**I - Como gerar um índice novo para o DeDup a partir de uma base de dados MySQL e estando logado no servidor.**

Para a geração de um índice que possa ser utilizado para o DeDup são necessários 4 coisas:

* O shell de execução (/home/javaapps/sbt-projects/DCDup/bin/local/MySQL2Lucene.sh) com as bibliotecas correspondentes. Estes podem ser encontrados tanto no servidor de teste (serverofi5) como no de produção.
* Acesso ao servidor MySQL (url, user, password, database) contendo os metadados dos documentos a serem utilizados na geração do índice.
* Arquivo (ou arquivos) contendo expressão sql a ser utilizada no MySQL para a extração dos documentos a serem indexados. O local padrão para tal arquivo é no diretório /home/javaapps/sbt-projects/DCDup/sql tanto no servidor de teste (serverofi5) como no de produção. A extensão padrão é \*.sql
* Arquivo de schema que define como cada campo do arquivo pipe gerado será utilizado pelo DeDup no processo de se encontrar documentos duplicados. O local padrão para tal arquivo é no diretório /home/javaapps/DeDup/work tanto no servidor de teste (serverofi5) como no de produção. A extensão padrão é \*.cfg

A geração do índice ocorre com a execução do shell MySQL2Lucene.sh com os seguintes parâmetros:

operacao@serverofi5:/home/javaapps/sbt-projects/DCDup/bin/local$ ./MySQL2Lucene.sh

MySQL2Lucene shell takes a list of documents retrieved from a MySQL database and

creates a local DeDup index with then. If such index already exists, it will rewritten.

usage: MySQL2Lucene

-host=<MySQL server> - IP or domain of the MySQL server

-user=<MySQL user> - MySQL user name

-pswd=<MySQL password> - MySQL user password

-dbnm=<MySQL database> - MySQL database name

-sqls=<MySQL sql file list> - comma separated sql file names

-index=<indexPath> - NGrams Lucene index path

-schema=<schemaFile> - NGram schema file

[-port=<MySQL\_Port>] - MySQL port

[-sqlEncoding=<sqlEncoding>] - sql file encoding. Default is utf-8

[-jsonField=<tag>[,<tag>,...,<tag>]] - if a column element is a json element, indicates which json elements

to retrieve the content. Default are text,\_f,\_e.

[-repetitiveField=<name>[.<name>,...,<name>]] - the name of the fields that should be broken into a new line

when the repetitive separator symbol is found. Default is title.

[-repetitiveSep=<separator>] - repetitive field string separator. Default is //@//

[-jsonLangField=<jsonLangField>] - the json field that store the language indicator. If present it will be

used to suffix the id field with the language.

[-idFieldName=<name>] - id field name of the mysql record]. Will be used to prefix the id with the language.

if the jsonLangField is specified.

[-schemaEncoding=<schemaFileEncoding>] - NGram schema file encoding. Default is utf-8.

[--append] - append documents to an existing Lucene index.

Segue uma explicação abreviada dos parâmetros obrigatórios:

Os parâmetros -host, -user, -pswd e -dbnm descrevem respectivamente o servidor do MySQL assim como o nome do usuário e do banco de dados.

O parâmetro -sqls especifica uma lista separada por vírgulas de um ou mais arquivos sql que serão utilizados para se extrair o arquivo pipe descrito pelo parâmetro -schema.

O parâmetro -schema indica o nome/caminho do arquivo do tipo schema que descreve todos os campos a serem utilizados na criação do índice e suas respectivas posições no arquivo pipe.

O parâmetro -index dá o nome e o caminho do índice Lucene a ser criado ou apendado.

Executado o shell e tendo sua execução terminado corretamente, o índice pode ser utilizado para a procura de documentos duplicados por outros shells. Tal índice, porém, ainda não será visível na aplicação web DeDup (<http://serverof5.bireme.br:8180/DeDup> no servidor de teste ou <https://dedup.bireme.org/> no servidor de produção).

Para tanto devemos informar a aplicação sobre a existência deste novo índice seguindo as explicações da seção “Inclusão de um novo índice na aplicação DeDup”.

**II - Como gerar um índice novo para o DeDup a partir de uma base de dados MySQL sem estar logado no servidor.**

Para a geração de um índice que possa ser utilizado para o DeDup são necessários 4 coisas:

* O shell de execução (/home/javaapps/sbt-projects/DCDup/bin/remote/MySQL2Lucene.sh) com as bibliotecas correspondentes no servidor de teste (serverofi5) ou no servidor de produção. Localmente fazendo-se um clone do projeto <https://github.com/bireme/DCDup> e executando-se os shells no diretório ‘bin’.
* Acesso ao servidor MySQL (url, user, password, database) contendo os metadados dos documentos a serem utilizados na geração do índice.
* Arquivo (ou arquivos) contendo expressão sql a ser utilizada no MySQL para a extração dos documentos a serem indexados. O local padrão para tal arquivo é no diretório /home/javaapps/sbt-projects/DCDup/sql tanto no servidor de teste (serverofi5) como no de produção. A extensão padrão é \*.sql
* Nome do schema que define os campos a serem utilizados pelo DeDup no processo de se encontrar documentos duplicados. O schema e o índice já devem estar registrados na aplicação DeDep, veja a seção “Inclusão de um novo índice na aplicação DeDup”.

A geração do índice ocorre com a execução do shell MySQL2Lucene.sh com os seguintes parâmetros:

operacao@serverofi5:/home/javaapps/sbt-projects/DCDup/bin/remote$ ./MySQL2Lucene.sh

MySQL2Lucene shell takes a list of documents retrieved from a MySQL database and

creates a remote DeDup index with them. If such index already exists, it will rewritten.

usage: MySQL2Lucene

-host=<MySQL server> - IP or domain of the MySQL server

-user=<MySQL user> - MySQL user name

-pswd=<MySQL password> - MySQL user password

-dbnm=<MySQL database> - MySQL database name

-sqls=<MySQL sql file list> - comma separated sql file names

-index=<indexName> - DeDup index name

-schema=<schemaName> - DeDup schema name

[-port=<MySQL\_Port>] - MySQL port

[-sqlEncoding=<sqlEncoding>] - sql file encoding. Default is utf-8

[-jsonField=<tag>[,<tag>,...,<tag>]] - if a column element is a json element, indicates which json elements

to retrieve the content. Default are text,\_f,\_e.

[-repetitiveField=<name>[.<name>,...,<name>]] - the name of the fields that should be broken into a new line

when the repetitive separator symbol is found. Default is title.

[-repetitiveSep=<separator>] - repetitive field string separator. Default is //@//

[-jsonLangField=<jsonLangField>] - the json field that store the language indicator. If present it will be

used to suffix the id field with the language.

[-idFieldName=<name>] - id field name of the mysql record]. Will be used to prefix the id with the language.

if the jsonLangField is specified.

[--append] - append documents to an existing Lucene index.

Segue uma explicação abreviada dos parâmetros obrigatórios:

Os parâmetros -host, -user, -pswd e -dbnm descrevem respectivamente o servidor do MySQL assim como o nome do usuário e do banco de dados.

O parâmetro -sqls especifica uma lista separada por vírgulas de um ou mais arquivos sql que serão utilizados para se extrair o arquivo pipe descrito pelo parâmetro -schema.

O parâmetro -schema indica o nome do schema que descreve todos os campos a serem utilizados na criação do índice.

O parâmetro -index dá o nome do índice Lucene a ser criado ou apendado.

Executado o shell e tendo sua execução terminado corretamente, o índice pode ser utilizado para a procura de documentos duplicados por outros shells.

**III - Inclusão de um novo índice na aplicação DeDup**

Para que um novo índice seja incluído no DeDup é necessária a edição do arquivo /home/javaapps/DeDup/config\_LILACS.cfg

Neste arquivo adiciona-se caso não exista, o elemento xml <schema> contendo o nome do schema e o caminho do mesmo. O nome apresentado na aplicação será o mesmo utilizado neste parâmetro.

Por exemplo, para o índice ‘meuIndice’ o elemento seria:

<schema>

<name>schema para o meu índice</name>

<path>work/configMeuIndice.cfg</path>

<encoding>UTF-8</encoding>

</schema>

Depois adiciona-se o elemento xml <index> onde se coloca o nome do índice (tal como se deseja que se apresente na aplicação) e o seu caminho tal qual utilizado no parâmetro <indexName> utilizado na geração do índice. O caminho é relativo em relação ao diretório do DeDup.

Por exemplo, para o índice ‘meuIndice’ o elemento seria:

<index>

<name>o meu índice</name>

<path>work/meuIndice</path>

</index>

Uma vez acrescentados os dois parâmetros e gravado o arquivo, faz-se necessário reiniciar a aplicação DeDup no Tomcat correspondente (presente no diretório /usr/local/tomcat8 tanto no servidor de teste como no de produção). Para executar tal tarefa, faz-se necessário o usuário estar cadastrado como autorizado na aplicação gráfica ‘manager’ do Tomcat ou pela linha de comando.

O mais fácil é pedir ao ITI realizar para realizar tal tarefa.

Agora o índice deve estar presente no DeDup ([http://serverof5.bireme.br:8180/DeDup/services](http://serverof5.bireme.br:8180/DeDup) (no servidor de teste ou [https://dedup.bireme.org/services](https://dedup.bireme.org/) no servidor de produção).

**IV - Como gerar um índice novo para o DeDup a partir de um arquivo pipe e estando logado no servidor.**

Para a geração de um índice que possa ser utilizado para o DeDup são necessários 3 coisas:

* O shell de execução (/home/javaapps/sbt-projects/DCDup/bin/local/Pipe2Lucene.sh) com as bibliotecas correspondentes. Estes podem ser encontrados tanto no servidor de teste (serverofi5) como no de produção.
* Arquivo pipe contendo os campos necessários para a criação do índice e que siga a descrição do mesmo feita pelo respectivo schema armazenado no arquivo de schema.
* Arquivo de schema que define como cada campo do arquivo pipe gerado será utilizado pelo DeDup no processo de se encontrar documentos duplicados. O local padrão para tal arquivo é no diretório /home/javaapps/DeDup/work tanto no servidor de teste (serverofi5) como no de produção. A extensão padrão é \*.cfg

A geração do índice ocorre com a execução do shell Pipe2Lucene.sh com os seguintes parâmetros:

operacao@serverofi5:/home/javaapps/sbt-projects/DCDup/bin/local$ ./Pipe2Lucene.sh

Pipe2Lucene shell takes a list of documents (defined by an schema

file) from a piped file and insert then into a local DeDup index.

usage: Pipe2Lucene

-index=<indexPath> - DeDup index path

-schema=<schemaFile> - DeDup schema file

-pipe=<pipeFile> - pipe file

-pipeEncoding=<pipeEncoding> - pipe file encoding

[-schemaFileEncod=<schemaFileEncoding>] - NGram schema file encoding. Default is utf-8

[--append] - append documents to an existing DeDup index

Segue uma explicação abreviada dos parâmetros obrigatórios:

O parâmetro -index dá o nome e o caminho do índice Lucene a ser criado ou apendado.

O parâmetro -schema indica o nome/caminho do arquivo do tipo schema que descreve todos os campos a serem utilizados na criação do índice e suas respectivas posições no arquivo pipe.

O parâmetro -pipe indica o nome/caminho do arquivo pipe contendo o conteúdo dos campos a serem indexados.

O parâmetro -pipeEncoding indica a codificação do texto utilizada no arquivo pipe.

Executado o shell e tendo sua execução terminado corretamente, o índice pode ser utilizado para a procura de documentos duplicados por outros shells. Tal índice, porém, ainda não será visível na aplicação web DeDup (<http://serverof5.bireme.br:8180/DeDup> no servidor de teste ou <https://dedup.bireme.org/> no servidor de produção).

Para tanto devemos informar a aplicação sobre a existência deste novo índice seguindo as explicações descritas acima na seção “Inclusão de um novo índice na aplicação DeDup”.

**V - Como gerar remotamente um índice novo para o DeDup a partir de um arquivo pipe.**

Para a geração de um índice que possa ser utilizado para o DeDup são necessários 3 coisas:

* O shell de execução (/home/javaapps/sbt-projects/DCDup/bin/remote/Pipe2Lucene.sh) com as bibliotecas correspondentes. Estes podem ser encontrados tanto no servidor de teste (serverofi5) como no de produção.
* Arquivo pipe contendo os campos necessários para a criação do índice seguindo as especificações feitas no respectivo arquivo de schema.
* O índice a ser criado assim como o seu respectivo schema já definidos na aplicação DeDup como explicado na seção “Inclusão de um novo índice na aplicação DeDup”.

A geração do índice ocorre com a execução do shell Pipe2Lucene.sh com os seguintes parâmetros:

operacao@serverofi5:/home/javaapps/sbt-projects/DCDup/bin/remote$ ./Pipe2Lucene.sh

Pipe2Lucene shell takes a list of documents (defined by an schema

file) from a piped file and insert then into a remote DeDup index.

usage: Pipe2Lucene

-dedupUrl=<DeDupBaseUrl> - DeDup url service (for ex: http://dedup.bireme.org/services)

-index=<indexPath> - DeDup index name

-schema=<schemaFile> - DeDup schema name

-pipe=<pipeFile> - pipe file

-pipeEncoding=<pipeEncoding> - pipe file encoding

[--append] - append documents to an existing DeDup index

Segue uma explicação abreviada dos parâmetros obrigatórios:

O parâmetro -dedupUrl indica a localização (URL) da aplicação DeDup

O parâmetro -index dá o nome do índice Lucene a ser criado ou apendado.

O parâmetro -schema indica o nomedo schema que descreve todos os campos a serem utilizados na criação do índice e suas respectivas posições no arquivo pipe.

O parâmetro -pipe indica o nome/caminho do arquivo pipe contendo o conteúdo dos campos a serem indexados.

O parâmetro -pipeEncoding indica a codificação do texto utilizada no arquivo pipe.

Executado o shell e tendo sua execução terminado corretamente, o índice pode ser utilizado para a procura de documentos duplicados por outros shells. O novo índice será visível na aplicação web DeDup (<http://serverof5.bireme.br:8180/DeDup> no servidor de teste ou <https://dedup.bireme.org/> no servidor de produção).

**VI – Como gerar um arquivo pipe a partir de um banco de dados MySQL**

Para a geração de um arquivo pipe que possa ser utilizado para a criação de um índice Lucene utilizado pelo DeDup são necessários 3 coisas:

* O shell de execução (/home/javaapps/sbt-projects/DCDup/bin/MySQL2Pipe.sh) com as bibliotecas correspondentes. Estes podem ser encontrados tanto no servidor de teste (serverofi5) como no de produção ou localmente clonando-se o projeto <https://github.com/bireme/DCDup> e indo no diretório ‘bin’.
* Acesso ao servidor MySQL (url, user, password, database) contendo os metadados dos documentos a serem utilizados na geração do índice.
* Arquivo (ou arquivos) contendo expressão sql a ser utilizada no MySQL para a extração dos documentos a serem indexados. O local padrão para tal arquivo é no diretório /home/javaapps/sbt-projects/DCDup/sql tanto no servidor de teste (serverofi5) como no de produção. A extensão padrão é \*.sql

A geração do índice ocorre com a execução do shell MySQL2Pipe.sh com os seguintes parâmetros:

operacao@serverofi5:/home/javaapps/sbt-projects/DCDup/bin/local$ ./MySQL2Pipe.sh

MySQL2Pipe shell takes a list of documents retrieved from a MySQL database

and creates a local pipe file following a given schema file.

usage: MySQL2Pipe

-host=<MySQL server> - IP or domain of the MySQL server

-user=<MySQL user> - MySQL user name

-pswd=<MySQL password> - MySQL user password

-dbnm=<MySQL database> - MySQL database name

-sqls=<MySQL sql file list> - comma separated sql file names

-pipe=<pipeFile> - output piped file

[-port=<MySQL\_Port>] - MySQL port

[-sqlEncoding=<sqlEncoding>] - sql file encoding. Default is utf-8

[-pipeEncoding=<pipeEncoding>] - output piped file encoding. Default is utf-8

[-jsonField=<tag>[,<tag>,...,<tag>]] - if a column element is a json element, indicates

which json elements to retrieve the content

[-repetitiveField=<name>[.<name>,...,<name>]] - the name of the fields that should be broken into a new

line when the repetitive separator symbol is found

[-repetitiveSep=<separator>] - repetitive field string separator. Default is //@//

[-jsonLangField=<jsonLangField>] - the json field that store the language indicator.

If present it will be used to suffix the id field with the language

[-idFieldName=<name>] - id field name of the mysql record]. Will be used to prefix the id with the language

if the jsonLangField is specified

Segue uma explicação abreviada dos parâmetros obrigatórios:

Os parâmetros -host, -user, -pswd e -dbnm descrevem respectivamente o servidor do MySQL assim como o nome do usuário e do banco de dados.

O parâmetro -sqls especifica uma lista separada por vírgulas de um ou mais arquivos sql que serão utilizados para se extrair o arquivo pipe descrito pelo parâmetro -schema.

O parâmetro -pipe indica o nome/caminho do arquivo pipe a ser criado contendo o conteúdo dos campos a serem indexados.