Name: Yan He
Andrew ID: yanhe
September 23, 2014

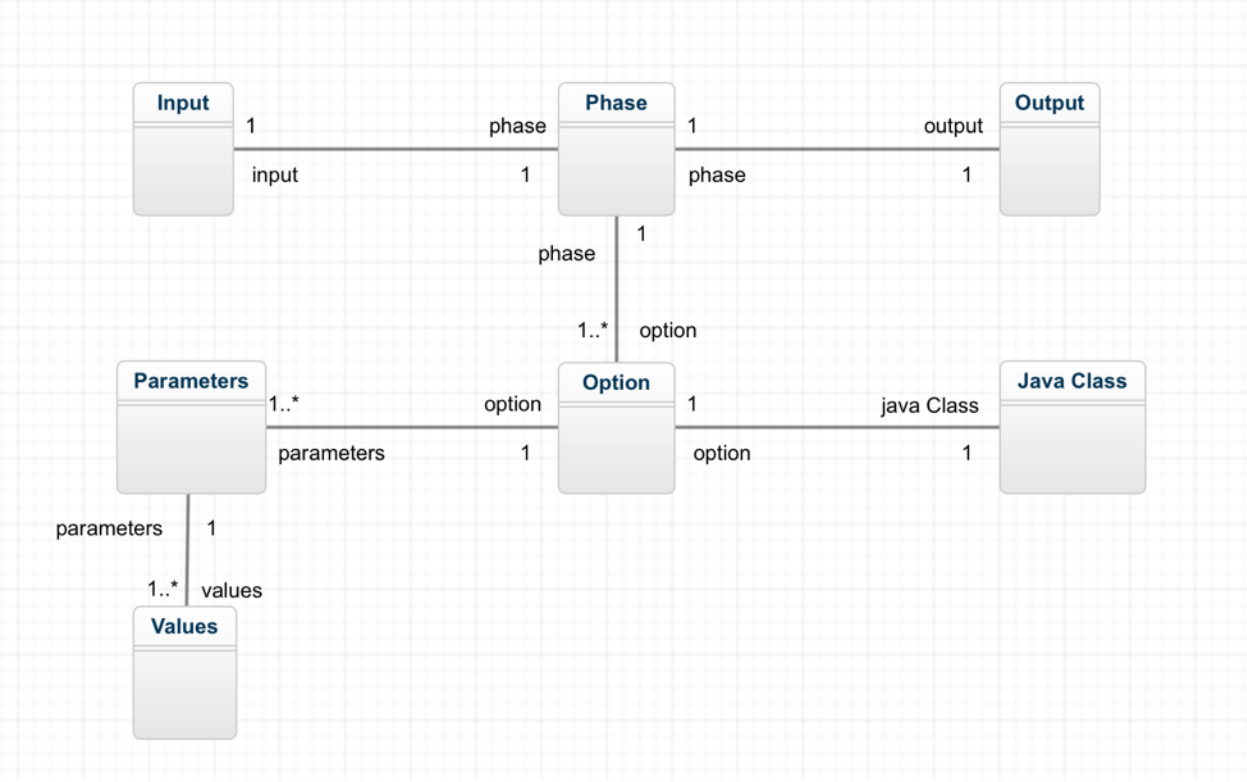
Software Methods for Biotechnology

Homework 1 - Final Report

1. UML Design

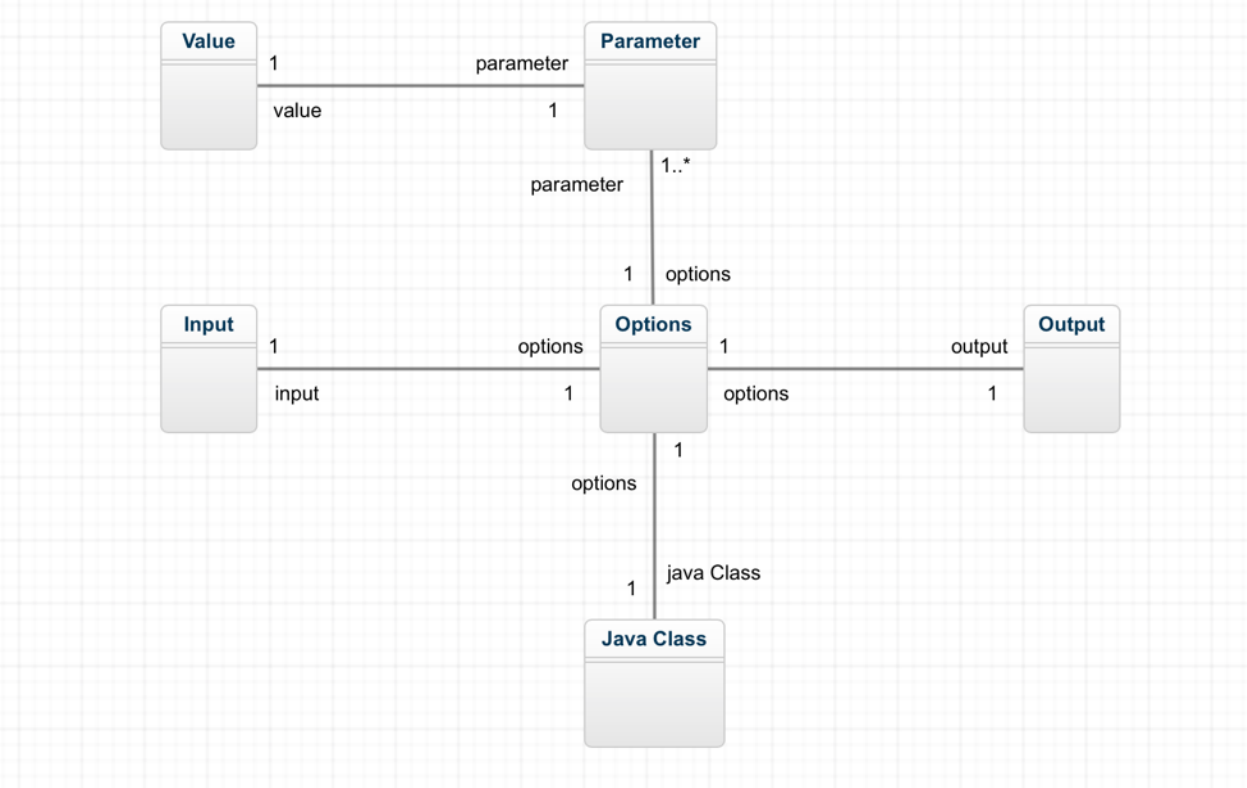
1.1 Domain diagram for IntelligentInformationSystem

Here is the class diagram for Intelligent Information System:



1.2 Domain diagram for AnalysisEngine

Here is the class diagram for Analysis Engine:



2. NLP techniques/components

For this assignment, I've discussed with my classmate and decided to use the LingPipe tool kit instead of Stanford NLP to find the gene name in the content of the file. The reason that I chose LingPipe is that it could improve the accuracy of the identify process. It has provided the model of human gene package and I could use it to compare with the content in the input file and find the target gene name. Then the program could also store the final result into a local file in organized format.

3. General data flow in the system

In this gene identify system, the first thing that the system do is to import the file in the Collection Reader and convert the content of the file into string and store it inside the system. Then it has been passed to the Annotator and split by each sentence. Also, the Annotator has also split it by the ID of the sentence at the first of each line with the other sentence content right after the ID. After that, the data has been transferred to geneFinder which identify the gene in the content via LingPipe tool kit. The gene name and ID, as well as the index of the start and end of the gene name has been extracted and calculated,

which has been processed by CAS Consumer finally. Consumer got all the necessary data it needed and then wrote it inside a local file. To be honest, I was planning to compare the result with the sample and calculate the identify performance of this system, but due to the lack of time, I think I could try it out after the deadline. After all, it's worth it.