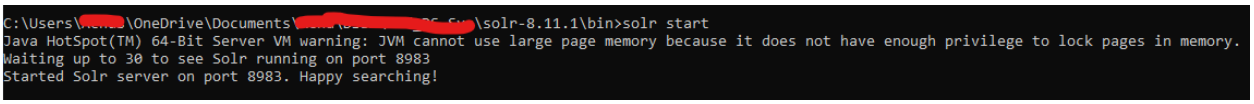
Steps to install Solr and index Wikipedia data dump

PRE REQUISTE CHECK : Make sure you have Java Runtime Environment (JRE) version 1.8 or higher in your machine Refer here for more information on system requirements.

Download and Install Solr

1. Download the latest solr version from this link and unzip in the local machine.
   1.1. Choose the 'Binary' release only.

2. Use Command prompt to go to bin (cd ..\solr-8.11.1\bin) and start running solr using the command ***solr start***
   2.1. Once solr starts running successfully you will see the message *"Started Solr server on port 8983. Happy searching!"*



Solr- Create Core and Index exampledocs

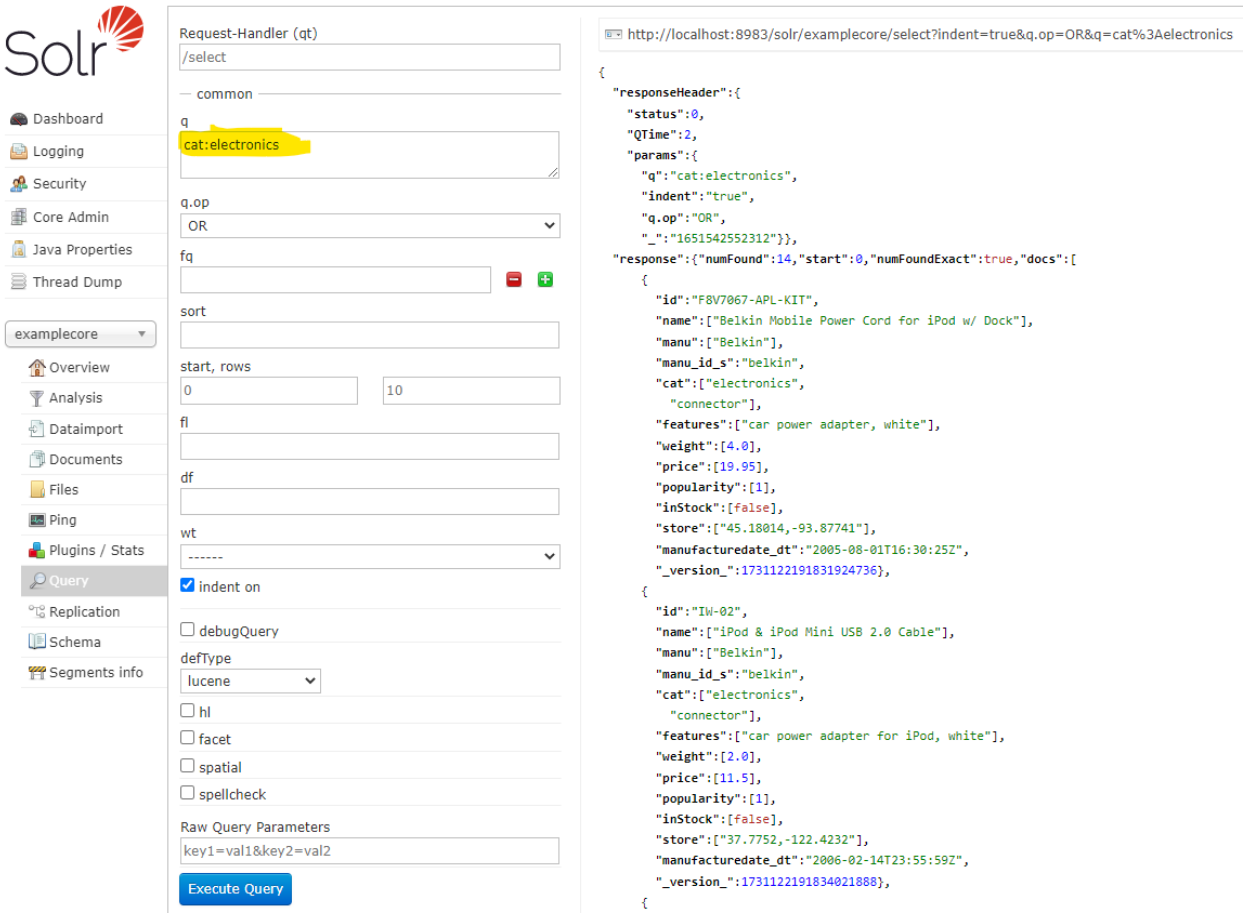
3. First step is to create a core.
   3.1. Create the test core using the command ***solr create -c <core_name>***

4. Next step is indexing the documents under exampledocs.
   4.1. For this you need to go back up to solr-8.11.1 in the command prompt and on Windows machine, run the command
   java -jar -Dc=<core_name> -Dauto example\exampledocs\post.jar example\exampledocs*

Note:(Linux/Mac) ***bin/post -c <core_name> example/exampledocs/****

Solr - Search on exampledocs

5. Now it is time to view the Solr search user interface
   5.1. Admin console can be initiated using the url: http://localhost:8983/solr/#/
   5.2. Choose the core from the "Core Selector" dropdown
   5.3. More options show up under the section for Core Selector. Click on Query
   5.4. By default, wild card search is set .So click on Execute Query to see the results
   5.5. For specific searches, please follow the syntax **cat:electronics** (this indicates the string 'electronics' to be searched against an attribute 'cat' (meaning category).



Note: Query interface can also be accessed http://localhost:8983/solr/#/<core_name>/query

Download Wikipedia datadump

6. Now that we know Solr is up and running we will work on extracting wiki data dump and indexing the files. You can download the data dump (xml) from wikipedia using https://dumps.wikimedia.org/.

Solr - Create Core for wikipedia search

7. The next step is to create a new core for the wiki data dump using the using the command ***solr create -c <wiki_core_name>***

Solr - Setting up input file for wikipedia search

8. Now go to the core folder (C:\Users\....\solr-8.11.1\server\solr\<wiki_core_name>\)that was created locally on the machine and create a new folder - ***input***

9. Unzip the downloaded wiki data dump file and save the .xml file in the input folder (folder path- “C:\Users\....\solr-8.11.1\server\solr\<wiki_core_name>\input) created above.

10. Go to the input folder (folder path- “C:\Users\....\solr-8.11.1\server\solr\<wiki_core_name>\input) and confirm the file - enwiki-2022xxxx-pages-articles-multistream*.xml exist

Solr - Configuration changes for wikipedia search

11. Next step is to work on the configuration files. **Please see attached copies of solrconfig.xml , dat-config.xml and schema.xml files for reference**. Go to the conf folder (C:\Users\....\solr-8.11.1\server\solr\<wiki_core_name>\conf)

11.1. Rename the Managed-schema.xml to schema.xml

11.2. Open schema.xml in a text editor and add the following lines of code provided below.

```
<field name="_version_" type="plong" indexed="true" stored="true"/>
<field name="id" type="string" indexed="true" stored="true" required="true"/>
<field name="title" type="string" indexed="true" stored="true"/>
<field name="revision" type="pint" indexed="true" stored="true"/>
<field name="user" type="string" indexed="true" stored="true"/>
<field name="userId" type="pint" indexed="true" stored="true"/>
<field name="text" type="text_en" indexed="true" stored="true"/>
```

Snippet 11.2

11.3. Open solrconfig.xml in a text editor and add the following snippets of code.

11.3.1. Within the config section (towards beginning) add the following line

```
<lib dir="${solr.install.dir:../../../..}/dist/" regex="solr-dataimporthandler-\d.*\.jar" />
```

Snippet 11.3.1

11.3.2. Add the following line of code to make sure that solr uses the file that we renamed to schema.xml

```
<schemaFactory class="ClassicIndexSchemaFactory"/>
```

Snippet 11.3.2

11.3.3. Search for “update.autoCreateFields” and set it to false as shown in the code snippet below.

```
<updateRequestProcessorChain name="add-unknown-fields-to-the-schema"
default="${update.autoCreateFields:false}"
        processor="uuid,remove-blank,field-name-mutating,parse-boolean,parse-long,parse-d
ouble,parse-date,add-schema-fields">
  <processor class="solr.LogUpdateProcessorFactory"/>
  <processor class="solr.DistributedUpdateProcessorFactory"/>
  <processor class="solr.RunUpdateProcessorFactory"/>
</updateRequestProcessorChain>
```

Snippet 11.3.3

11.3.4. Add the following lines of code in the section for request handler.The code below initiates the data input handler class and reads the data-config.xml file

```
<requestHandler name="/dihupdate" class="org.apache.solr.handler.dataimport.DataImportHandler"
startup="lazy" >
  <!-- initialization args may optionally be defined here -->
  <lst name="defaults">
    <str name="config">data-config.xml</str>
  </lst>
</requestHandler>
```

Snippet 11.3.4

11.4. Create a new blank xml file and name it as data-config.xml. Paste the following snippet of code into the xml file and update the highlighted folder path for URL in data-config.xml to match the folder structure on your local machine

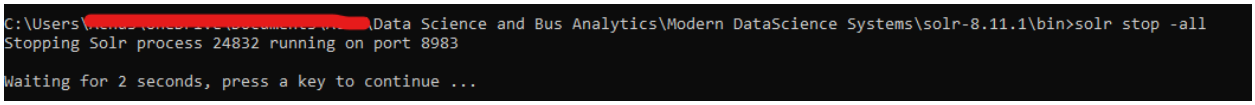
```
<dataConfig>
        <dataSource type= "FileDataSource" encoding="UTF-8" />
        <document>
        <entity name="page"
                                        processor="XPathEntityProcessor"
                                        stream ="true"
                                        forEach="/mediawiki/page/"
                                        url="<complete URL including the .xml file extension to the downloaded
wikipedia xml file placed in the input folder>"
                                        transformer="RegexTransformer,DateFormatTransformer" >
                                <field column ="id"  xpath="/mediawiki/page/id" />
                                <field column ="title"  xpath="/mediawiki/page/title" />
                                <field column ="revision"  xpath="/mediawiki/page/revision/id" />
                                <field column ="user"  xpath="/mediawiki/page/revision/contributor/username" />
                                <field column ="userId"  xpath="/mediawiki/page/revision/contributor/id" />
                                <field column ="text"  xpath="/mediawiki/page/revision/text" />
```

```
                        <field column ="timestamp"  xpath="/mediawiki/page/revision/timestamp"
dateTimeFormat="yyyy-MM-dd'T'hh:mm:ss'Z'"/>
                        <field column ="$skipDoc"  regex="^#REDIRECT .*" replaceWith="true"
sourceColName="text"/>
            </entity>
            </document>
</dataConfig>
```

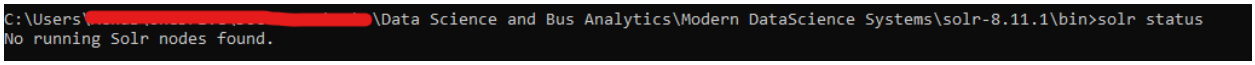
Snippet 11.4

Solr - Restart Solr for starting/set up of wikipedia core

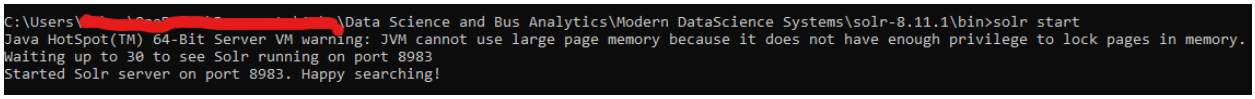
12. Now that configuration files have been updated, we need to re-start solr.
    12.1. Stop all instances of solr using the command *bin/solr stop -all*



    12.2. Check the status of solr using the command *bin/solr status*



    12.3. Start solr using the command *bin/solr start*



Solr - Admin console and Indexing wikipedia xml file

13. Access the Admin console using the url: http://localhost:8983/solr/#/

    13.1. If we configured everything correctly, the Admin console page would open without any error.If not, there would be error messages that would need to be fixed before you can proceed further.

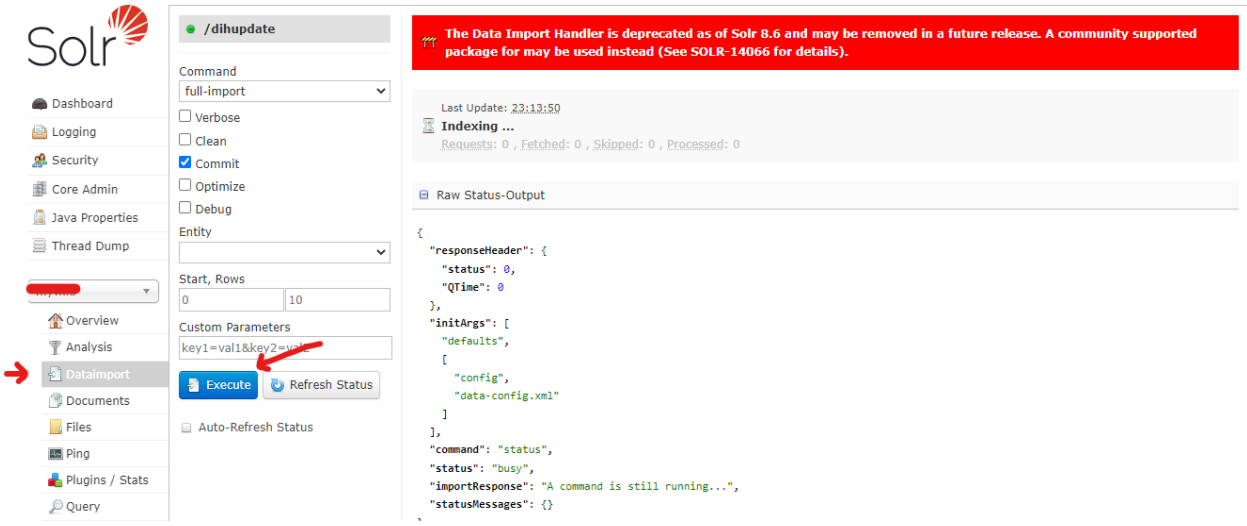
    13.2. Call the Request handler we configured using the url below to do a full import and index the wikipages.
    http://localhost:8983/solr/<wiki_core_name>/dihupdate?command=full-extract

    13.3. Check the status of the Request Handler by using the following code
    http://localhost:8983/solr/<wiki_core-name>/dihupdate

    13.4. Go to the Core_Selector on the Admin console, select the core - <wiki_core_name>,

13.4.1. click on **DataImport** on the left navigation bar under the section where you choose core (see screen shot below)

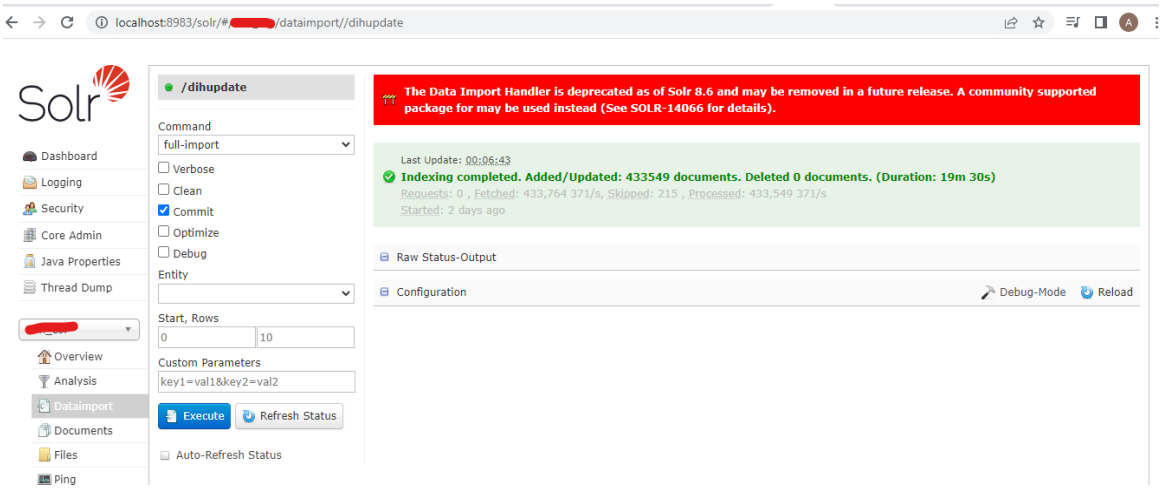
13.4.2. click **Execute** button to start indexing the documents



13.5. To check the status of the import and indexing, keep clicking on the refresh Status button beside the “Execute” button

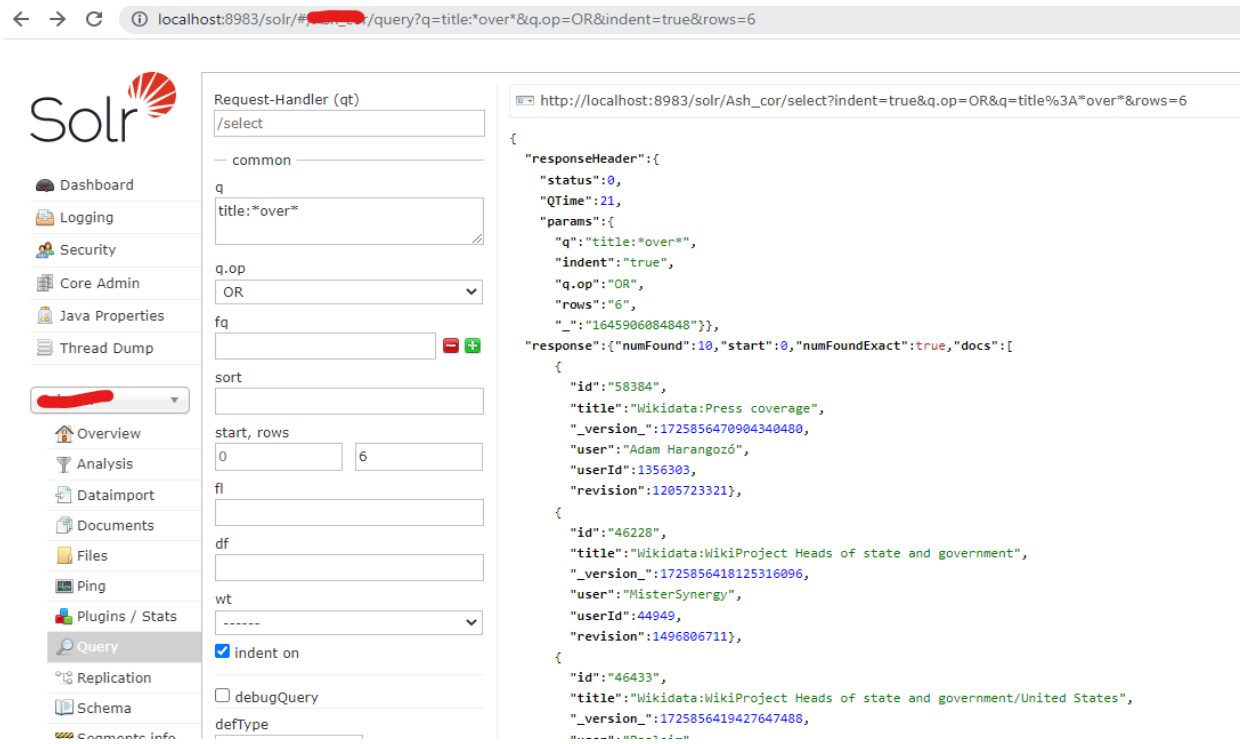
13.6. Instead of steps 13.4 and 13.5, we can also index wiki data dump by using the following url http://localhost:8983/solr/<wiki_core_name>/dihupdate?command=full-import

13.7. When solr finishes the indexing, you will get the following message on the screen. Please note that the Red message in the screenshot below is a warning and can be ignored for now.



Solr - Admin console and Search on wikipedia datadump

13.8. Now we can go to the Query tab on the Admin console, and start searching. The example query shown below searches for any words containing the text 'over' in the title. The format for the search is: title:*over*



13.9. You can see that the results include tiles like "Wikidata:Press coverage" , "Wikidata:WikiProject Heads of state and government" etc

Solr - User Interface using Velocity plug-in

14. Now we will set up the Velocity Response Writer (more intuitive UI provided by Solr)

14.1. Copy the following jar files from the folder - C:\Users\....\solr-8.11.1\contrib\velocity to the folder C:\Users\....\solr-8.11.1\server\solr-webadpp\webapp\WEB-INF\lib

- commons-lang3-3.10.jar
- velocity-engine-core-2.3.jar
- velocity-tools-generic-3.1.jar
- velocity-tools-view-3.1.jar
- velocity-tools-view-jsp-3.1.jar

Solr - Configuration changes for Velocity User Interface

14.2. Now open the solrconfig.xml under the core for wiki (folder path-

"C:\Users\....\solr-8.11.1\server\solr\<wiki_core_name>\config")

14.3. Go to the QueryResponseWriter section(in solrconfig.xml) and add the following snippet of code

```
<queryResponseWriter name="velocity" class="solr.VelocityResponseWriter">
<!--<str name="template.base.dir">${velocity.template.base.dir:}</str>-->
<str name="template.base.dir">< folder path to solr-8.11.1\server\solr-webadpp\webapp\WEB-INF\lib</str>
<str name="init.properties.file">velocity-init.properties</str>
<bool name="params.resource.loader.enabled">false</bool>
<bool name="solr.resource.loader.enabled">true</bool>
</queryResponseWriter>
```

Snippet 14.3

14.4. Add the following snippet of code in the RequestHandler section (in solrconfig.xml) to add a '/browse' handler that can enable the velocity UI template

```
<requestHandler name="/browse" class="solr.SearchHandler" useParams="query,velocity,browse">
  <lst name="defaults">
    <str name="echoParams">explicit</str>

                        <!-- VelocityResponseWriter settings -->
                        <str name="wt">velocity</str>
                        <str name="v.template">browse</str>
                        <str name="v.layout">layout</str>
                        <str name="title">Solritas</str>

                        <!-- Query settings -->
                        <str name="defType">edismax</str>
                        <str name="qf"> title^10.0 id^10.0 user^5.0 text^9.0
                        </str>
                        <str name="df">title</str>
                        <str name="mm">100%</str>
                        <str name="q.alt">*:*</str>
                        <str name="rows">10</str>
                        <str name="fl">*,score</str>
                        <str name="mlt.qf"> _version_^1.0 title^10.0 id^10.0 user^2.0 revision^1.5 text^9.0
                        </str>
                        <str name="mlt.fl">title,id,user,_version_,userId,revision,text</str>
                        <int name="mlt.count">3</int>

                        <!-- Faceting defaults -->
                        <str name="facet">off</str>
                        <str name="facet.missing">true</str>
                        <str name="facet.field">user</str>
                        <!--<str name="facet.field">manu_exact</str>
                        <str name="facet.field">content_type</str>
```

```
                        <str name="facet.field">author_s</str>-->
                        <!--<str name="facet.mincount">1</str> -->

                        <!-- Highlighting defaults -->
                        <str name="hl">on</str>
                        <str name="hl.fl">title id user revision</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.title.hl.fragsize">0</str>
                        <str name="f.title.hl.alternateField">title</str>
                        <str name="f.name.hl.fragsize">0</str>
                        <str name="f.name.hl.alternateField">name</str>
                        <str name="f.content.hl.snippets">3</str>
                        <str name="f.content.hl.fragsize">200</str>
                        <str name="f.content.hl.alternateField">content</str>
                        <str name="f.content.hl.maxAlternateFieldLength">750</str>

                        <!-- Spell checking defaults -->
                        <str name="spellcheck">on</str>
                        <str name="spellcheck.extendedResults">false</str>
                        <str name="spellcheck.count">5</str>
                        <str name="spellcheck.alternativeTermCount">2</str>
                        <str name="spellcheck.maxResultsForSuggest">5</str>
                        <str name="spellcheck.collate">true</str>
                        <str name="spellcheck.collateExtendedResults">true</str>
                        <str name="spellcheck.maxCollationTries">5</str>
                        <str name="spellcheck.maxCollations">3</str>
                    </lst>
  </requestHandler>
  <!-- SearchHandler
```

Snippet 14.4

Solr - Restart Solr to use the Velocity User Interface

14.5. Now that configuration files have been updated, we need to re-start solr.

14.5.1. Stop all instances of solr using the command *bin/solr stop -all*

```
C:\Users\[redacted]\Data Science and Bus Analytics\Modern DataScience Systems\solr-8.11.1\bin>solr stop -all
Stopping Solr process 24832 running on port 8983

Waiting for 2 seconds, press a key to continue ...
```

14.5.2. Check the status of solr using the command *bin/solr status*

```
C:\Users\[redacted]\Data Science and Bus Analytics\Modern DataScience Systems\solr-8.11.1\bin>solr status
No running Solr nodes found.
```

14.5.3. Start solr using the command *bin/solr start*

```
C:\Users\[redacted]\Data Science and Bus Analytics\Modern DataScience Systems\solr-8.11.1\bin>solr start
Java HotSpot(TM) 64-Bit Server VM warning: JVM cannot use large page memory because it does not have enough privilege to lock pages in memory.
Waiting up to 30 to see Solr running on port 8983
Started Solr server on port 8983. Happy searching!
```

Solr - Search on Velocity User Interface

14.6. Now you can access the user friendly interface by using the following code http://localhost:8983/solr/<wiki_core-name>/browse

14.6.1. Just type in what you need to search in the Find/search bar and we get the results of our search