Final Report

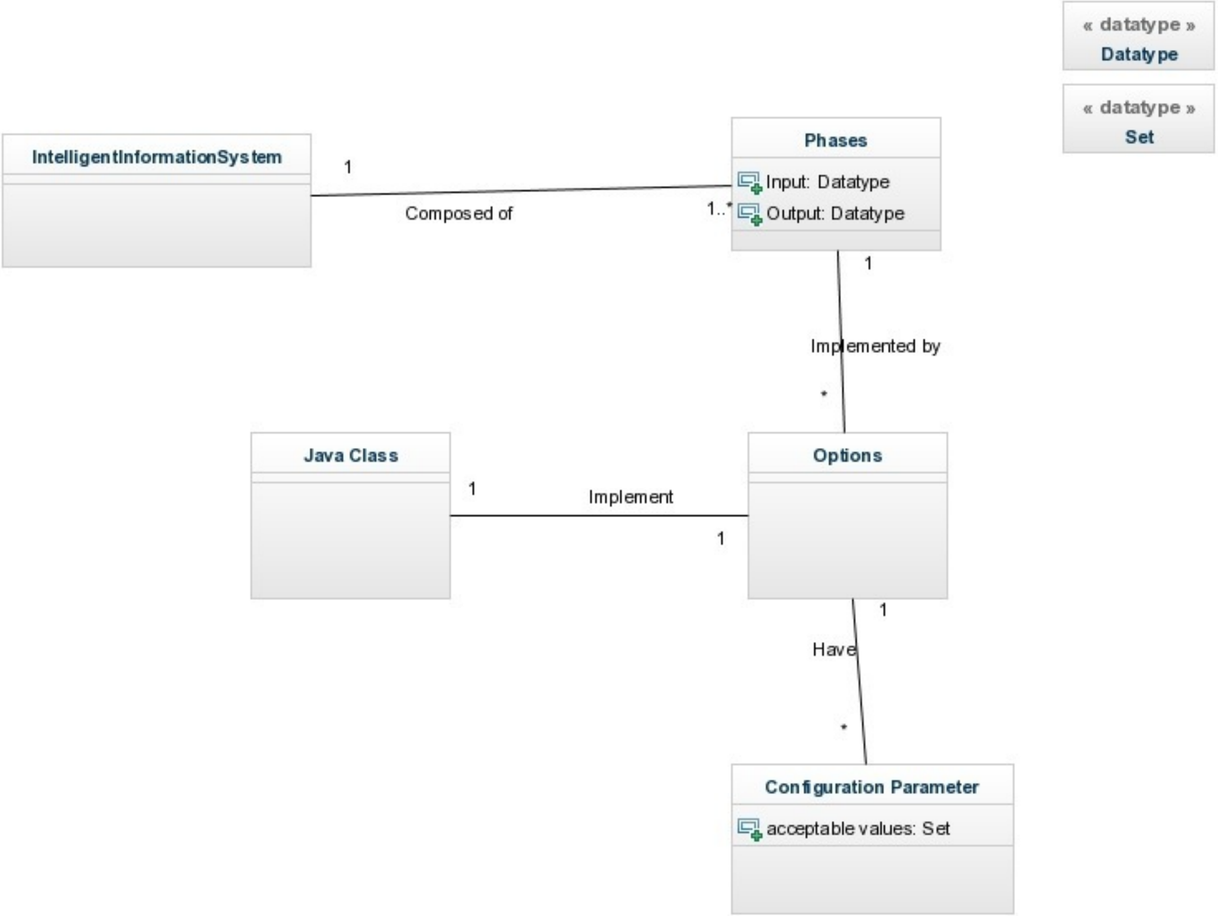
Andrew ID : yichenca

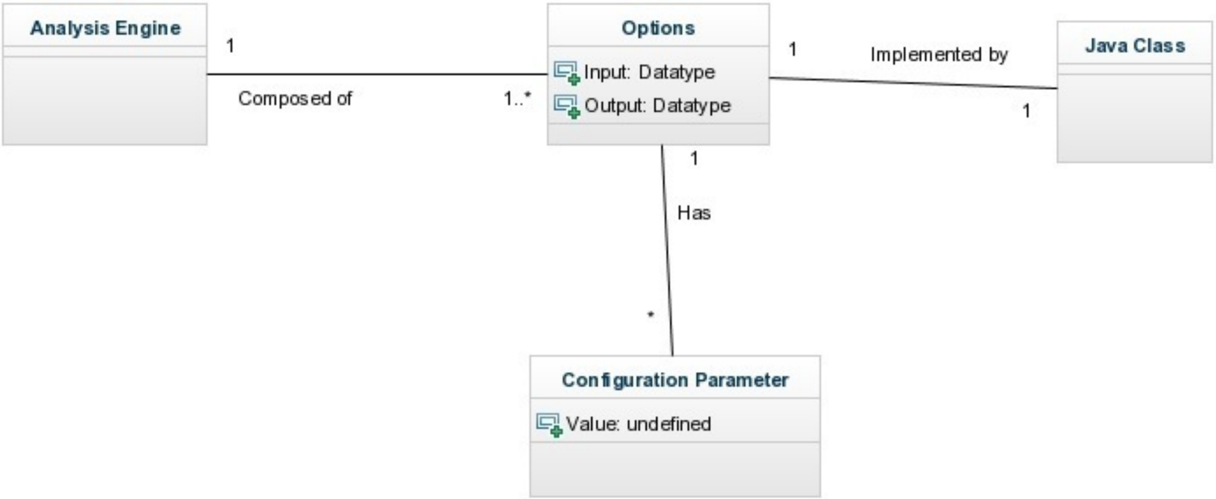
Author: Cai Yichen

Date: 2014-09-18

Task1 :

UML Diagram




Analysis Engine
1
Composed of
1..*
Options
Input: Datatype
Output: Datatype
1
Implemented by
1
Java Class
1
Has
*
Configuration Parameter
Value: undefined

Implementation methods

UIMA Architecture :

Basically, the TypeSystem defines two annotations, one is Text which contains two features: id and sentence text; another one is Genetype which contains two features : id and recognized gene.

The CollectionReader class reads a sentence each time from file, and split the sentence into id and sentence text, then put them into Text annotation, respectively.

The GeneAnnotator class gets the feature of Text annotation, and uses LingPipe to recognize specific gene, then put the id, start point, end point and gene into Genetype annotation, respectively.

The CasConsumer class gets the features of Genetype annotation, and write the features to a given output file.

For the name entity recognition, the LingPipe runs a statistical named entity Recognizer, specifically for gene, which is "ne-en-bio-genetag.HmmChunker" , the object chunker can load the begin position and end position of recognized gene.

I use sample.in to test this method, as a result, it could recognize specific genes with high precision.

File path:
All the .java files such as CollectionReader, GeneAnnotator , CasConsumer, Type are in /src/main/java.
All the .xml files such as asDescriptor, casConsumerDescriptor, collectionReaderDescriptor, typeSystemDescriptors are in /src/main/resources/descriptors.
The input file directory is /src/main/resources/In ; The output file directory is /src/main/resources/out, and the output filename is hw1-yichenca.out.
The file /src/main/resources/ne-en-bio-genetag.Hmmchunker is recognizer used by LingPipe.