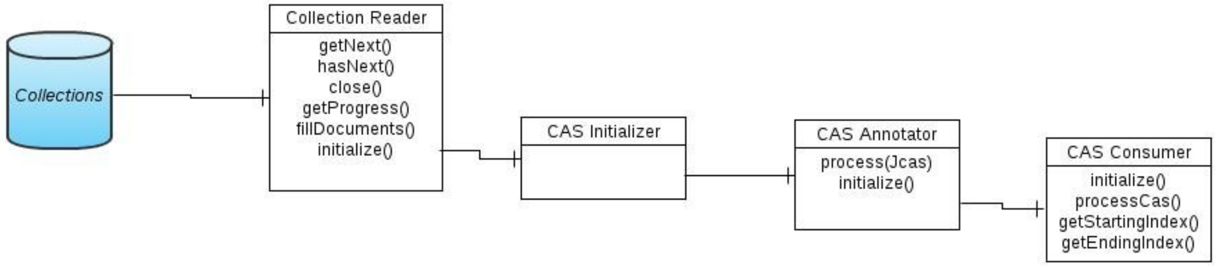
NER Report

Anika Gupta
anikag

Design

UML Diagram



UIMA Components

1. Type System
2. Collection Reader
3. CAS Annotator
4. CAS Consumer

UIMA Components Description

1. Type System
   NamedEntity Type has been derived from the uimas.tcas.Annotation. Besides the inherited features(start,end and sofa feature), the DocID feature of Range type String is added to it to store the documentID.
2. Collection Reader
   The parameter InputFile is to give the path to the input file. The collection reader is inherited from the `CollectionReader_ImplBase`. It initializes the FileReader and stores the documents in an ArrayList. The getNext(CAS casObj) function selects a document from the array list at a time ,annotates it and sends it to the CAS consumer. The same process is repated for all the documents in the array list.
3. CAS Annotator
   The TrainedModel parameter to the CAS Annotator stores the path to the model file used for the annotation of the text. In the process function, it reads a Jcas object, the document is retrieved from it. The annotations for the model are produced(more on this later – in the Algorithms section). The document text is

annotated by setting the beginning and end of the span. The type object is created and the annotated text is indexed into it.

4. CAS Consumer
   The initializer process initializes the FileWriter by reading the location of the output file from the parameter OutputFilePath. In the processCas function, it loads a CAS object and retrieves all the annotations associated with it. The function calls a function to convert the start and end of annotation span into the format as per required by the question. The output is appended onto a file.

Algorithm Used

The model for Genetag in the Named Entity Recognizer for LingPipe is the trained model used. The trained model is based on Hidden Markov Model and character language-based chunkers to extract the mention of genes. A full description of the algorithm can be found in the following paper[1].

References

[1] http://www.colloquial.com/carp/Publications/biocreative-8-alias-i.pdf