Homework 1 report of Software methods for biotechnology-11693
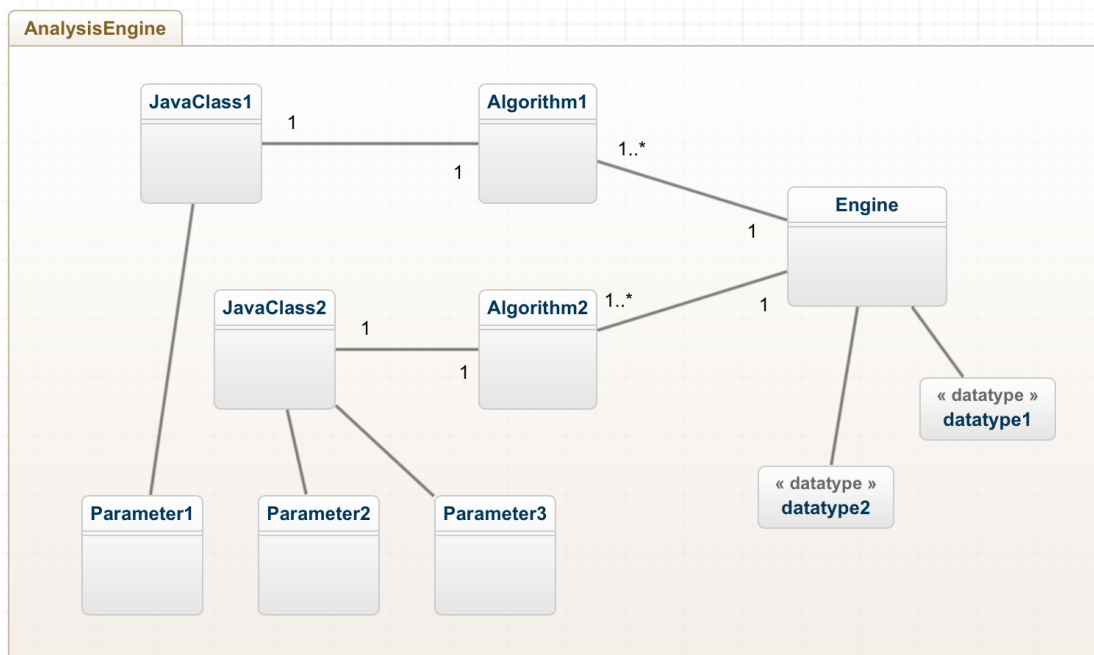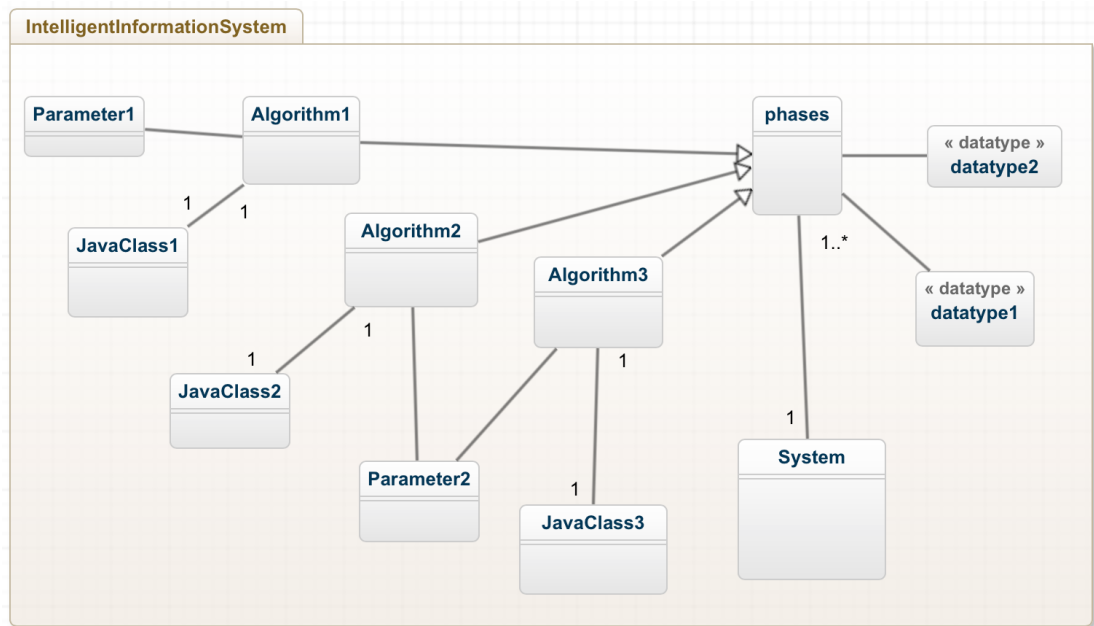
Name: xiaoxulu

AndrewID: Xiaoxul

I have got a lot of useful information about LINGPIPE (http://alias-i.com/lingpipe/index.html) and a lot of help regarding the homework from Xiao Han. As one of my major class of homework, geneAnalysis.class, which is gene annotator of the CPE, was spired by him a lot. And Victor Zhao helped me a lot about all this git and maven things, even though I still failed to hand in my homework in the desired way.)

Task1

UML documents

IntelligentInformationSystem

Parameter1 — Algorithm1 — phases — « datatype » datatype2

1 1 1..* « datatype » datatype1

JavaClass1 Algorithm2 Algorithm3

1 1

JavaClass2 1 System

Parameter2 1

1 JavaClass3

AnalysisEngine

JavaClass1 — 1 — Algorithm1 — 1..* — Engine

1 1 « datatype » datatype1

JavaClass2 — 1 — Algorithm2 — 1..* — 1
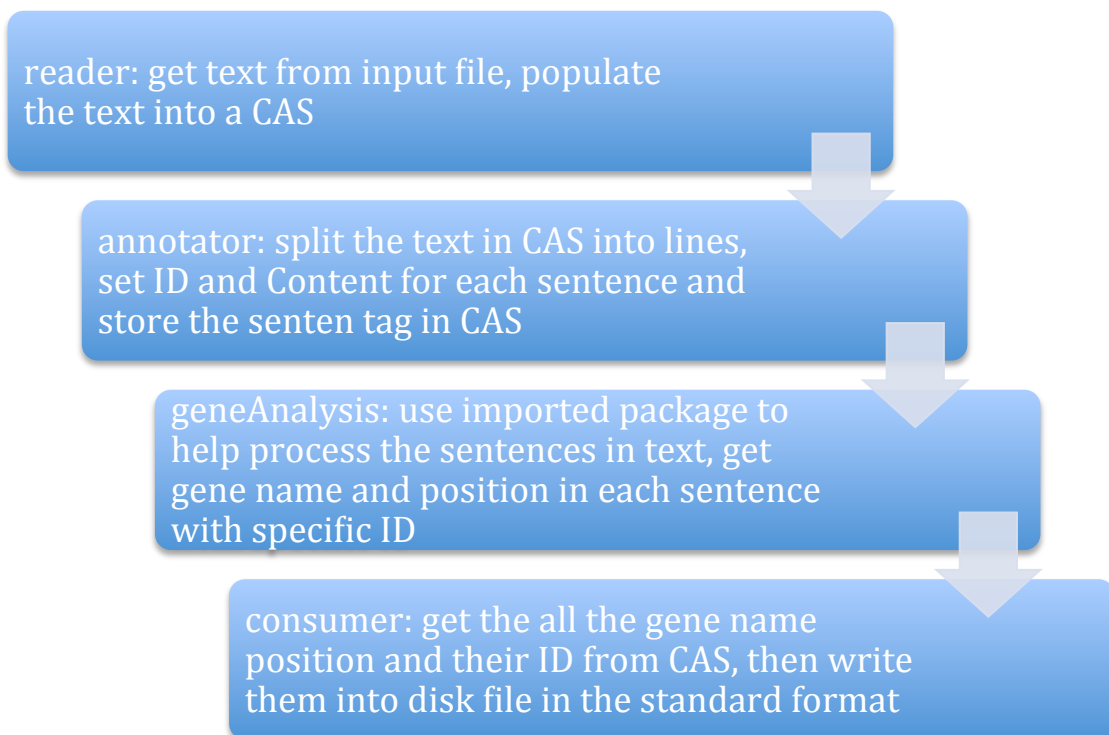
1 1 « datatype » datatype2

Parameter1 Parameter2 Parameter3

Task2
Design and Implementation of IIS

UIMA provides a standard frame for us to analysis totally unstructured dataset. By definition of tutorial, CPE is composed of three parts: Collection Reader, Analysis Engine and CAS Consumer. In my designed program, Collection Reader is implemented by **reader.class**; Analysis Engine is implemented by **annotator.class**, together with **geneAnalysis.class**; and CAS Consumer is implemented by **consumer.class**. These 4 classes communicate via an object of class CAS and here is the flow.

reader: get text from input file, populate the text into a CAS

annotator: split the text in CAS into lines, set ID and Content for each sentence and store the senten tag in CAS

geneAnalysis: use imported package to help process the sentences in text, get gene name and position in each sentence with specific ID

consumer: get the all the gene name position and their ID from CAS, then write them into disk file in the standard format

The SimpleRunCPE.class initializes a CPE and process the input file in the flow.