1. Установка Eclipse и плагина текстовой аналитики

Процедура скачивания и установки среды разработки Eclipse и плагина текстовой аналитики описана в:

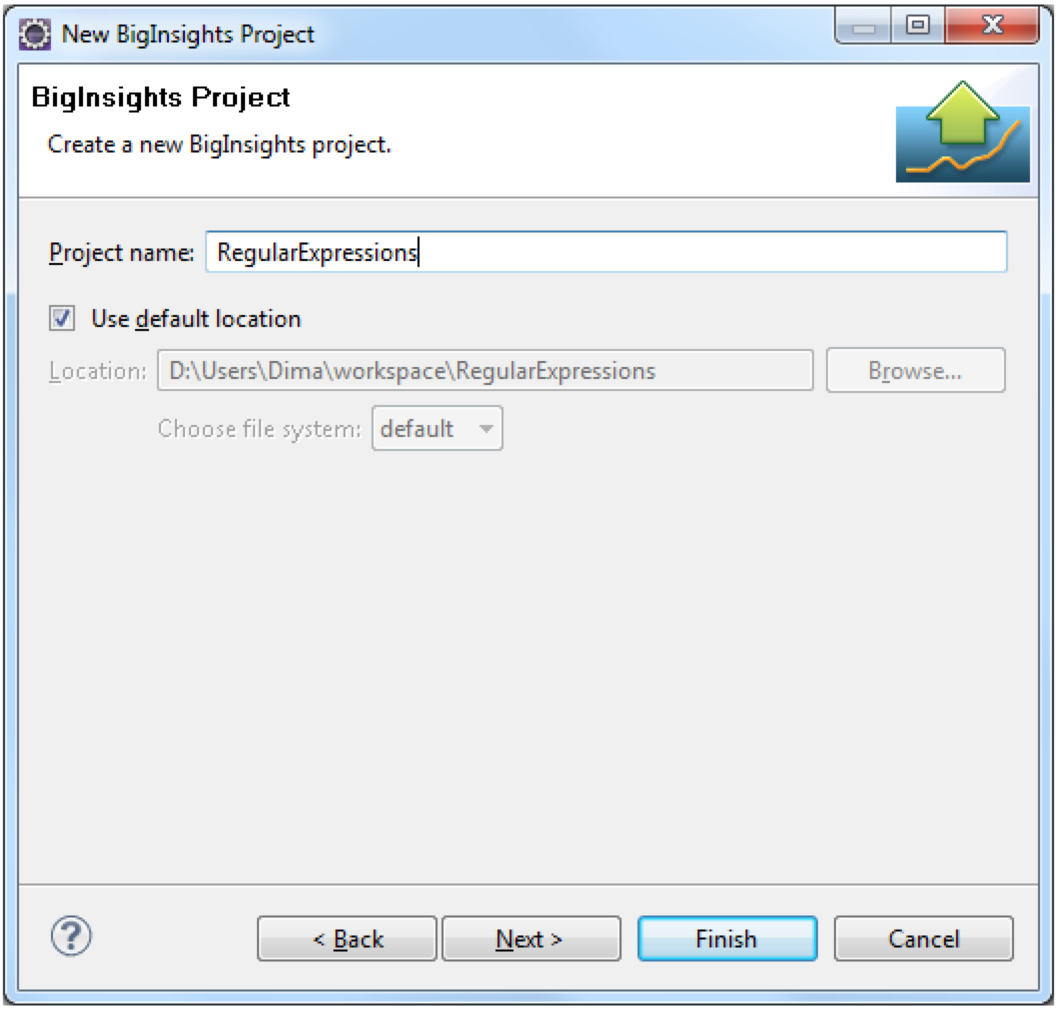
http://synthesis.ipi.ac.ru/synthesis/student/BigData/seminar-hadoop/lectures2014/06_MapReduce_Lab.pdf

В качестве сервера будут использоваться те же самые данные что и прошлый раз:

2. Создание проекта

Выбрать: File -> New -> BigInsight Project

Ввести имя проекта.

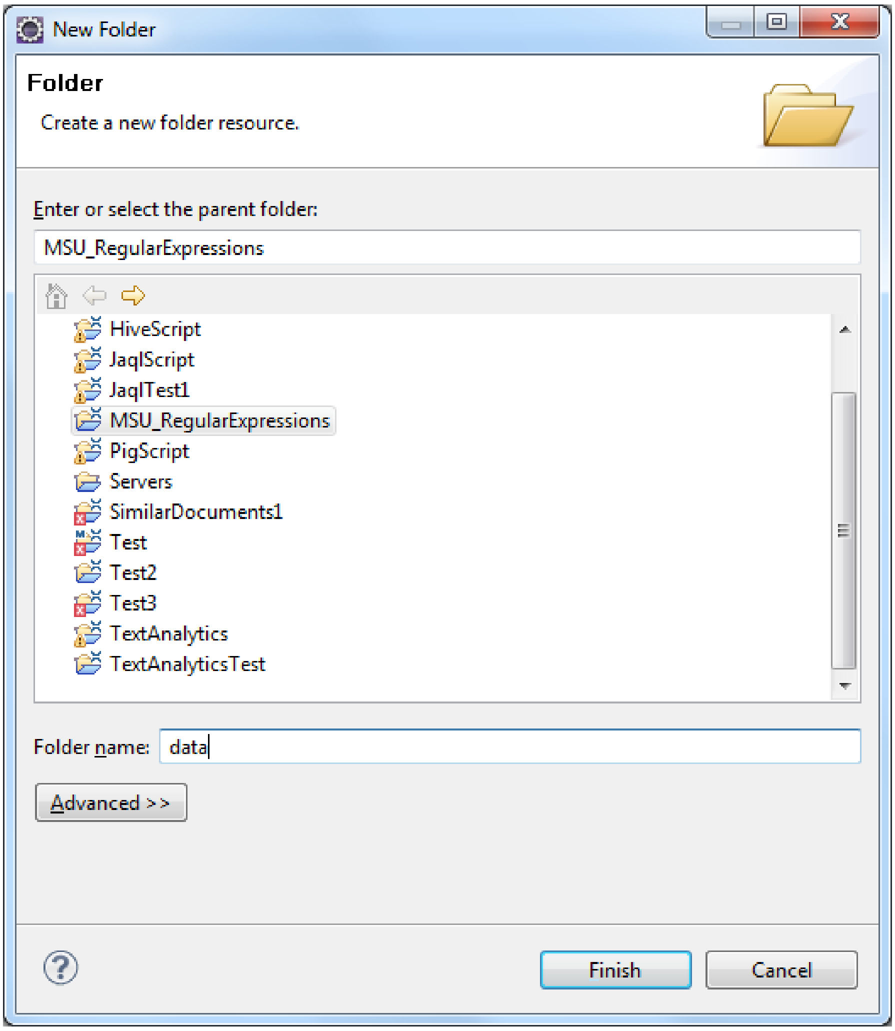


Создание директории data

Кликнуть правой кнопкой мыши на название проекта в закладке Project Explorer.

Выбрать New -> Folder

Ввести имя директории data.



Импорт данных

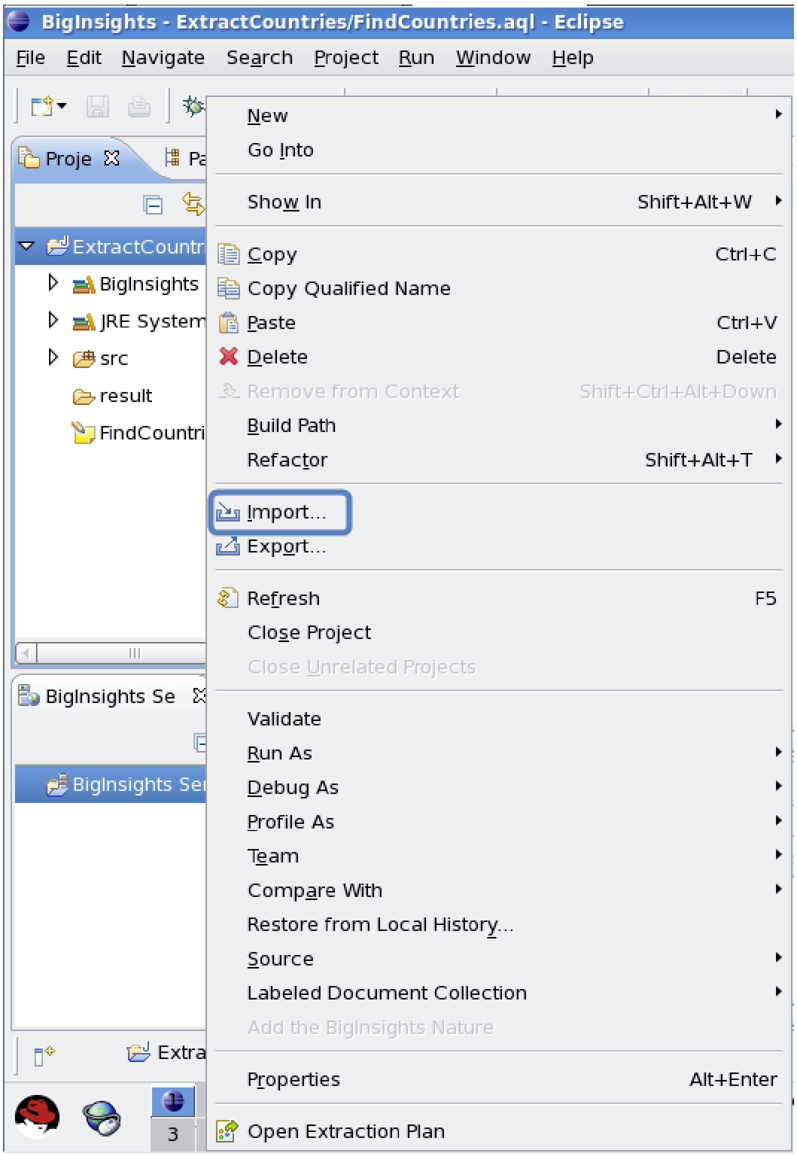
Скачать файл Facts.txt http://synthesis.ipi.ac.ru/synthesis/student/BigData/lectures-ie/Facts.txt

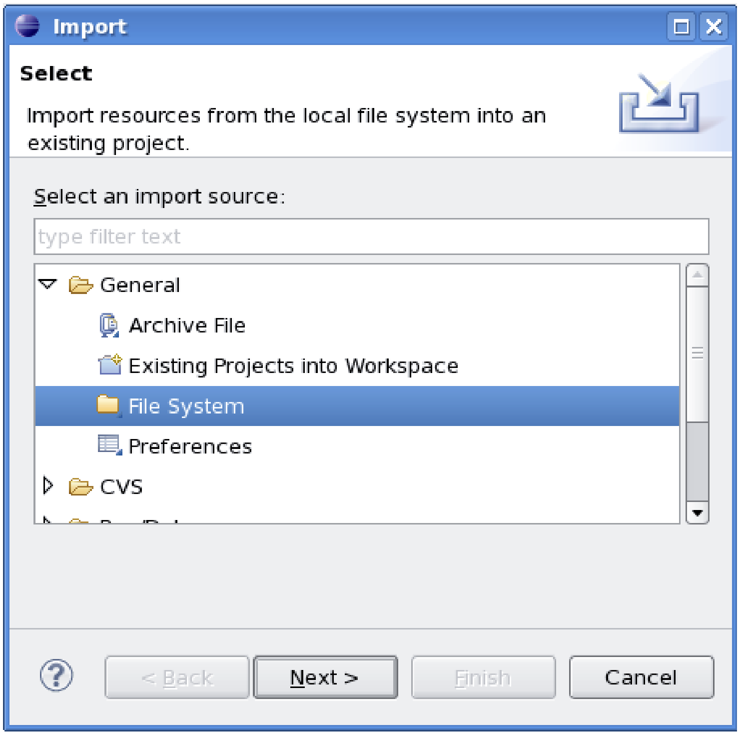
Кликнуть правой кнопкой мыши на директорию data в закладке Project Explorer.

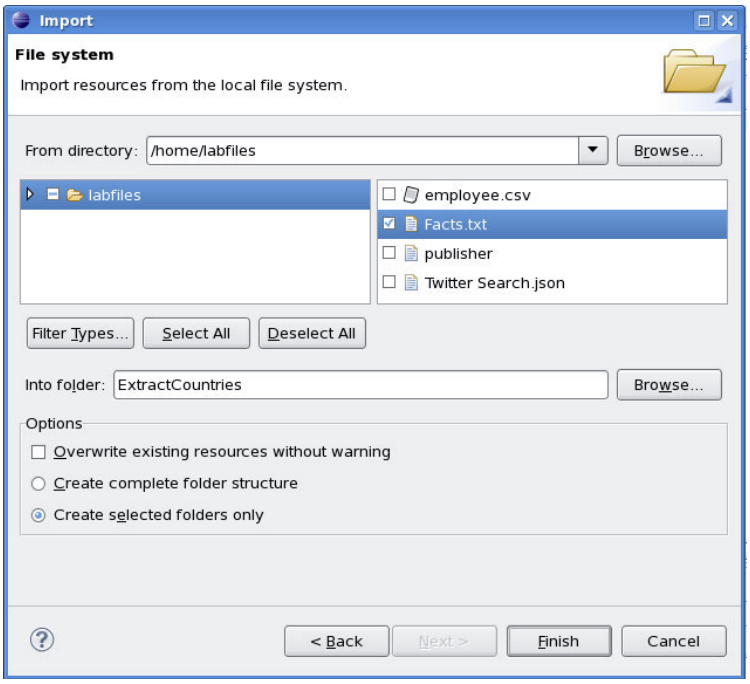
Выбрать Import

Выбрать General -> File System

Указать директорию, где сохранили файл, Выбрать этот файл и нажать Finish.


BigInsights - ExtractCountries/FindCountries.aql - Eclipse
File Edit Navigate Search Project Run Window Help
Proje
ExtractCountr
BigInsights
JRE System
src
result
FindCountri
BigInsights Se
BigInsights Se
Extra
New
Go Into
Show In Shift+Alt+W
Copy Ctrl+C
Copy Qualified Name
Paste Ctrl+V
Delete Delete
Remove from Context Shift+Ctrl+Alt+Down
Build Path
Refactor Shift+Alt+T
Import...
Export...
Refresh F5
Close Project
Close Unrelated Projects
Validate
Run As
Debug As
Profile As
Team
Compare With
Restore from Local History...
Source
Labeled Document Collection
Add the BigInsights Nature
Properties Alt+Enter
Open Extraction Plan

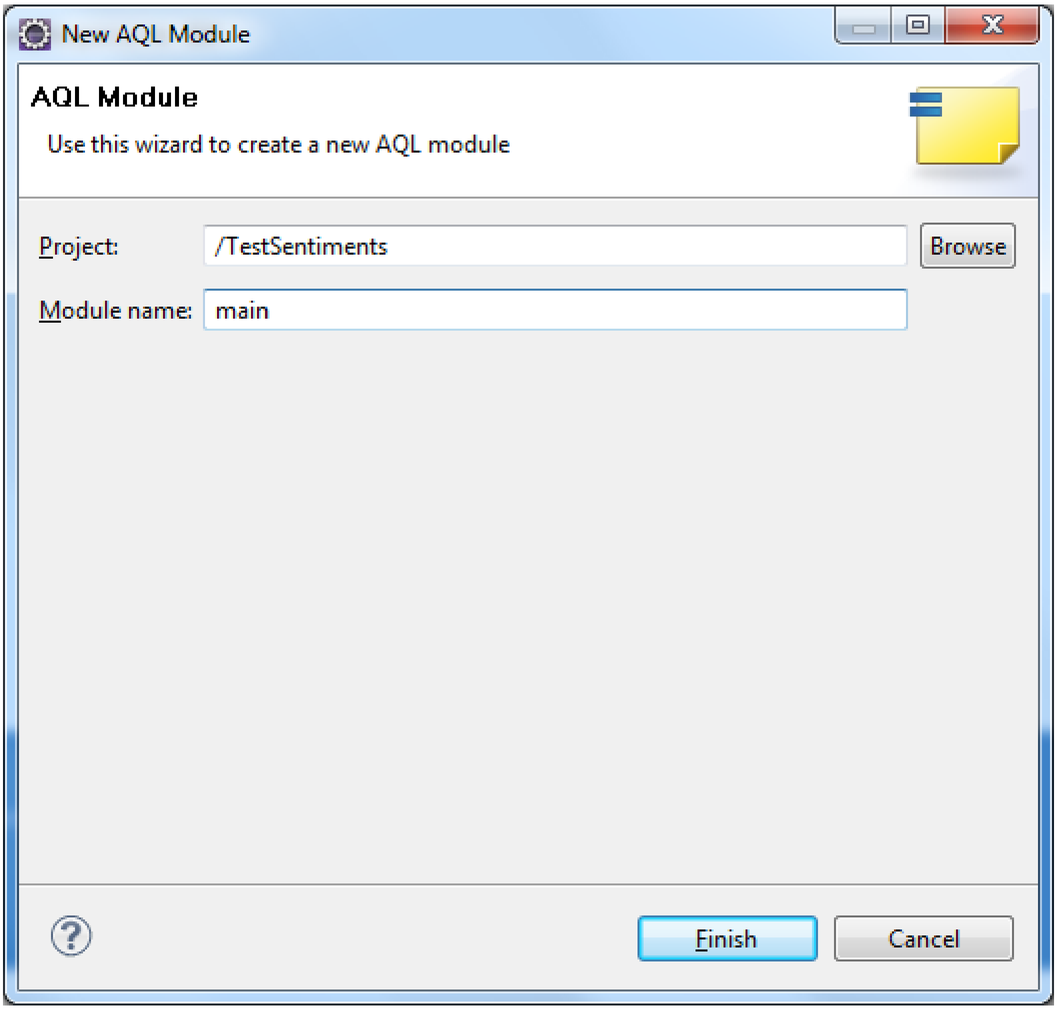

Import
Select
Import resources from the local file system into an existing project.
Select an import source:
type filter text
General
Archive File
Existing Projects into Workspace
File System
Preferences
CVS
< Back
Next >
Finish
Cancel



Создание AQL файла

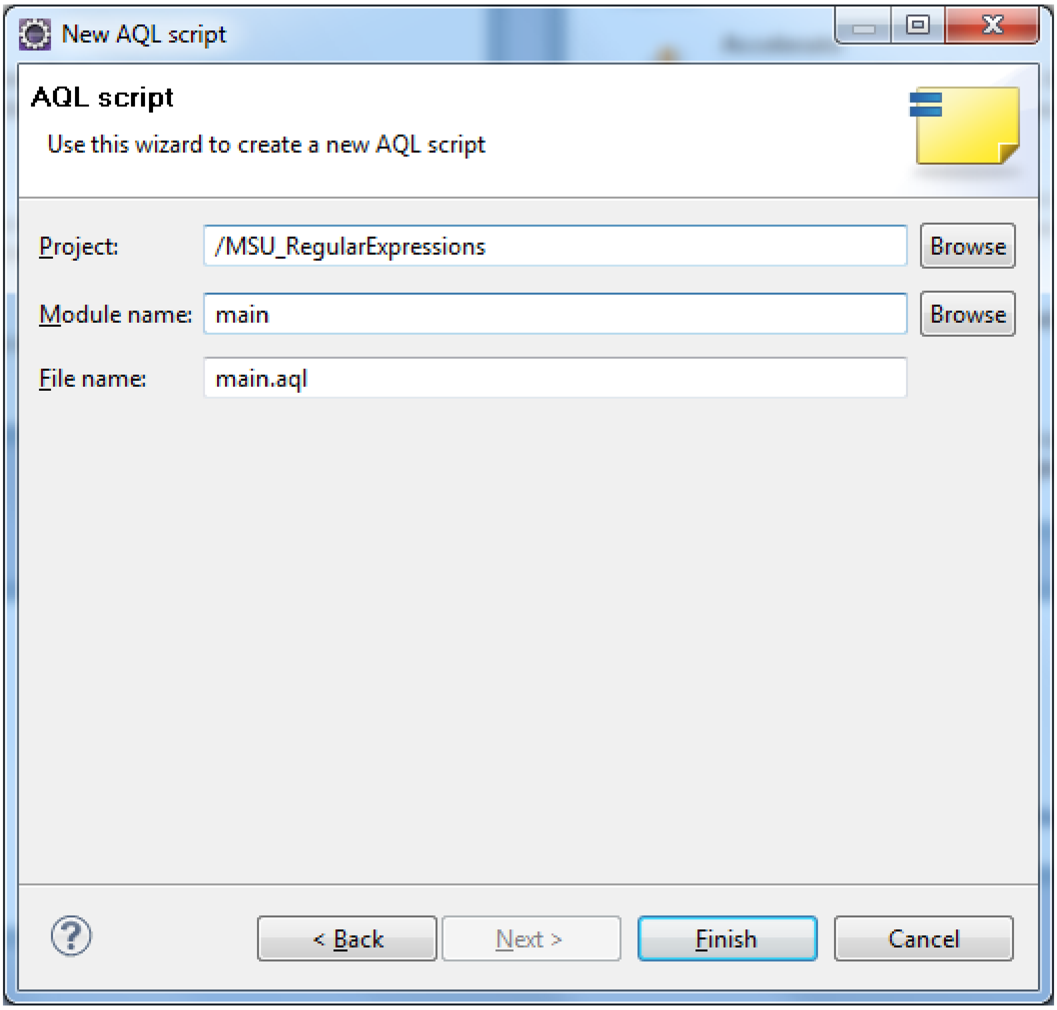
Создать AQL модуль main

- Выбрать: File -> New -> AQL Module
- Ввести имя модуля



Создать AQL файл main.aql

- Выбрать: File -> New -> AQL Script
- Выбрать модуль main
- Ввести имя файла main.aql



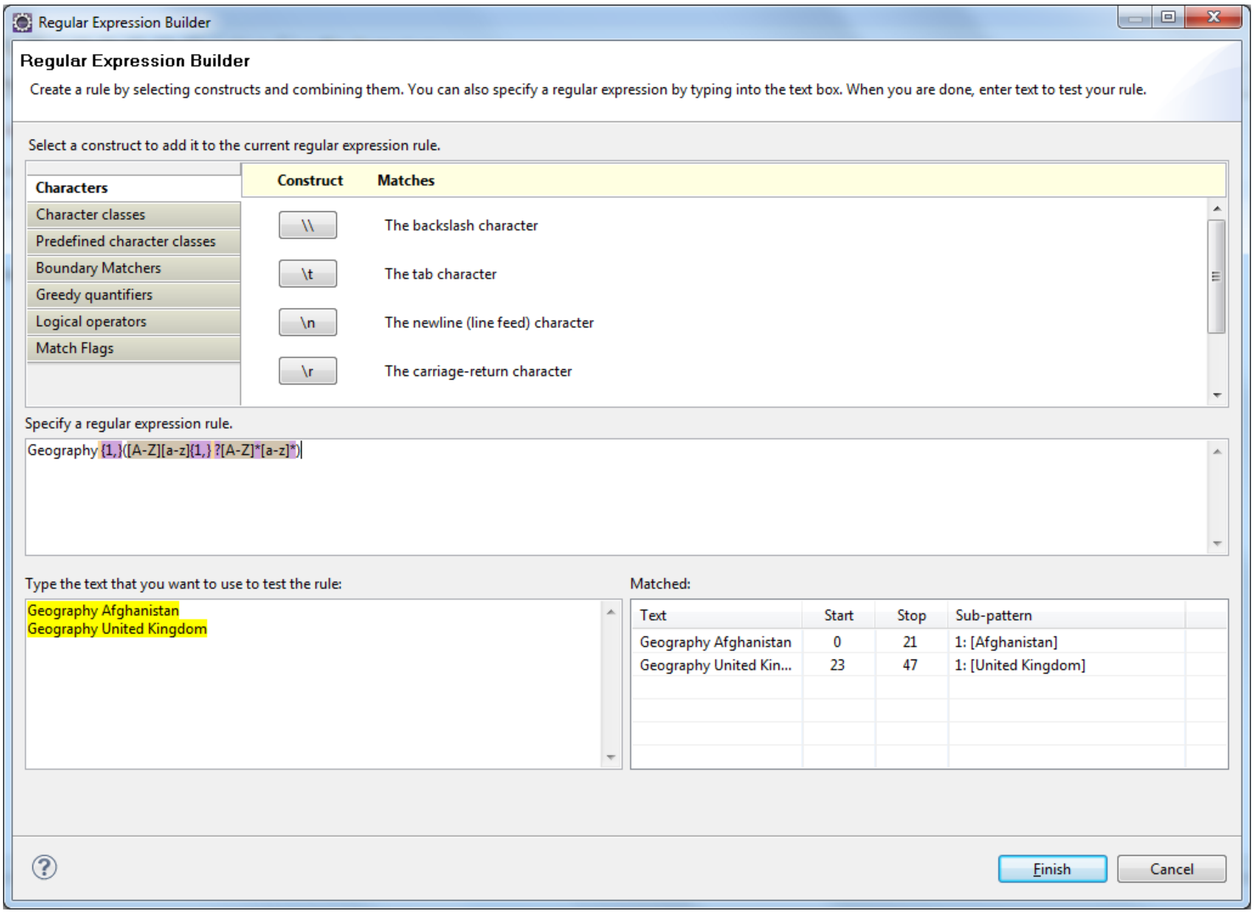
Создание представления Countries

Открыть файл main.aql

Ввести:

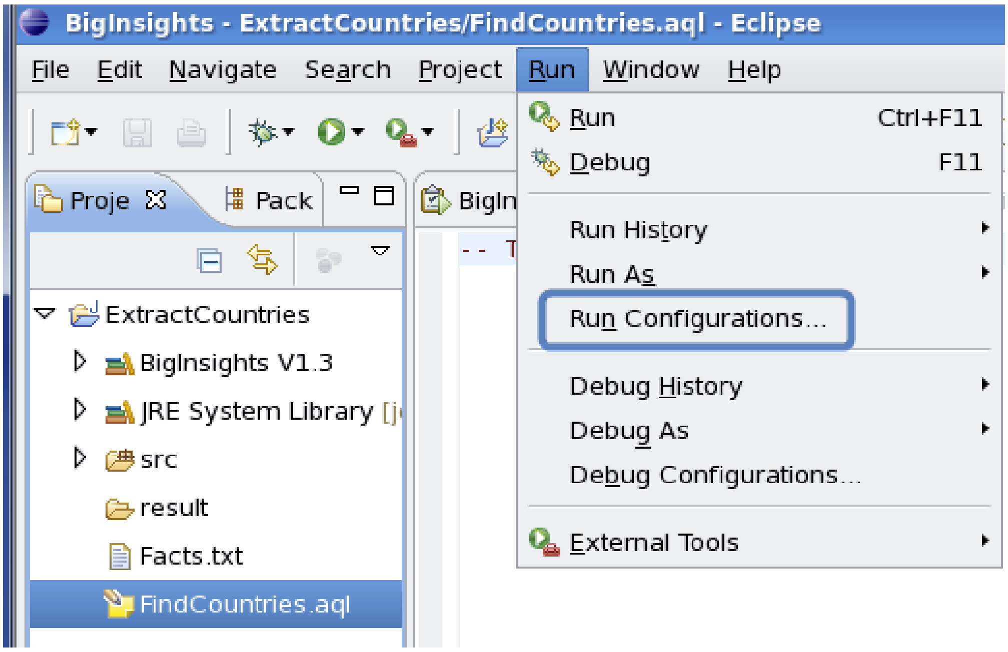
```
module main;
create view Countries as
extract regex /Geography {1,}([A-Z][a-z]{1,} ?[A-Z]*[a-z]*)/ on D.text
return group 1 as country
from Document D;
output view Countries;
```

Для редактирования регулярных выражений можно воспользоваться конструктором регулярных выражений:

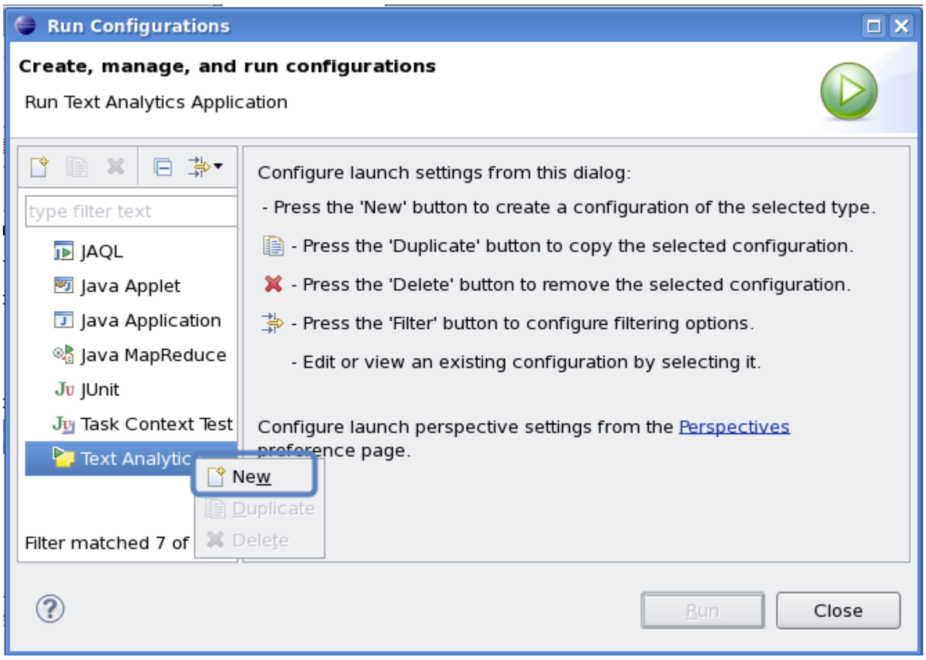


Конфигурация выполнения

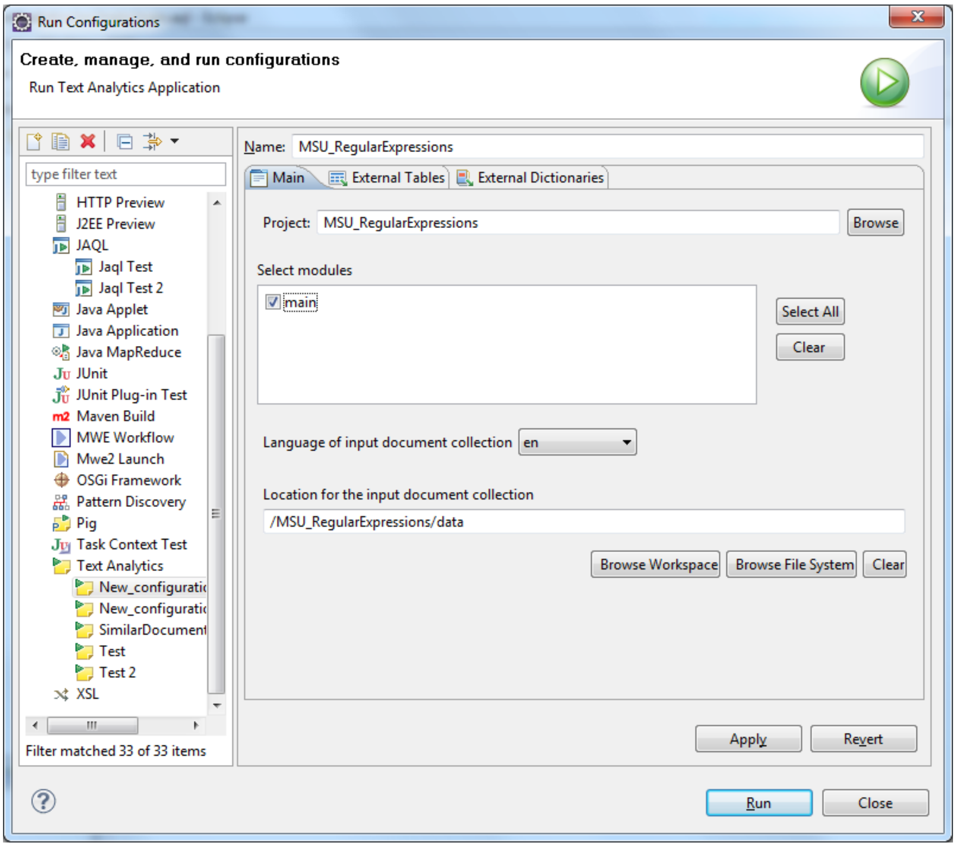
Выбрать меню Run -> Run Configurations



Кликнуть правой кнопкой мыши на Text Analytics и выбрать New



Выбрать проект, AQL модуль main, указать директорию с файлом данных Facts.txt

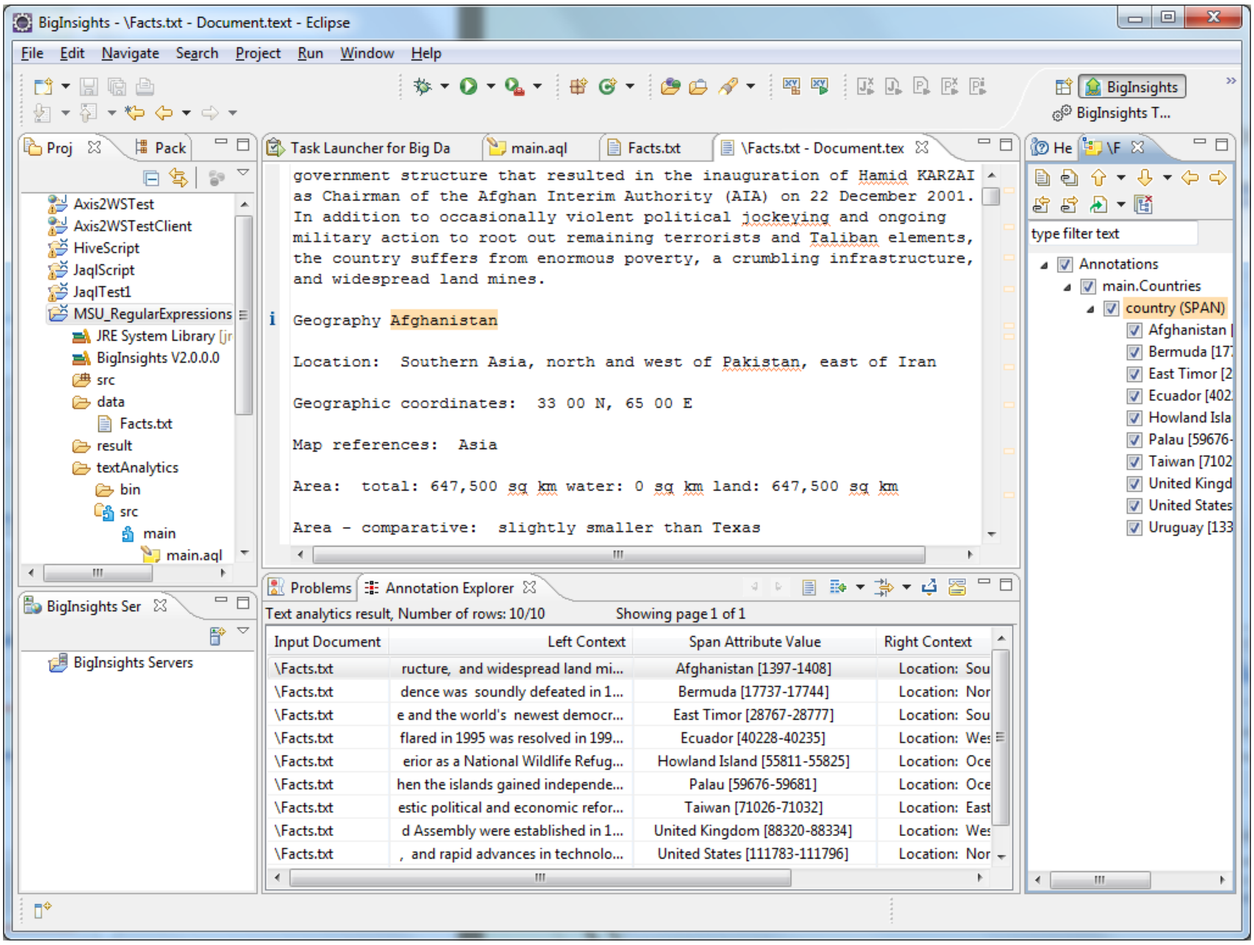


Нажать Apply

Результат выполнения экстракторов

В конфигурации выполнения нажать Run

После выполнения откроется страница с результатами:


BigInsights - \Facts.txt - Document.text - Eclipse
File Edit Navigate Search Project Run Window Help
BigInsights
BigInsights T...
Proj
Pack
Axis2WSTest
Axis2WSTestClient
HiveScript
JaqlScript
JaqlTest1
MSU_RegularExpressions
JRE System Library [jr
BigInsights V2.0.0.0
src
data
Facts.txt
result
textAnalytics
bin
src
main
main.aql
BigInsights Ser
BigInsights Servers
Task Launcher for Big Da
main.aql
Facts.txt
\Facts.txt - Document.tex
government structure that resulted in the inauguration of Hamid KARZAI
as Chairman of the Afghan Interim Authority (AIA) on 22 December 2001.
In addition to occasionally violent political jockeying and ongoing
military action to root out remaining terrorists and Taliban elements,
the country suffers from enormous poverty, a crumbling infrastructure,
and widespread land mines.
Geography Afghanistan
Location: Southern Asia, north and west of Pakistan, east of Iran
Geographic coordinates: 33 00 N, 65 00 E
Map references: Asia
Area: total: 647,500 sq km water: 0 sq km land: 647,500 sq km
Area - comparative: slightly smaller than Texas
He
\F
type filter text
Annotations
main.Countries
country (SPAN)
Afghanistan [
Bermuda [17
East Timor [2
Ecuador [402
Howland Isla
Palau [59676-
Taiwan [7102
United Kingd
United States
Uruguay [133
Problems
Annotation Explorer
Text analytics result, Number of rows: 10/10
Showing page 1 of 1
Input Document
Left Context
Span Attribute Value
Right Context
\Facts.txt
ructure, and widespread land mi...
Afghanistan [1397-1408]
Location: Sou
\Facts.txt
dence was soundly defeated in 1...
Bermuda [17737-17744]
Location: Nor
\Facts.txt
e and the world's newest democr...
East Timor [28767-28777]
Location: Sou
\Facts.txt
flared in 1995 was resolved in 199...
Ecuador [40228-40235]
Location: Wes
\Facts.txt
erior as a National Wildlife Refug...
Howland Island [55811-55825]
Location: Oce
\Facts.txt
hen the islands gained independe...
Palau [59676-59681]
Location: Oce
\Facts.txt
estic political and economic refor...
Taiwan [71026-71032]
Location: East
\Facts.txt
d Assembly were established in 1...
United Kingdom [88320-88334]
Location: Wes
\Facts.txt
, and rapid advances in technolo...
United States [111783-111796]
Location: Nor

3. Примеры AQL экстракторов

Извлечение координат

```
create view Locations as
extract regex /Geographic coordinates: {1,}((\d{1,2} \d{2} [NS]), (\d{1,3}
\d{2} [EW]))/ on D.text
    return group 1 as longlat
    and group 2 as longitude
    and group 3 as latitude
from Document D;
```

Работа со словарями

```
create dictionary MonthsDict as (
'January', 'February', 'March', 'April', 'May', 'June', 'July', 'August',
'September', 'October', 'November', 'December'
);
create view Months as
extract
    dictionary 'MonthsDict'
    on D.text as month
from Document D;
```

Работа с частями речи

```
create view Nouns as
extract part_of_speech 'NN' and 'NNS' with language 'en'
on D.text as noun
from Document D;
```

Работа с блоками

```
create view BlockNouns as
extract blocks
with count between 2 and 3
and separation between 0 and 50 characters
on n.noun
as BlockedNouns
from Nouns n;
```

Работа с паттернами

```
create view Countries as
extract regex /Geography {1,}([A-Z][a-z]{1,} ?[A-Z]*[a-z]*)/ on D.text
return group 1 as country

from Document D; create view CountryLocation as
extract
pattern (<C.country>) <Token>{5,35} (<L.longlat>)
return group 1 as country
and group 2 as location
from Countries C, Locations L;
output view CountryLocation;
```

Работа с Select

```
create view Textnames as
select GetText(C.country) as country, GetText(C.location) as location
```

```
from CountryLocation C;
```

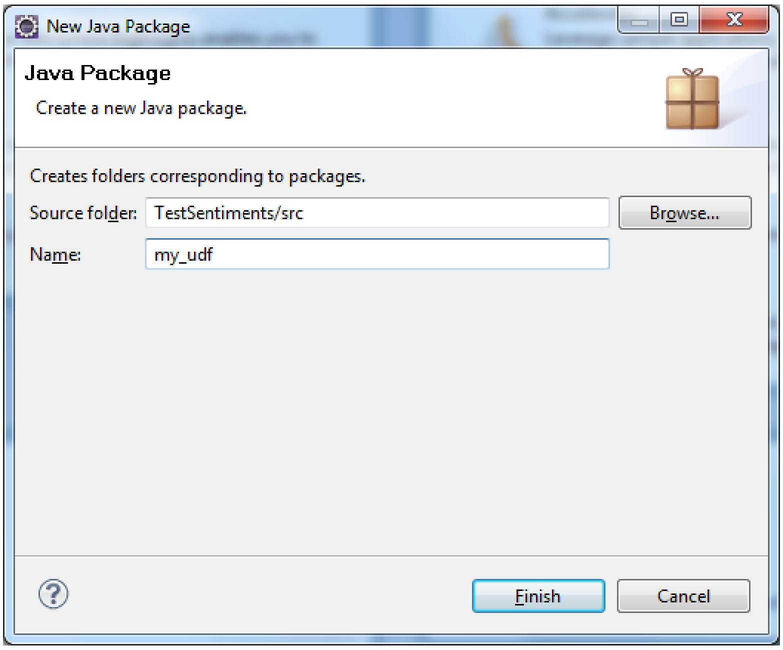
4. Пример задачи сентиментального анализа

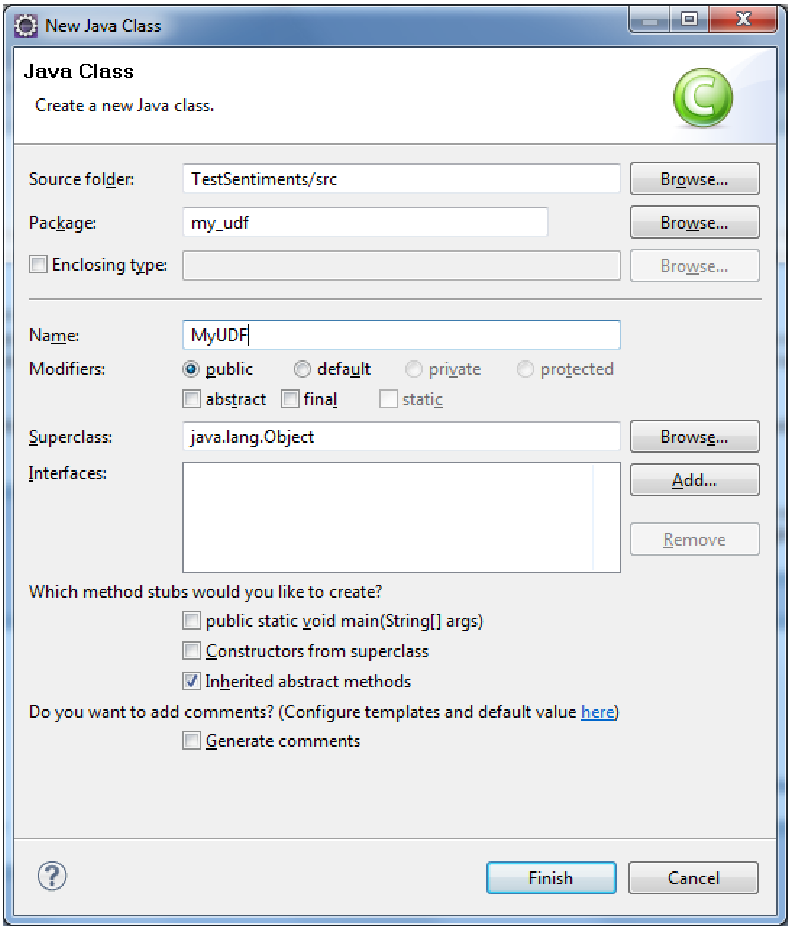
```
create dictionary CountryDict as (
'Afghanistan', 'Bermuda', 'East Timor', 'Ecuador', 'Howland Island', 'Palau',
'Taiwan', 'United Kingdom', 'United States', 'Uruguay'
);

create dictionary NegativeSentimentsDict as (
'reduce', 'poor', 'fall', 'insufficient'
);

create dictionary PositiveSentimentsDict as (
'growth', 'efficient', 'intensive', 'good', 'positive'
);

create view Country as
extract
      dictionary 'CountryDict'
      on D.text as country
from Document D;

create view Paragraph as
extract
    split using B.boundary
      retain right split point
      on B.text
      as text
      from (
            extract
                  D.text as text,
                  regex /(\n\s*\n)/ on D.text as boundary
                  from Document D
      ) B
;

output view Paragraph;

create view EconomyOverview as
extract regex /Economy - overview:/ on D.text
return group 0 as text
from Document D;

create view EconomyOverviewParagraph as
select
      P.text
from Paragraph P, EconomyOverview EO
where
      Contains(P.text, EO.text)
;

output view EconomyOverviewParagraph;


create view CountryEconomyOverview as
select
      C.country
      , P.text
from EconomyOverviewParagraph P, Country C
where
      Follows(C.country, P.text, 1, 10)
;

output view CountryEconomyOverview;
```

```
create view EconomyPositive as
extract
	dictionary 'PositiveSentimentsDict'
	on CEO.text as positive
from CountryEconomyOverview CEO;

output view EconomyPositive;

create view CountryEconomyPositive as
select
	CEO.country
	, EP.positive
	, 1 as cnt
from CountryEconomyOverview CEO, EconomyPositive EP
where
	Contains(CEO.text, EP.positive)
;

output view CountryEconomyPositive;

create view EconomyNegative as
extract
	dictionary 'NegativeSentimentsDict'
	on CEO.text as negative
from CountryEconomyOverview CEO;

output view EconomyNegative;

create view CountryEconomyNegative as
select
	CEO.country
	, EN.negative
	, -1 as cnt
from CountryEconomyOverview CEO, EconomyNegative EN
where
	Contains(CEO.text, EN.negative)
;

output view CountryEconomyNegative;

create view CountrySentiment as
(select
	CEP.country
	, CEP.cnt
from CountryEconomyPositive CEP)
union all
(select
	CEN.country
	, CEN.cnt
from CountryEconomyNegative CEN)
;

output view CountrySentiment;
```

5. Работа с пользовательскими функциями

Создание Java класса





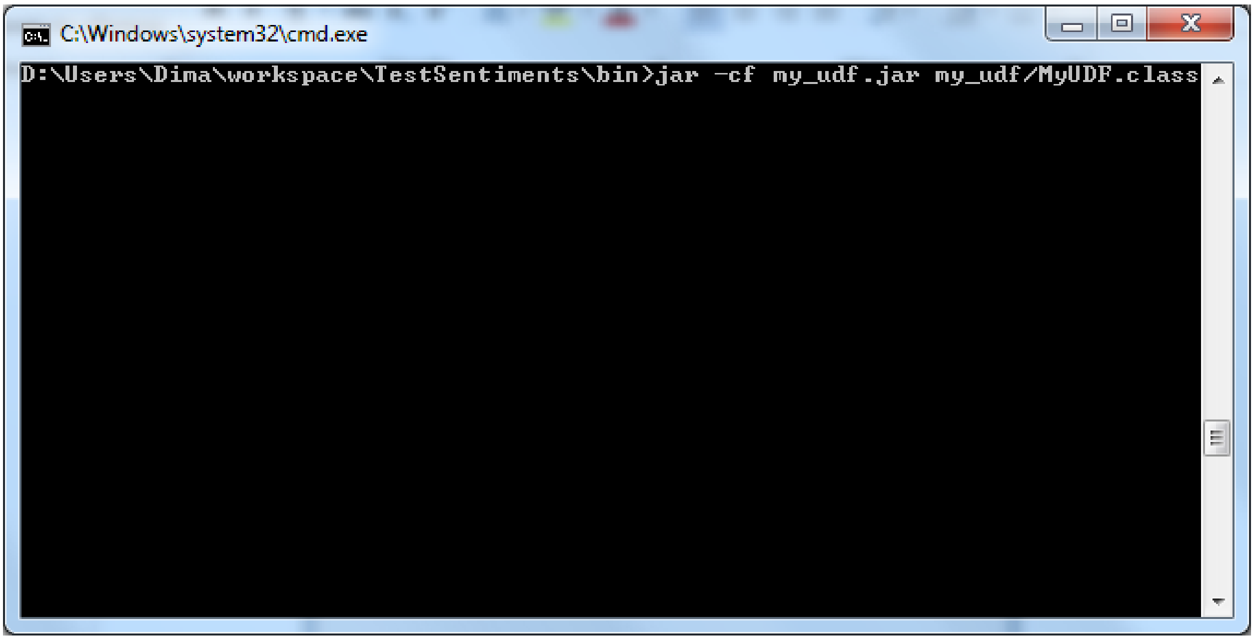
Редактирование класса. Создание функции toUpperCase

```
package my_udf;

import com.ibm.avatar.algebra.datamodel.*;

public class MyUDF {
```

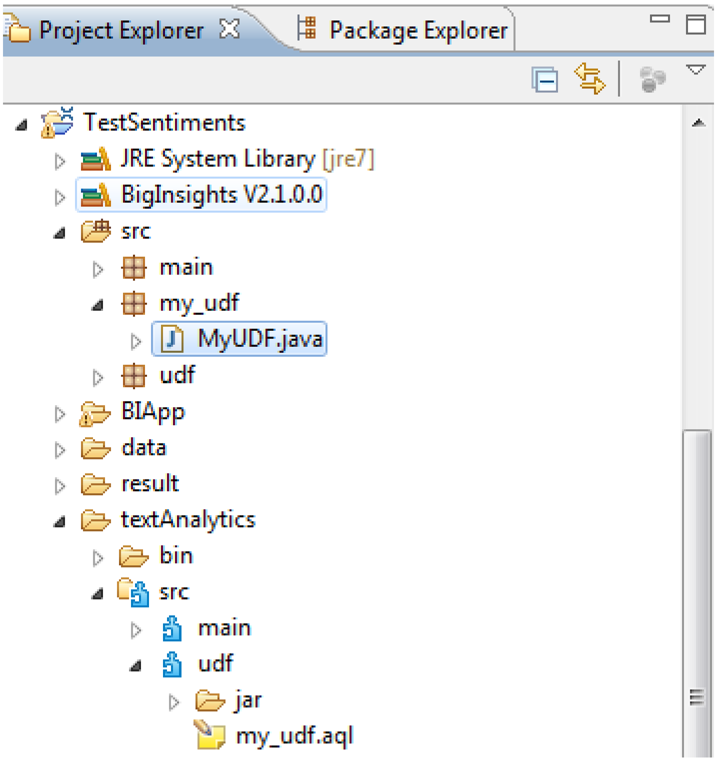
```java
        public String toUpperCase(Span s) {
              return s.getText().toUpperCase();
        }

}
```

Компилирование проекта и создание jar файла

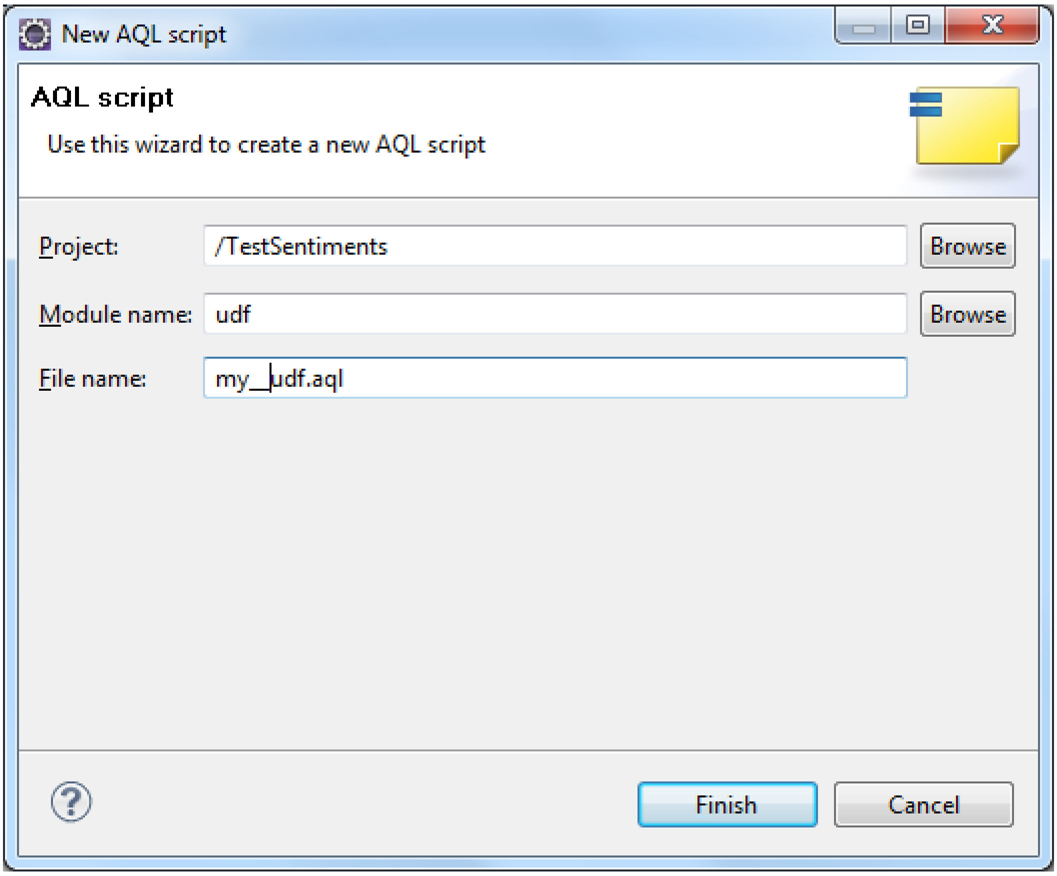
1) Откомпилировать проект
   Project -> Build Project
2) Открыть командную строку с директорией %EclipseWorkspace%\%ProjectName%\bin
3) Запустить команду создания jar файла:
   jar –cf my_udf.jar my_udf/MyUDF.class



4) Создать AQL module udf и в нем создать папку jar



5) Создать AQL скрипт my_udf.aql



6) Редактировать скрипт my_udf.aql

```
module udf;
-- TODO: Add AQL content here

create function udfToUpperCase(p1 Span)
return String
external_name 'jar/my_udf.jar:my_udf.MyUDF!toUpperCase'
language java
deterministic
return null on null input;

export function udfToUpperCase;
```

7) Редактировать скрипт main.aql. Добавить в него следующее представление

```
import function udfToUpperCase from module udf as udfToUpperCase;

...

create view CountrySentimentUpper as
select
      udfToUpperCase(CS.country) as country
      , CS.cnt
from CountrySentiment CS
;
output view CountrySentimentUpper;
```