**Setup & Run document for Scraper Plugins Application**

Below are the list of projects and input files required to run Scraper Plugin for the document type “**Proxy Statement 14A**” for the Company **DDR**,

1. holmes-ekyc-scraperplugins,
2. cognitive\_ekyc\_features
3. TextProcessor\_v1
4. DocumentAnalysis
5. Entityextractor
6. holmes-ekyc-pdfannotationservices
7. holmes-ekyc-utils
8. Sample document ‘Proxy Statement 14A’ for company DDR (DDR-CORP-DEF-14A.pdf)
9. Input JSON file (Ownership\_Info\_Extraction\_JSON\_ObjOne.txt)
10. Pattern file

a) INDIVIDUAL.pattern

b) ORGANIZATION.pattern

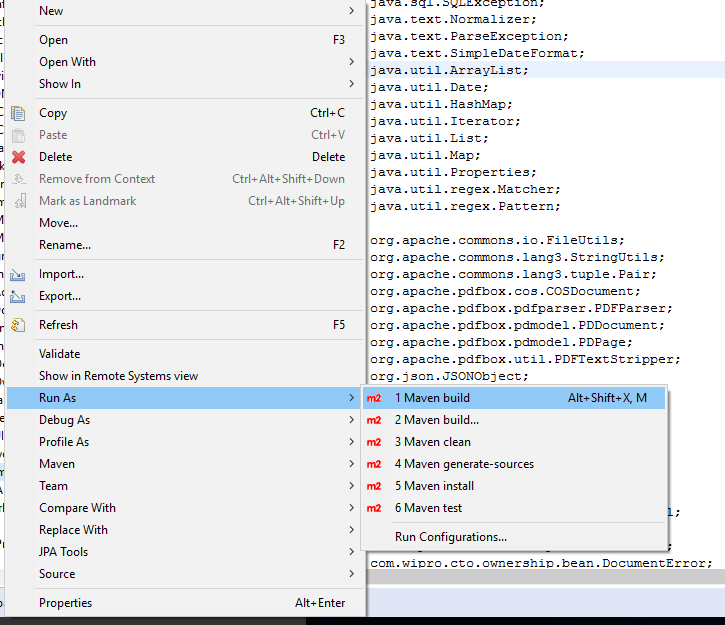
**Build the required Applications: -**

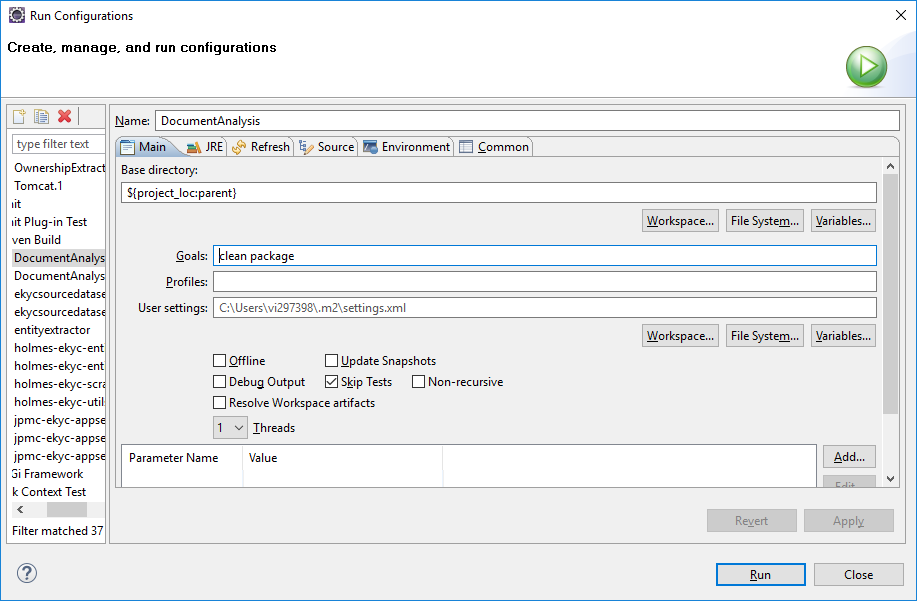
After checked out all mentioned projects, need to **build and packed** each one,

Select the **pom.xml** file of parent directory of each project,

right click on pom.xml -- > Run as -- > Maven build

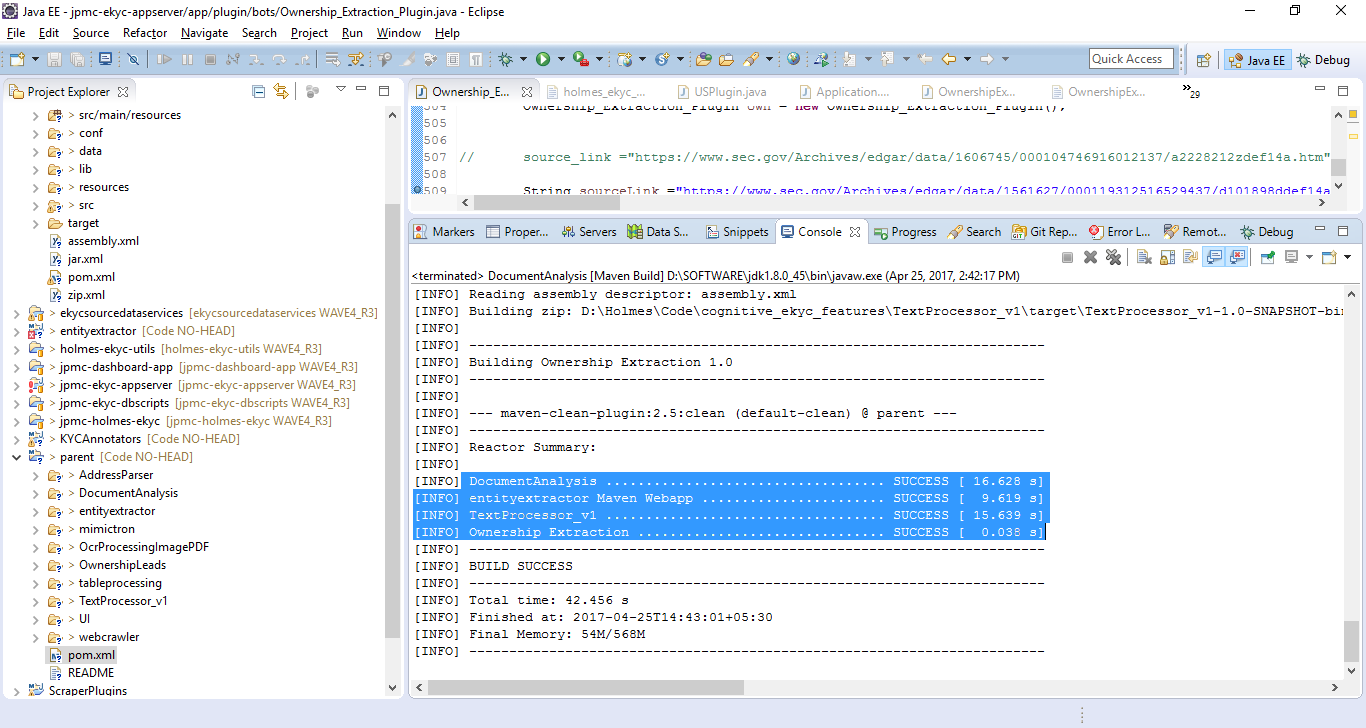
Set Goal as **clean package** as per below screen shots and click on **Run** to build the project.



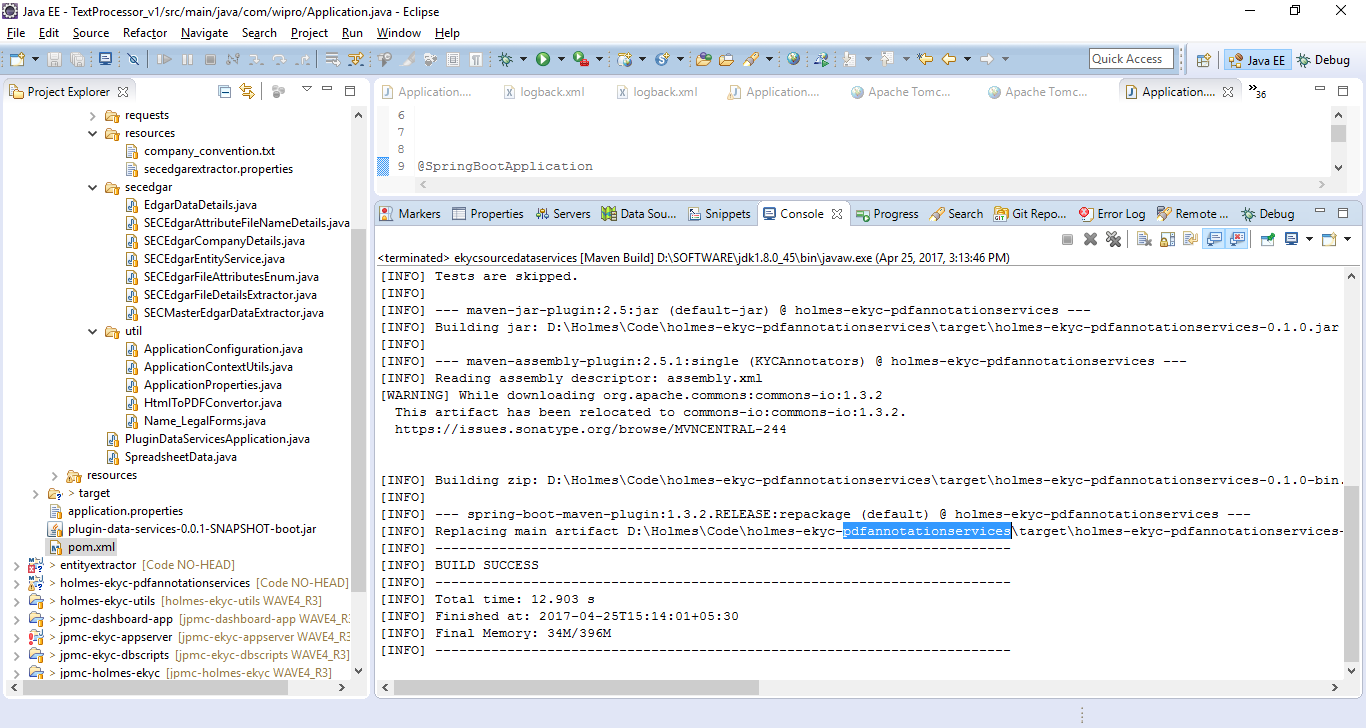


Once done Successfully build you will get below screen shots,

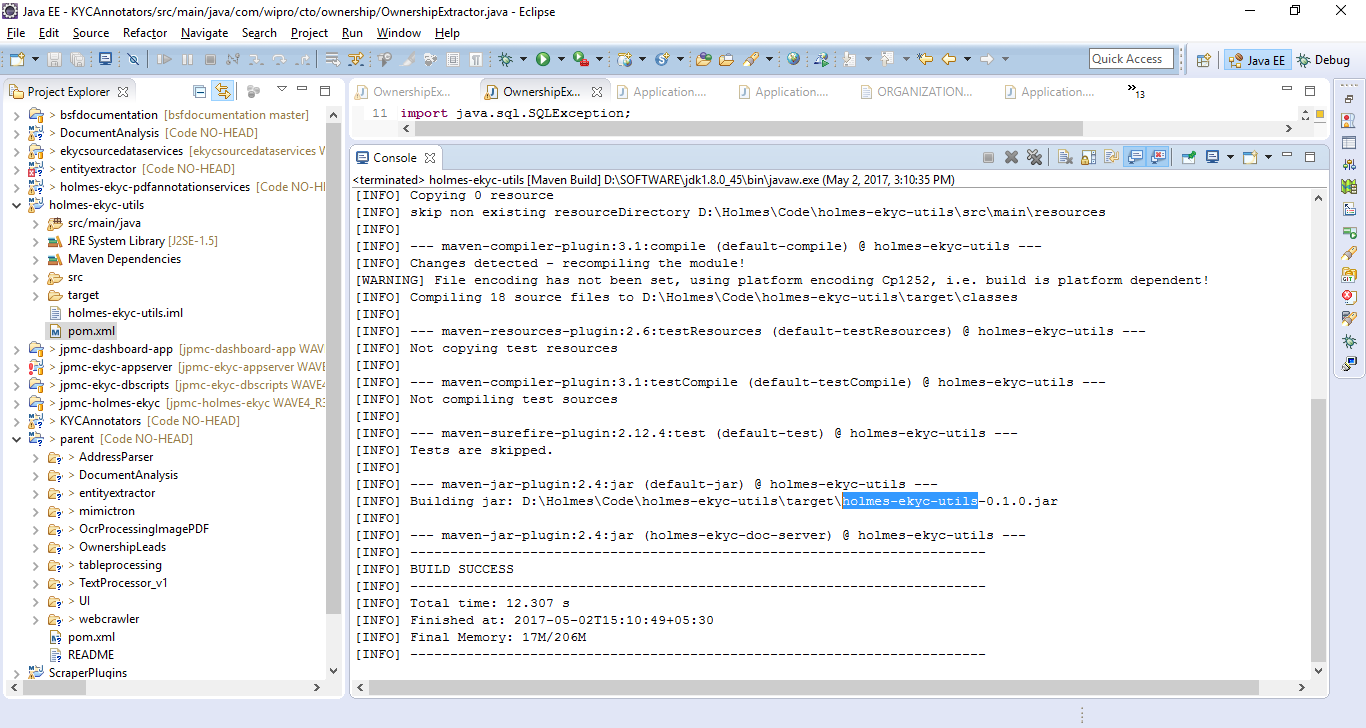
1. **cognitive\_ekyc\_features**



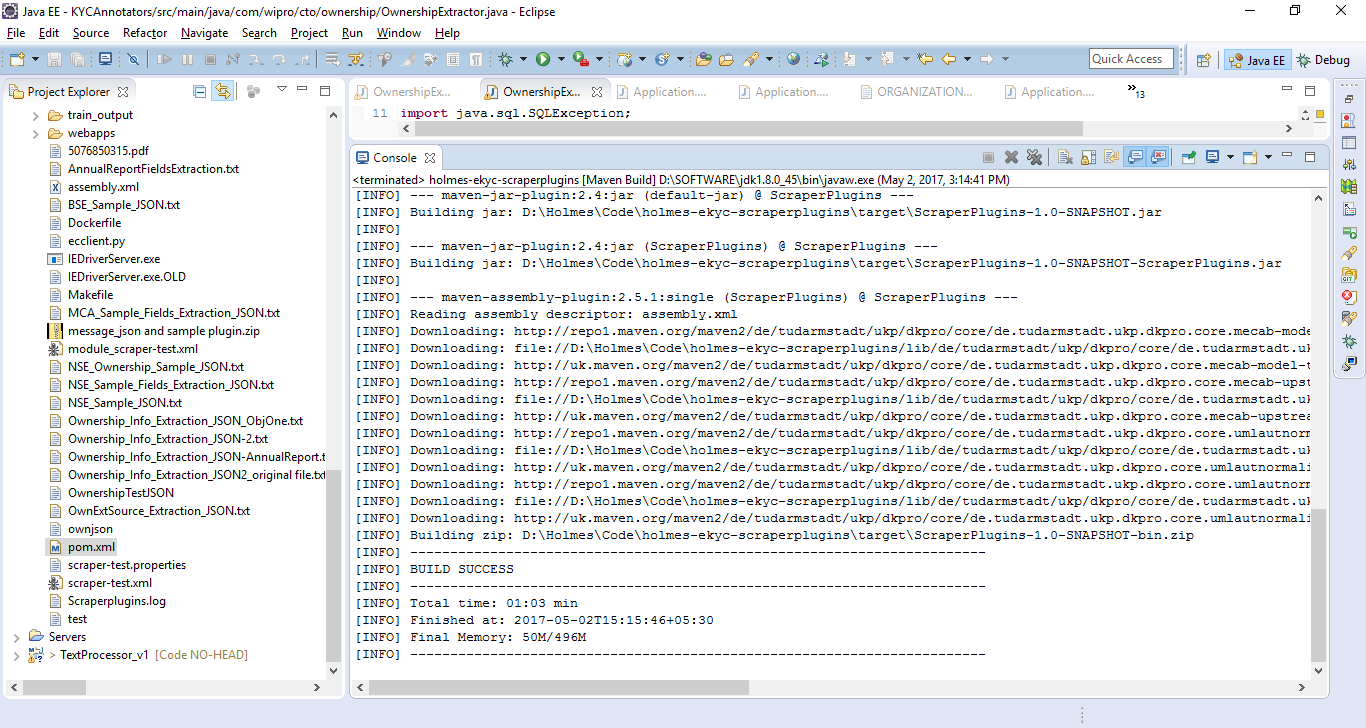
1. **holmes-ekyc-pdfannotationservices**



1. **holmes-ekyc-utils**



1. **holmes-ekyc-scraperplugins**



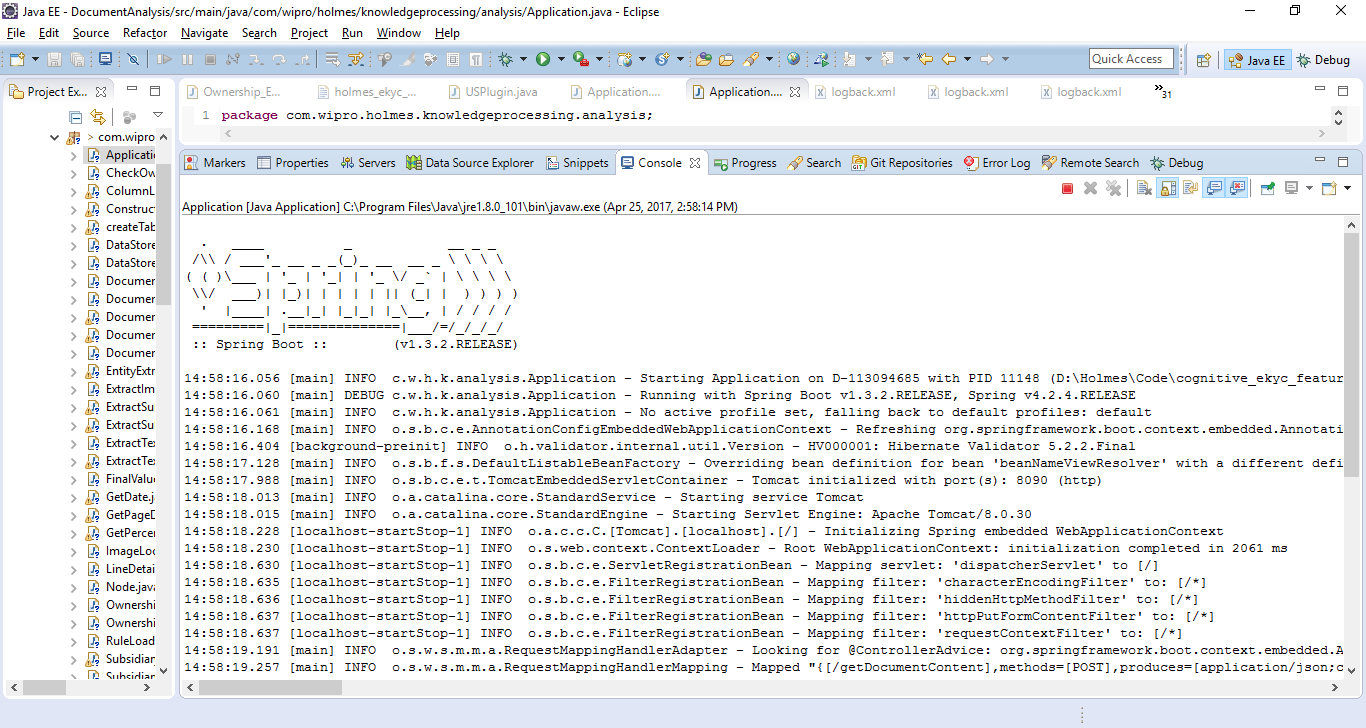
**Running the required Applications: -**

Below are services used in Scraper Plugin. Before running the Scraper plugin, we need to run those all services or application.

1. **DocumentAnalysis**

To run this application, open the file below and **Run as** -- > **Java Application**

DocumentAnalysis/src/main/java/com/wipro/holmes/knowledgeprocessing/analysis/**Application.java**



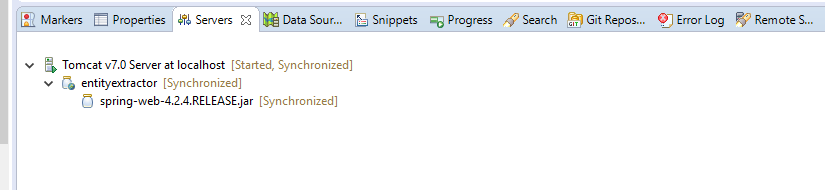
1. **Entityextractor**

To run this application, select the project and **Run As** -- > **Run on Server**

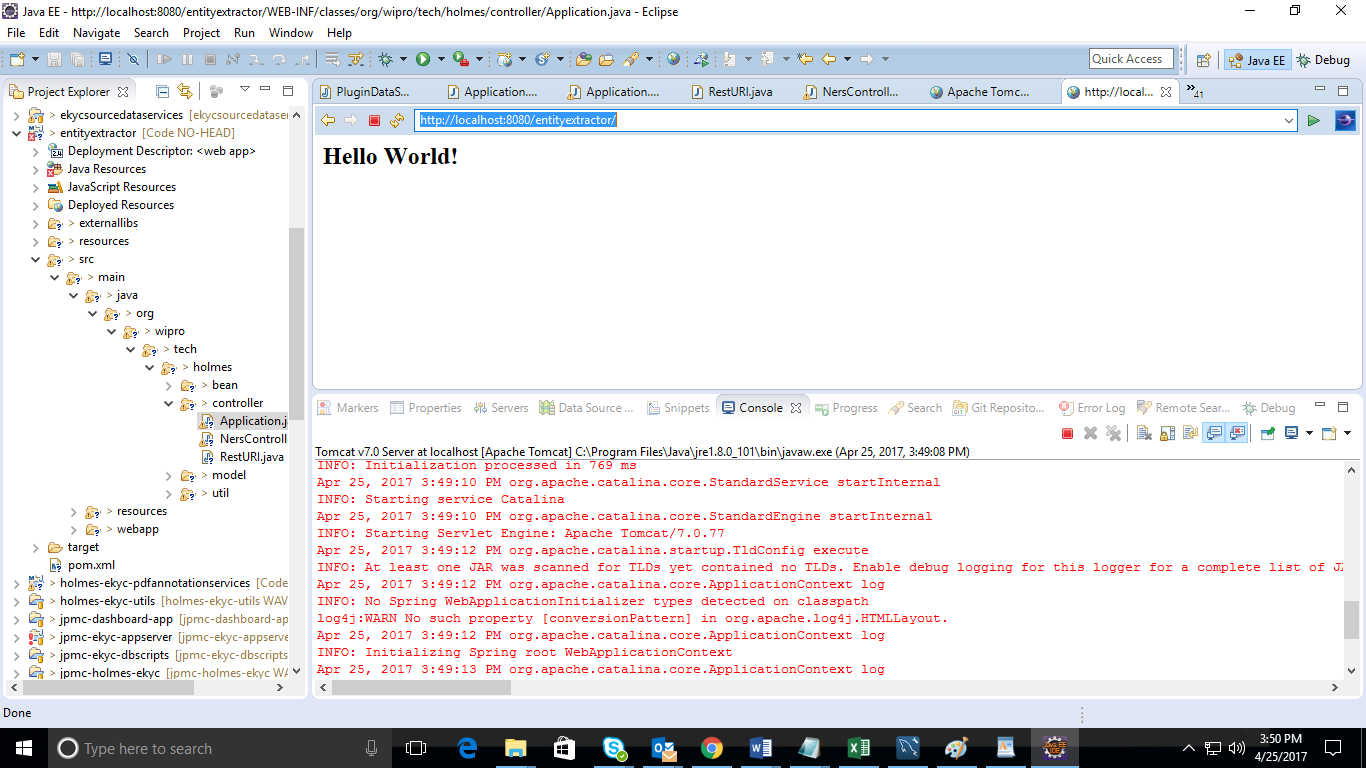
For this you need to set up server ‘**Tomcat v7.0 Server’** and set the port number as 8080 in the properties file below,

/ScraperPlugins/conf/holmes\_ekyc\_config.properties

<http://localhost:8080/entityextractor/services/entities>



Once it will run on server you will get the below screen,

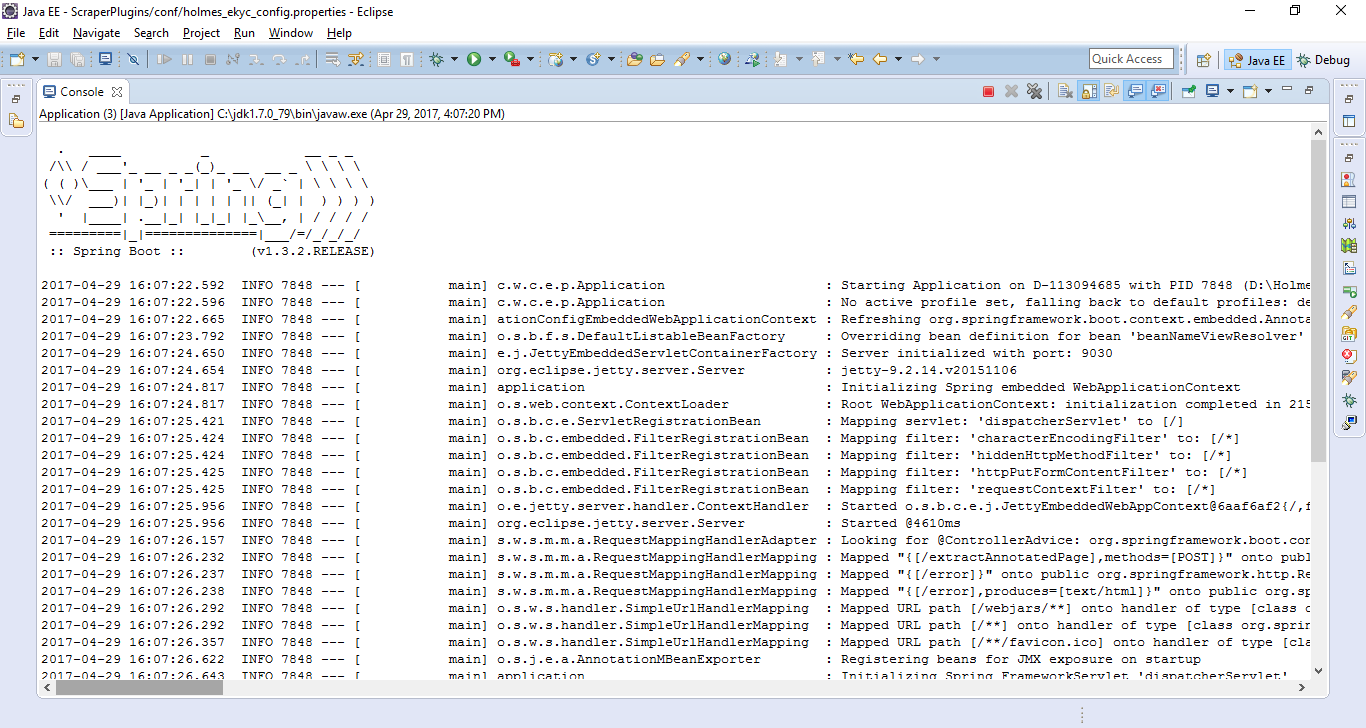


1. **holmes-ekyc-pdfannotationservices**

To run this application, select the file below and **Run As** -- > **Java Application**,

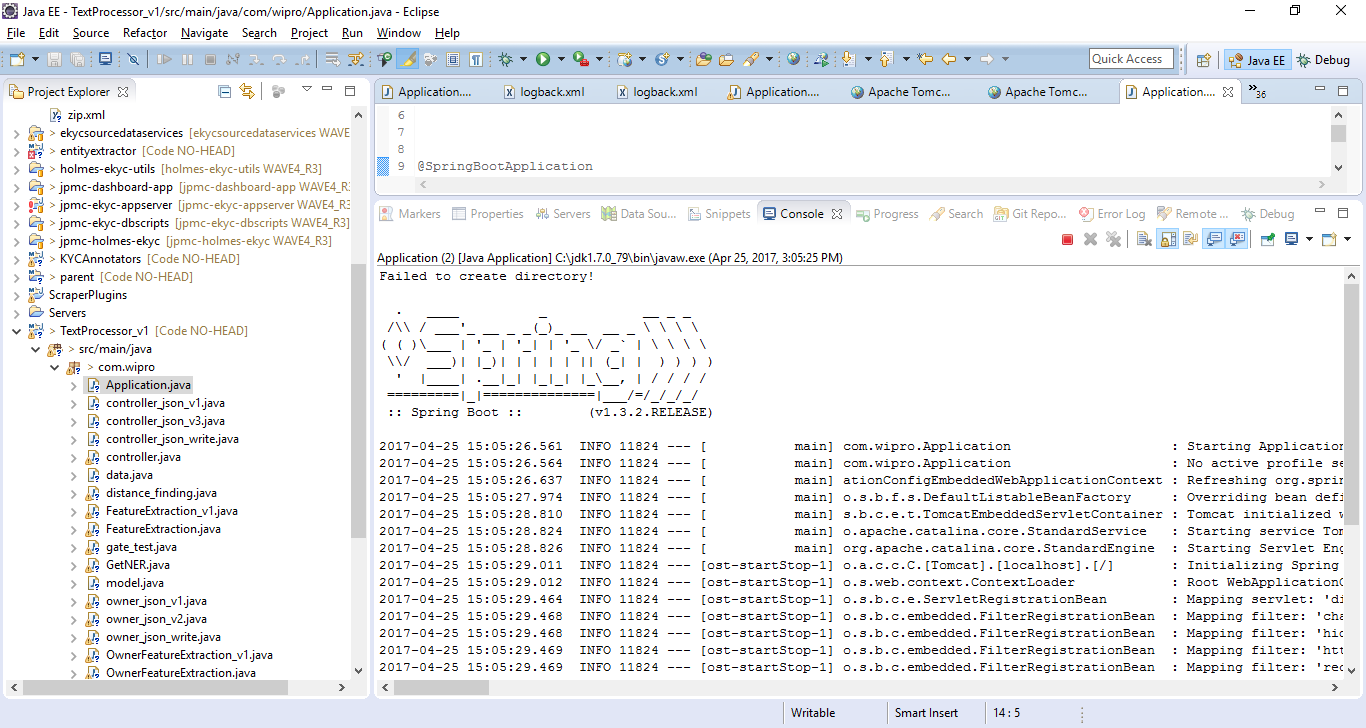
holmes-ekyc-pdfannotationservices/src/main/java/com/wipro/cto/ekyc/pdfannotationservices/**Application.java**

Once done you will get below screen shot,



1. **TextProcessor\_v1**

To run this application, select the below file and **Run As** -- > **Java Application,** Once done you will get below screen shot,



**Required Code changes to run the Scraper Plugin:-**

Below are the files where you required to change

1. /ScraperPlugins/conf/**holmes\_ekyc\_config.properties**

**Changes**: - Please make sure mentioned below folder will be available on mentioned path before running the scraper plugin ,

DB\_URL=jdbc:mysql://10.113.173.135:3306/holmes\_kyc\_config?user=ekycuser&password=ekycuser

USERDB\_URL=jdbc:mysql://10.113.173.135:3306/kyc\_user\_schema?user=ekycuser&password=ekycuser

entityExtractor=http://localhost:8080/entityextractor/services/entities

jsonLogPath=D:/TestDoc

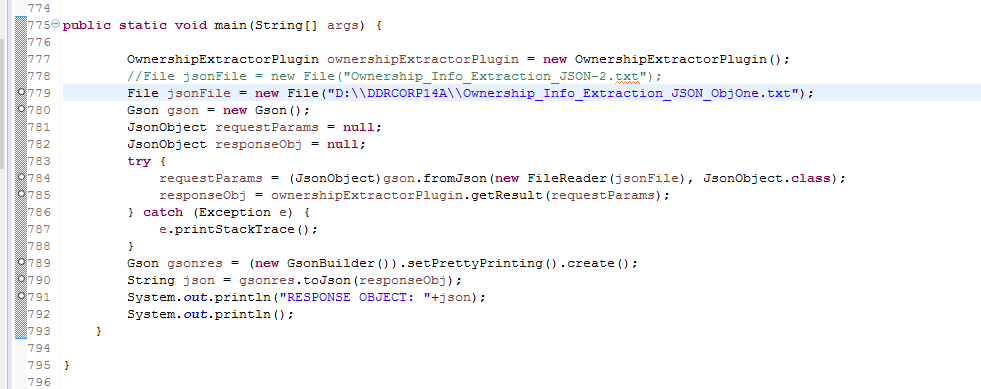
annualReportPath=D:/TestDoc

SURNAMES\_DIRECTORY=D:/resources/Surnames

1. /ScraperPlugins/src/main/java/plugin/OWNERSHIP\_EXTRACTOR/**OwnershipExtractorPlugin.java**

**Changes: -**

This is the main class from where Scraper Plugin is running,



Here I have mentioned the input JSON file with it’s path. Please set accordingly.

Input JSON File Ownership\_Info\_Extraction\_JSON\_ObjOne.txt shown below, put it inside

D:\DDRCORP14A folder,



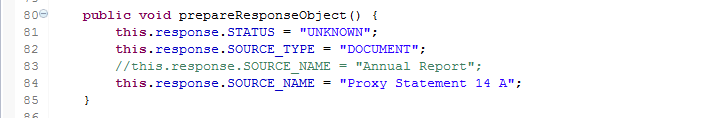
Inside the JSON input file I have mentioned the path for DOCUMENT\_LINK as D:\\DDRCORP14A\\DDR-CORP-DEF-14A.pdf

Please make sure this file will be present at mentioned path.

Set the variable as **true** for **pdfProxyFlowEnabled,**

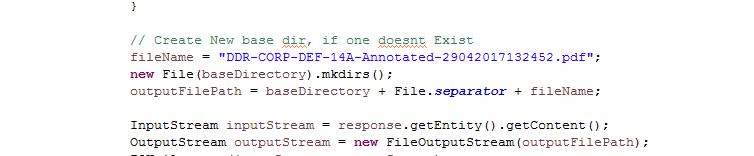


Set **SOURCE\_NAME** as ‘Proxy Statement 14 A’,

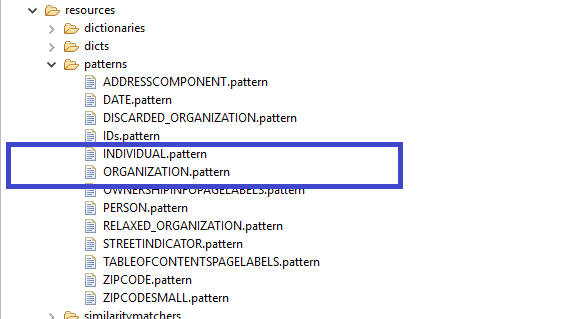


1. /holmes-ekyc-utils/src/main/java/com/wipro/cto/**PDFAnnotationServiceUtil.java**

Here I have mentioned the file name as below, you can change as per the name of the file generated.



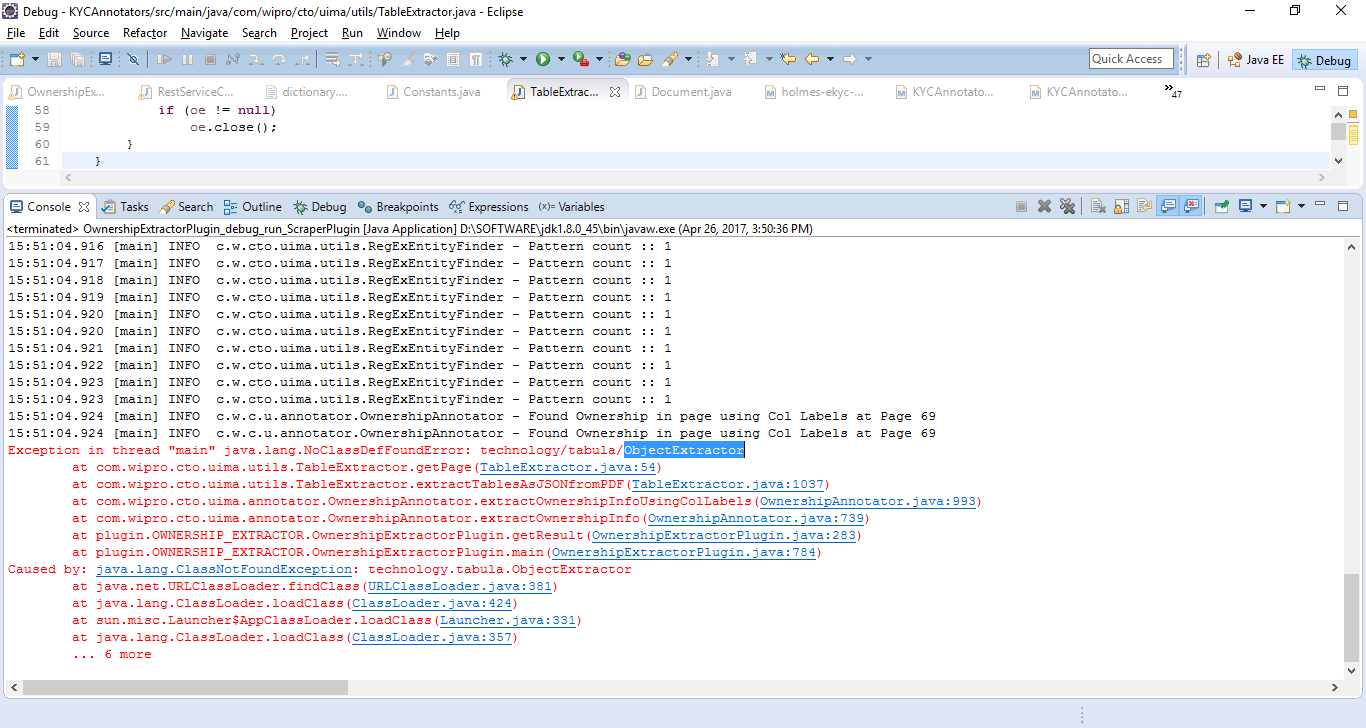
1. Please make sure that patterns files(INDIVIDUAL.pattern and ORGANIZATION.pattern) will be available at path “/ScraperPlugins/resources/patterns”



**TRAILS & ERRORS: -**

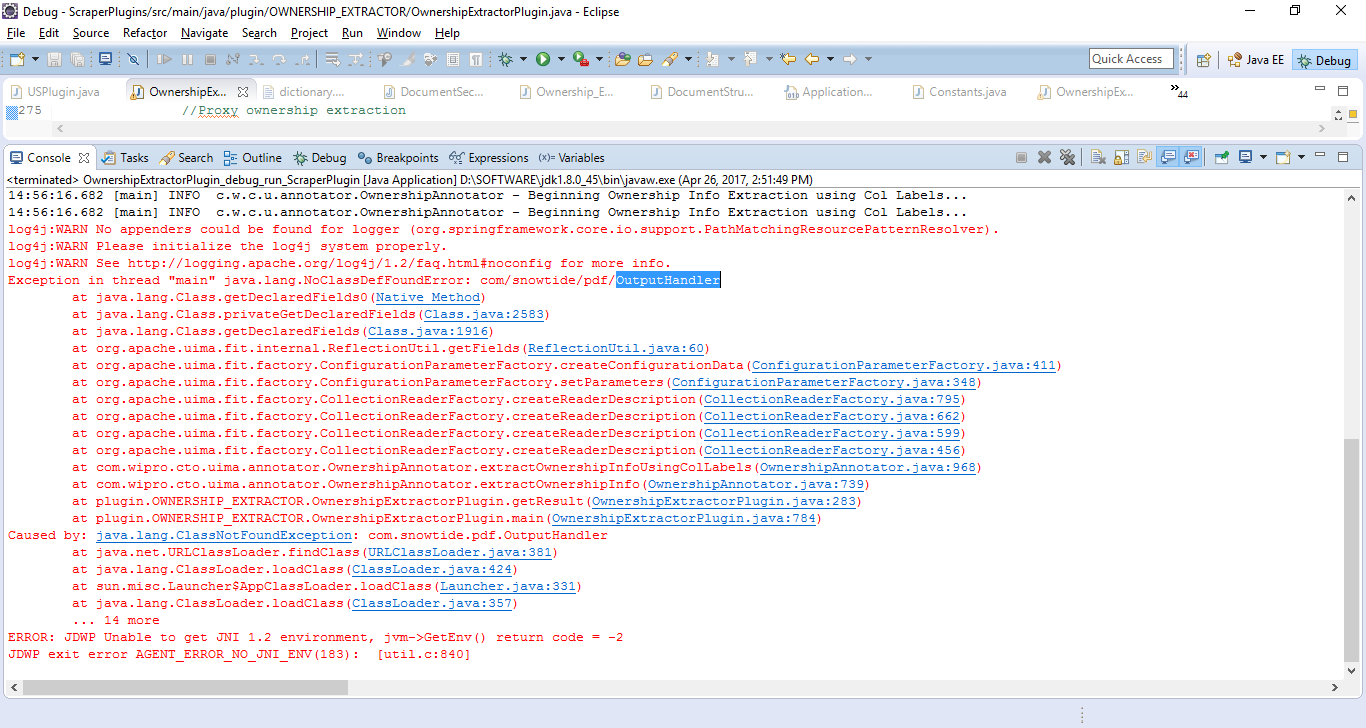
You may get some errors related to missing jar files,

1. tabula jar file,



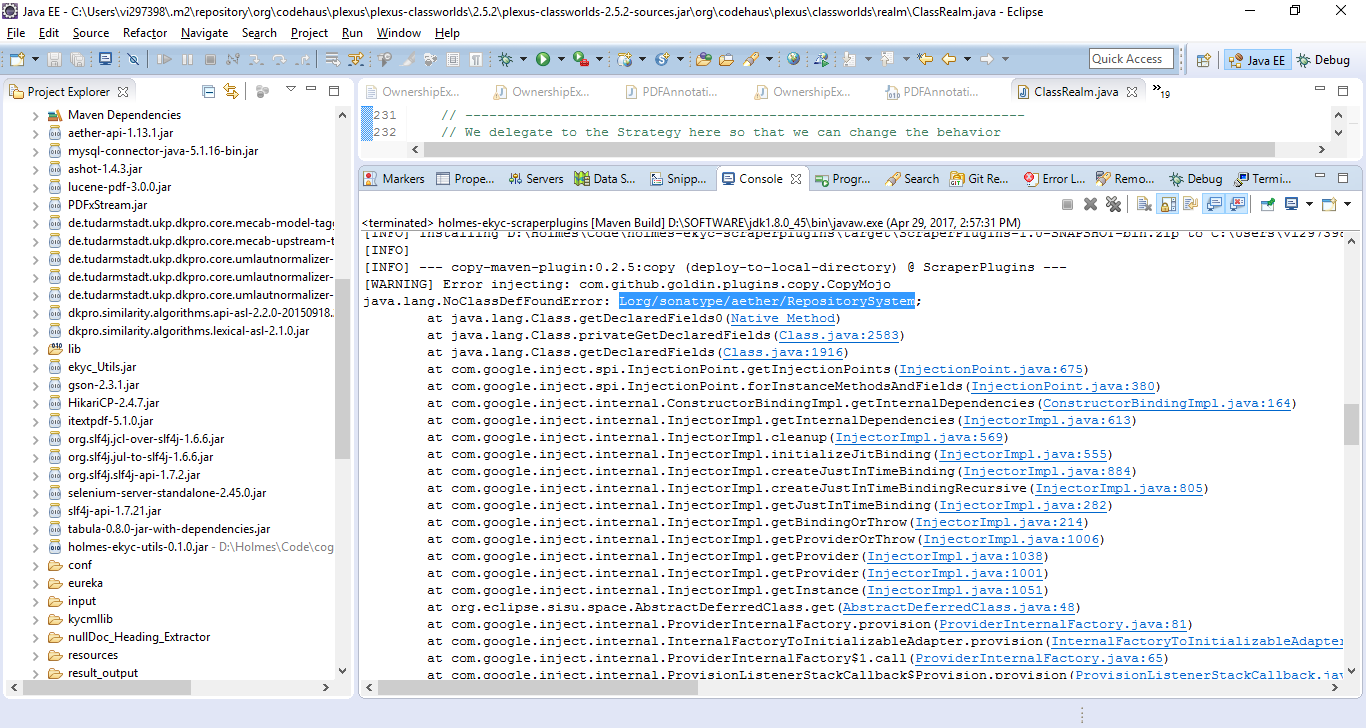
To Resolve this issue download tabula-0.8.0-jar-with-dependencies.jar jar file add it to Scraper Plugin build path.

1. OutputHandler Error



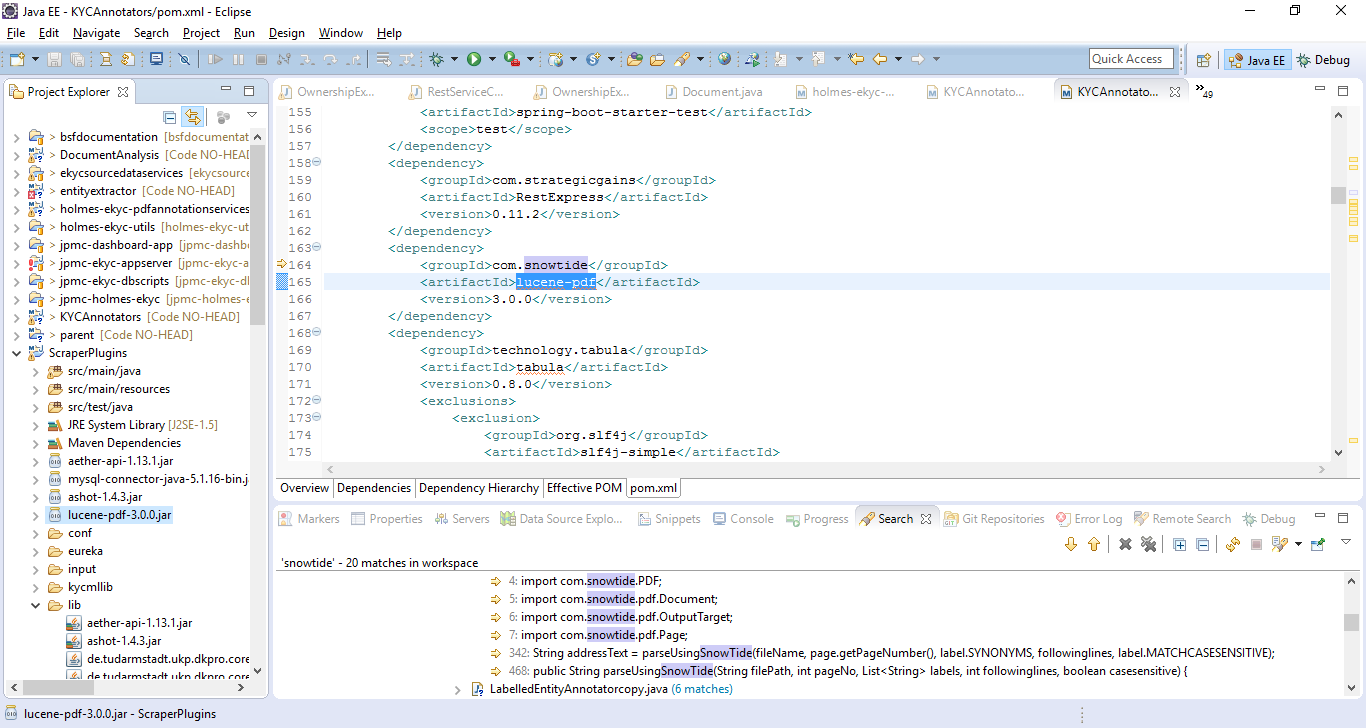
To Resolve this download the jar file Lucene-pdf-3.0.0.jar and add it to Scraper Plugin build path.

1. Add PDFxStream.jar to Scraper Plugin build path,
2. RepositorySystem errors,

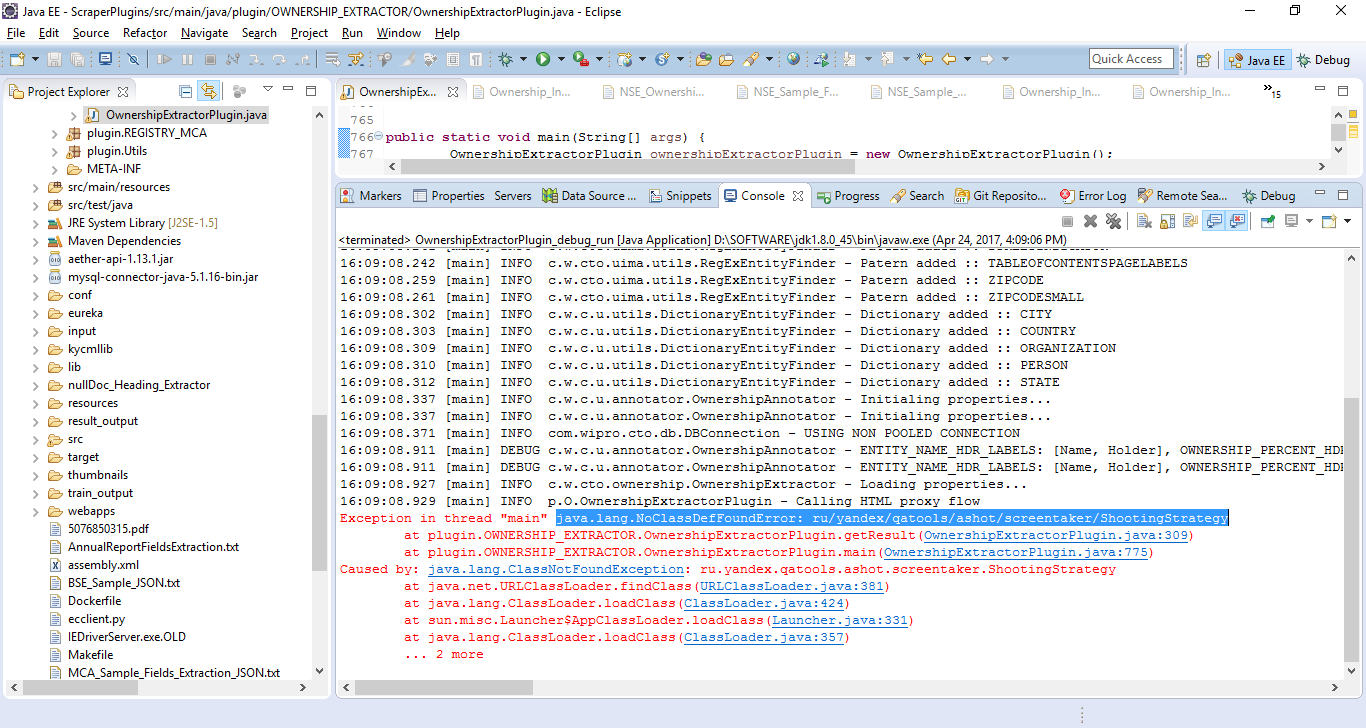


To Resolve this issue download and add jar file aether-api-1.13.1.jar to Scraper Plugin build path.

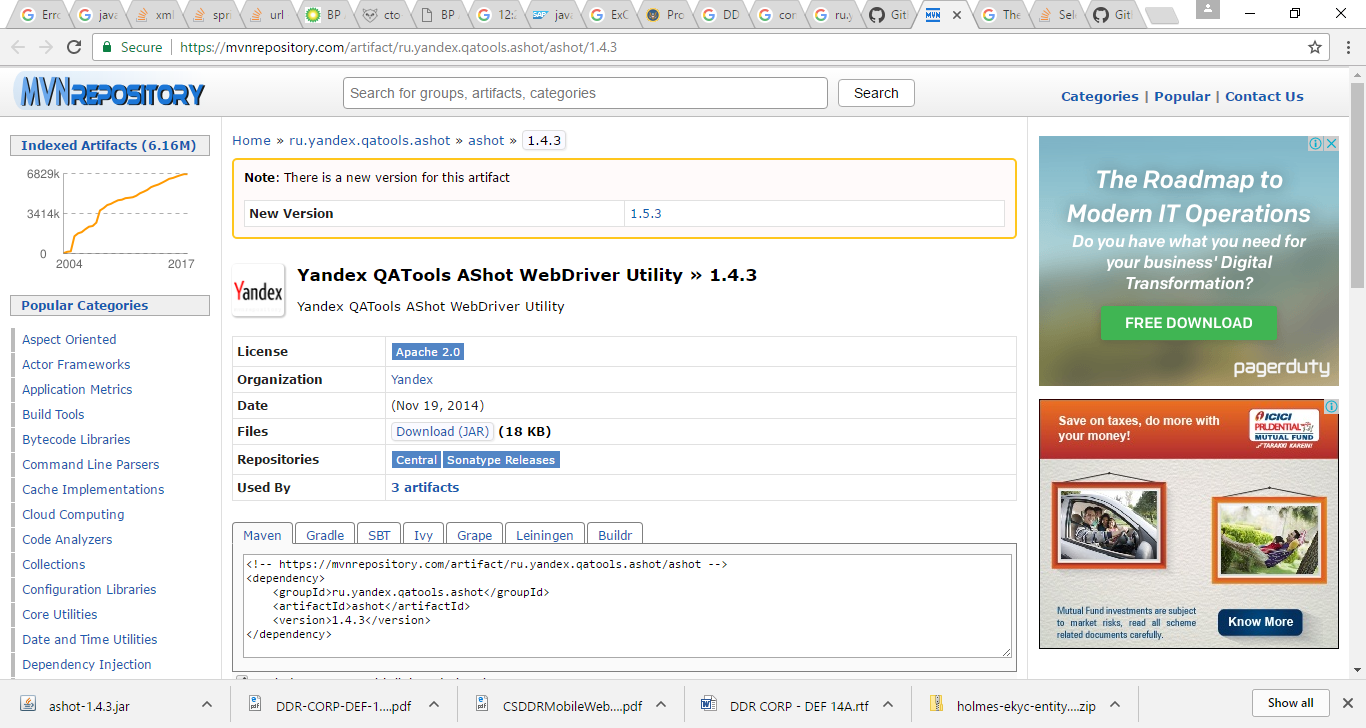




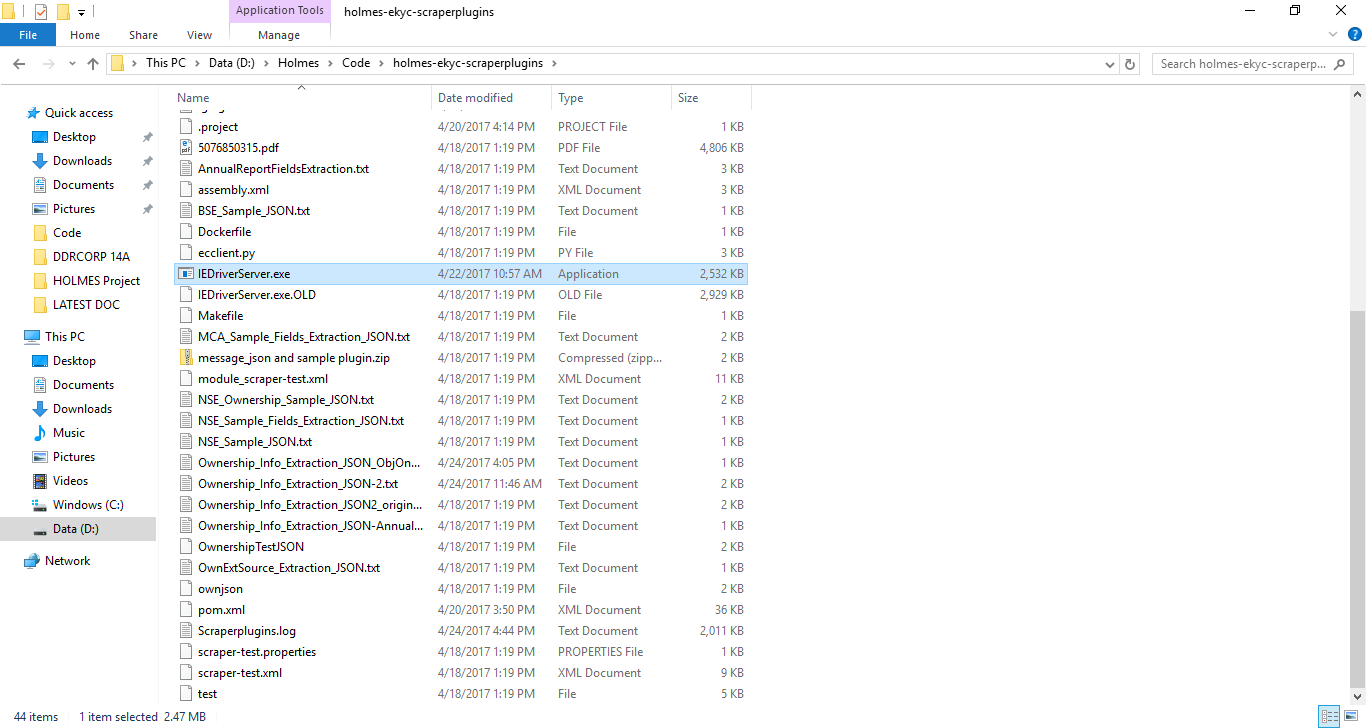
1. Add jar file holmes-ekyc-utils-0.1.0.jar to build path for scraper plugin which is a generated jar file from the project holmes-ekyc-utils.
2. Yandex jar file related error,



To resolve this issue download jar file as shown in below screen shot,



1. If you get IE Driver related issue then download jar file for IEDriverServer.exe add it to Scraper plugin build path,



---------------------------------------------------------------------End------------------------------------------------------