Programming Lab III, WS 2016/2017

Handout 2

Dr. Marc Zimmermann, Jens Dörpinghaus
2016-11-08

Second task – Combining CollectionReader and AnalysisEngine, learn the SCAI Typesystem

Has to be completed finally at 2016-11-15, 11am.

Now that you've (hopefully) developed a simple reader using UIMA, we'll use it for the next few weeks. If you have not successfully developed the reader pipelet yet you can use the one from the *standard_solutions* directory at the svn server .

The next task will be to create an **AnalysisEngine (AE)**, which parses gene sequences. There are many common conventions how to store gene sequences in different file formats. And there's help available in form of a Java framework, too:

BioJava is an open-source project dedicated to provide a Java framework for processing biological data. It provides analytical and statistical routines, parsers for common file formats and allows the manipulation of sequences and 3D structures. The goal of the BioJava project is to facilitate rapid application development for bioinformatics.
See http://biojava.org for further information.

Read the documentation of BioJava, especially the Cookbook published on their web site. There are many well explained examples which can help you finish the task.

Here's the list of topics you have to complete:

- Use your already developed CollectionReader (or create a new one; but bear in mind that it's also part of the course to reuse code that already exists).

- The pipelet should already be able to read files in the FASTA format and convert them to CAS (.xmi). We will place a FASTA file on the Gforge server, which includes a few data sets. Have a look to the FASTA file format specification, too (Google helps).

- Look into the code of the UIMACorePipelet and UIMATypeSystem (you'll find it under Maven Dependencies in your project folder), especially into the code of package *de.fraunhofer.scai.bio.extraction.types* and its sub-packages. This is the **SCAI Typesystem** for UIMA. Also read the documentation about type systems at https://uima.apache.org

- Create an AnalysisEngine (AE) pipelet which annotates each protein sequence stored in an .xmi (make use of your reader and Bio-Java here) of the FASTA data sets and add it to the annotation index as a so-called **NormalizedNamedEntity** (short *NNE*, a type defined in the SCAI Type system for UIMA). Hint: there is an Annotator template in the artifact! Add BioJava in the dependencies of your *pom.xml*.

- Install the UIMA CAS-Editor from Eclipse Tooling (https://uima.apache.org/downloads.cgi) to view the CAS files (yes, it is Eclipse with UIMA as plugin). Set the highlight color of the NNE in the CAS-Editor to green – make a screenshot and add this to your svn.

- Store the output files (.xmi) in folder *src/test/resources/output/task2*.

- Commit the new AnalysisEngine pipelet into your personal svn folder at SCAI GForge under task2.

- Think about a self-describing package name for the AE.

Good luck and have fun!