"
  ],
  "ano": "2005",
  "titulo_serie": "Chandos Information Professional Series",
  "editora": "Chandos Publishing",
  "autor": "Myburgh, Sue",
  "url": [
    "http://www.sciencedirect.com/science/book/9781843340874"
  ],
  "_version_": 1519109898465443847,
  "score": 3.7954252
},
{
  "id": "9780750688451",
  "titulo": [
    "The Professional Qualifying Examinations"
  ],

```

```

],
"ano": "2004",
"editora": "Butterworth-Heinemann",
"autor": "Eperjesi",
"url": [
  "http://www.sciencedirect.com/science/book/9780750688451"
],
"_version_": 1519109898476978201,
"score": 3.7954252
},
{
  "id": "9781597499934",
  "titulo": [
    "Professional Penetration Testing (Second Edition)"
  ],
  "ano": "2013",
  "editora": "Syngress",
  "url": [
    "http://www.sciencedirect.com/science/book/9781597499934"
  ],
  "_version_": 1519109897989390344,
  "score": 3.320997
},
{
  "id": "9780128005675",
  "titulo": [
    "Professional Issues in Forensic Science"
  ],
  "ano": "2015",
  "issn": "2352-6238",
  "titulo_serie": "AFSS/Advanced Forensic Science Series",
  "_version_": 1519109898176036898,
  "score": 3.320997
},
{
  "id": "9780080450339",
  "titulo": [
    "Law & Ethics for the Eye Care Professional"
  ],
  "ano": "2008",
  "editora": "Butterworth-Heinemann",
  "autor": "Pierscioneck",
  "url": [
    "http://www.sciencedirect.com/science/book/9780080450339"
  ],
  "_version_": 1519109898069082144,
  "score": 2.8465688
},
{
  "id": "16_NO_ISBN_317",
  "titulo": [
    "The Academic Librarian as Blended Professional"
  ],
  "ano": "2016",
  "_version_": 1519109898095296543,

```

```

"score": 2.8465688
},
{
  "id": "9781843346494",
  "titulo": [
    "Service Science and the Information Professional"
  ],
  "ano": "2015",
  "editora": "Chandos Publishing",
  "url": [
    "http://www.sciencedirect.com/science/book/9781843346494"
  ],
  "_version_": 1519109898163453970,
  "score": 2.8465688
},
{
  "id": "9781928994800",
  "titulo": [
    "Configuring and Troubleshooting Windows XP Professional"
  ],
  "ano": "2002",
  "editora": "Syngress",
  "autor": "Syngress",
  "url": [
    "http://www.sciencedirect.com/science/book/9781928994800"
  ],
  "_version_": 1519109898296623115,
  "score": 2.8465688
},
{
  "id": "9781855730069",
  "titulo": [
    "Professional Diver's Manual on Wet-Welding"
  ],
  "ano": "1990",
  "editora": "Woodhead Publishing",
  "autor": "Keats, D",
  "url": [
    "http://www.sciencedirect.com/science/book/9781855730069"
  ],
  "_version_": 1519109898319691802,
  "score": 2.8465688
},
{
  "id": "9780443103803",
  "titulo": [
    "Palliative Care: A Practical Guide for the Health Professional"
  ],
  "ano": "2008",
  "editora": "Churchill Livingstone",
  "autor": "Boog",
  "url": [
    "http://www.sciencedirect.com/science/book/9780443103803"
  ],
  "_version_": 1519109898070130690,

```

```

    "score": 2.3721406
  }
]
}

```

#### Documentos retornados sem o *boosting*:

```

"docs": [
  {
    "id": "9781843347705",
    "titulo": [
      "The Emerging Information Professional"
    ],
    "ano": "2015",
    "_version_": 1519109898179182607,
    "score": 1.8939004
  },
  {
    "id": "9780123855466",
    "titulo": [
      "Petroleum Rock Mechanics"
    ],
    "ano": "2012",
    "editora": "Gulf Professional Publishing",
    "url": [
      "http://www.sciencedirect.com/science/book/9780123855466"
    ],
    "_version_": 1519109897915990025,
    "score": 1.6233432
  },
  {
    "id": "9780123850799",
    "titulo": [
      "Foundations of Professional Psychology"
    ],
    "ano": "2011",
    "editora": "Elsevier",
    "url": [
      "http://www.sciencedirect.com/science/book/9780123850799"
    ],
    "_version_": 1519109897919135754,
    "score": 1.6233432
  },
  {
    "id": "9780123838483",
    "titulo": [
      "Petrophysics (Third Edition)"
    ],
    "ano": "2011",
    "editora": "Gulf Professional Publishing",
    "url": [
      "http://www.sciencedirect.com/science/book/9780123838483"
    ],
    "_version_": 1519109897919135775,
    "score": 1.6233432
  }
]

```

```

},
{
  "id": "9781597494250",
  "titulo": [
    "Professional Penetration Testing"
  ],
  "ano": "2009",
  "editora": "Syngress",
  "autor": "Wilhelm, Thomas",
  "url": [
    "http://www.sciencedirect.com/science/book/9781597494250"
  ],
  "_version_": 1519109897932767242,
  "score": 1.6233432
},
{
  "id": "9780123854759",
  "titulo": [
    "Offshore Structures"
  ],
  "ano": "2012",
  "editora": "Gulf Professional Publishing",
  "url": [
    "http://www.sciencedirect.com/science/book/9780123854759"
  ],
  "_version_": 1519109897971564548,
  "score": 1.6233432
},
{
  "id": "9781597499934",
  "titulo": [
    "Professional Penetration Testing (Second Edition)"
  ],
  "ano": "2013",
  "editora": "Syngress",
  "url": [
    "http://www.sciencedirect.com/science/book/9781597499934"
  ],
  "_version_": 1519109897989390344,
  "score": 1.6233432
},
{
  "id": "9780123878250",
  "titulo": [
    "Pipeline Integrity Handbook"
  ],
  "ano": "2014",
  "editora": "Gulf Professional Publishing",
  "url": [
    "http://www.sciencedirect.com/science/book/9780123878250"
  ],
  "_version_": 1519109897990438919,
  "score": 1.6233432
},
{

```

```

    "id": "9780123865458",
    "titulo": [
      "Enhanced Oil Recovery"
    ],
    "ano": "2013",
    "editora": "Gulf Professional Publishing",
    "url": [
      "http://www.sciencedirect.com/science/book/9780123865458"
    ],
    "_version_": 1519109898006167571,
    "score": 1.6233432
  },
  {
    "id": "9780124165847",
    "titulo": [
      "Arctic Pipeline Planning"
    ],
    "ano": "2013",
    "editora": "Gulf Professional Publishing",
    "url": [
      "http://www.sciencedirect.com/science/book/9780124165847"
    ],
    "_version_": 1519109898009313313,
    "score": 1.6233432
  },
  {
    "id": "9781856176897",
    "titulo": [
      "Subsea Engineering Handbook"
    ],
    "ano": "2010",
    "editora": "Gulf Professional Publishing",
    "url": [
      "http://www.sciencedirect.com/science/book/9781856176897"
    ],
    "_version_": 1519109898038673415,
    "score": 1.6233432
  },
  {
    "id": "16_NO_ISBN_317",
    "titulo": [
      "The Academic Librarian as Blended Professional"
    ],
    "ano": "2016",
    "_version_": 1519109898095296543,
    "score": 1.6233432
  },
  {
    "id": "9780128003909",
    "titulo": [
      "Unconventional Gas Reservoirs"
    ],
    "ano": "2014",
    "editora": "Gulf Professional Publishing",
    "url": [

```

```

    "http://www.sciencedirect.com/science/book/9780128003909"
  ],
  "_version_": 1519109898101587972,
  "score": 1.6233432
},
{
  "id": "9780128012567",
  "titulo": [
    "Crude Oil Fouling"
  ],
  "ano": "2015",
  "editora": "Gulf Professional Publishing",
  "url": [
    "http://www.sciencedirect.com/science/book/9780128012567"
  ],
  "_version_": 1519109898103685149,
  "score": 1.6233432
},
{
  "id": "9780080999715",
  "titulo": [
    "Natural Gas Processing"
  ],
  "ano": "2014",
  "editora": "Gulf Professional Publishing",
  "url": [
    "http://www.sciencedirect.com/science/book/9780080999715"
  ],
  "_version_": 1519109898114170890,
  "score": 1.6233432
}
]
}

```

#### Conclusão do teste:

Os documentos retornados quando a funcionalidade *boosting* estava habilitada possuíam o termo profissional no campo titulo. Por outro lado, quando a funcionalidade estava desabilitada, o termo profissional não estava presente no campo titulo em alguns documentos, mas somente nos campos editora, por exemplo.

Pode-se verificar, então, que habilitar o *boosting* para o campo titulo faz o efeito desejado, dando mais relevância a esse campo.

#### Conclusões acerca dos testes

Pode-se verificar que para todas as consultas, o número máximo de documentos retornados foi 15, como configurado no arquivo solrconfig.xml.

Além disso, como configurado, todos os documentos possuem um campo informando o *score* recebido pelo sistema de recuperação de informação para o *rankeamento*.

Para o *spellcheck* e *suggest*, é possível verificar que ambas as funcionalidades retornam somente uma sugestão, como requerido no arquivo de configuração.

Em relação aos *facets*, pelo teste 1 é possível afirmar que só são listados os anos e os tópicos dos documentos presentes no resultado. São excluídos da listagem de *facets* a quantidade de documentos que não possuem valores para o campo *ano/TOPICO*, como requerido na configuração.

Por fim, pode-se concluir que, pelos testes explicitados aqui (e diversos outros que não foram reportados), as configurações funcionam de modo harmonioso provendo um sistema de recuperação de informação satisfatório.

## Estatísticas

A coleção do projeto possui um total de 35972 livros. Os dados foram retirados de 11 arquivos .xlsx e 21 arquivos .sgm, como referenciado na seção "*Populando a coleção*".

Para os 11 arquivos .xlsx, são adicionados cerca de 14 mil documentos à coleção. Por outro lado, para os 21 arquivos .sgm são adicionados outros 21 mil documentos aproximadamente.



## Considerações finais

Ao final desse projeto o objetivo estipulado inicialmente foi concluído. Foi possível obter um bom conhecimento acerca do sistema de recuperação de informação *Solr* e também houve um aprendizado útil acerca dos procedimentos de configuração do mesmo.

Em suma, foi possível vincular diversas funcionalidades do *Solr* com conceitos vistos durante o decorrer da disciplina, como o *spellcheck*, *snippets*, *stopwords*, sinônimos, *rankeamento* por zonas de arquivo etc..