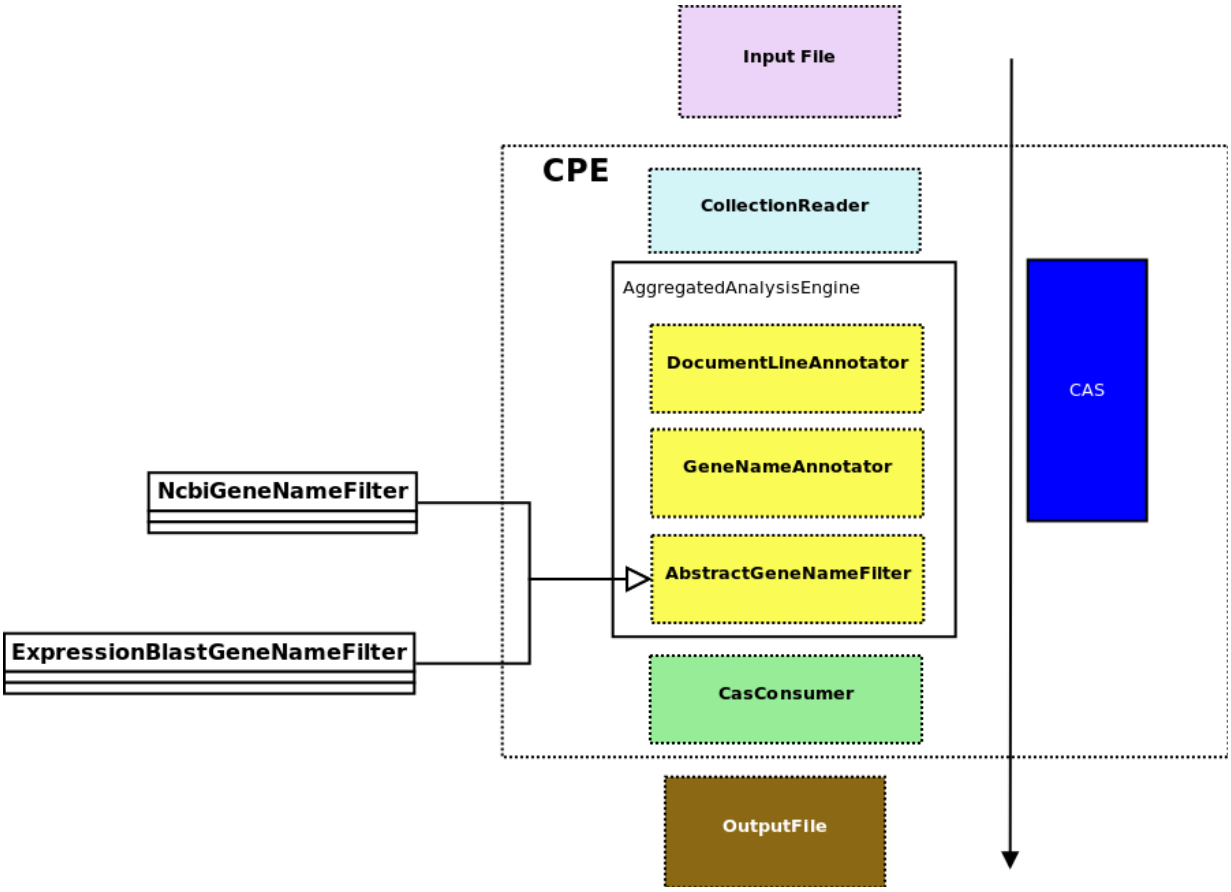
Homework 1 report

Bo Lei

System Design

The following diagram describes the system architecture.



The pipeline consists of 5 components: 1 Collection reader, 3 Annotators and one Cas consumer.

Type System

Two types are defined. They are DocumentLine and GeneName.

The DocumentLine class represents each line in the input file. It contains 2 attributes: sentenceId and content.

The GeneName class represents a gene name annotation. It contains the attributes begin, end which are inherited from Annotation class; and also sentenceId and name attributes.

Collection reader:

Collection reader loads the entire content of the input file into memory, creates one single Cas object that wraps the input file content and pass it to the downstream annotators.

Analysis Engine:

The Analysis Engine is composed of 3 primitive annotators. They follow a Fixed Flow order. The DocumentLineAnnotator annotates each line of the input file as an DocumentLine object. Each of the object is added into the Cas index.

The Cas is then passed to GeneName annotator. It parses each DocumentLine object and recognizes Gene Name Entities using “PosTagNamedEntityRecognizer” that is provided in homework1.

Because “PosTagNamedEntityRecognizer” is not accurate, a third annotator “GeneNameFilter” is used to filter out the gene names that are not correctly identified.
An “AbstractGeneNameFilter” is declared as an abstract class and provide the general logic to filter out gene names. Inside this abstract class, an abstract method "isGeneName(String name)“ is declared. which is a template method for the extended class to implement.

There are two classes that extend from AbstractGeneNameFilter and implement “isGeneName(String name)” method:

1. For each of the the raw gene names, the system queries the online NCBI database to see if it is indeed a gene name. This mechanism is implemented by the class “NcbiGeneNameFilter”. This approach gives better result. But it takes a long time to run all the input data (several hours) because each HTTP request takes a significant amount of time.
2. A list of gene names is acquired from ExpressionBlast.com (a project by Dr. Guy Zinman). Upon startup of the annotator “ExpressionBlastGeneNameFilter”, all the gene names are loaded from a file to a HashSet. Each raw gene name annotation is then compared against that inside the HashSet. This approach is simple and fast. However it doesn’t provide very good accuracy due to the limit number of gene names in the local gene name file.

Cas Consumer

The responsibility of Cas Consumer is very straightforward. It retrieves all the Gene Name Annotations from the Cas index, and prints them into the output file.