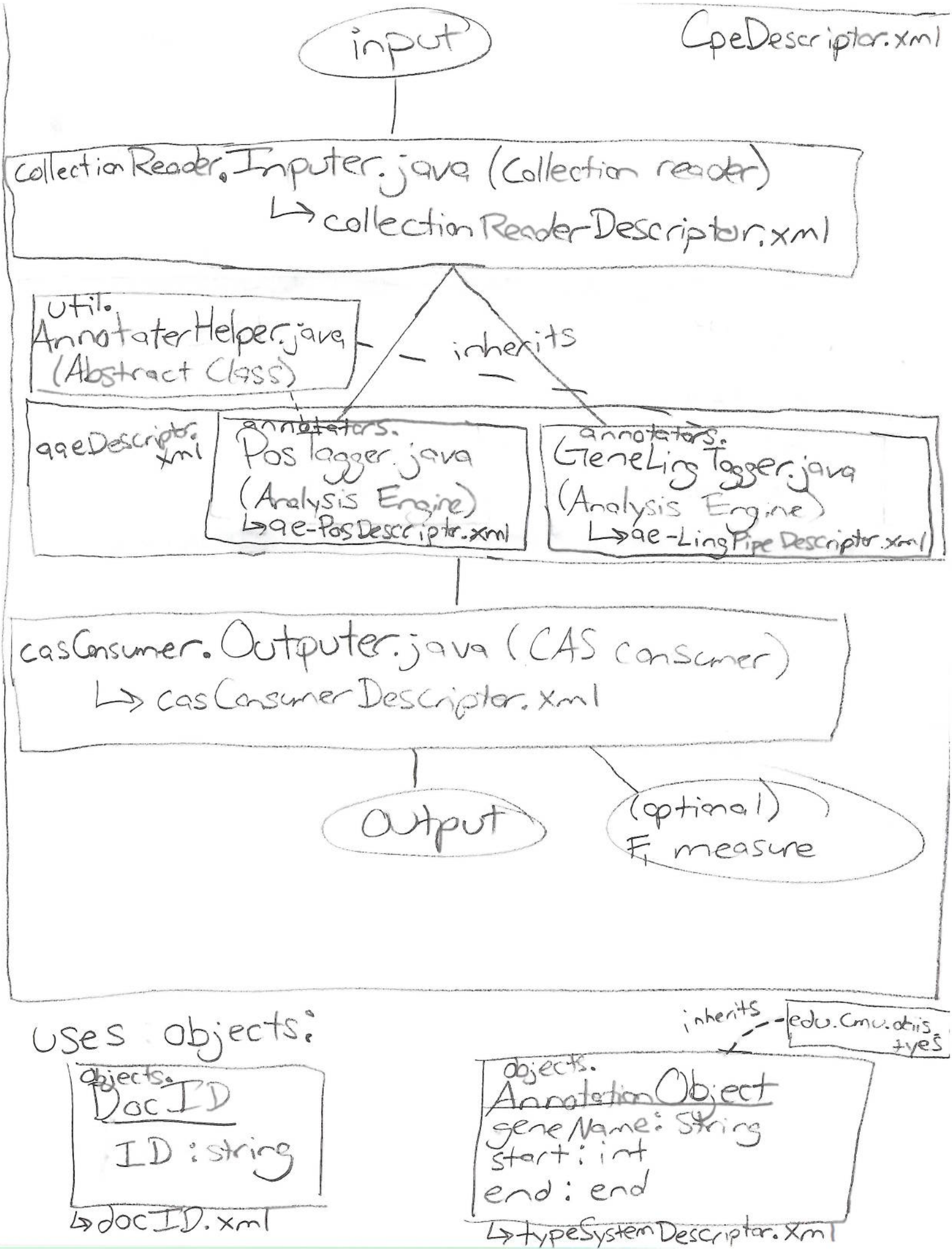
Lara Martin **HW2-ljmartin-report** Due: October 10, 2014

In my implementation of the gene-entity finder, I kept the part-of-speech tagger, which looks for nouns and adjectives, and then removes common words based off of a list of the most-common words found in English. Each of these annotations are given the same confidence: the F1 score that this annotator gets on the training data. For another analysis engine, I use a pre-trained HMM-chunker from LingPipe which finds gene mentions. I take the top 10 (max) chunks and if they have a confidence over 60%, they are added to the CAS. In the CAS consumer, both annotations are combined in the following way:

- If there are annotations from both analysis engines, add the confidence scores and store it.
- If there is no gene annotation from LingPipe, just ignore it.
- If the combined score of an annotation is greater than 80%, then write it to the output file.

The following diagram shows the structure of the program:


input
CpeDescriptor.xml
collectionReader.Inputer.java (Collection reader)
↳ collectionReaderDescriptor.xml
util.
AnnotaterHelper.java
(Abstract Class)
inherits
aaeDescriptor.xml
annotators.
Pos Tagger.java
(Analysis Engine)
↳ ae-PosDescriptor.xml
annotators.
GeneLing Tagger.java
(Analysis Engine)
↳ ae-LingPipeDescriptor.xml
casConsumer.Outputer.java (CAS consumer)
↳ casConsumerDescriptor.xml
output
(optional)
F measure
uses objects:
objects.
DocID
ID : string
↳ docID.xml
inherits
edu.cmu.tiis.types
objects.
AnnotationObject
geneName: String
start: int
end: end
↳ typeSystemDescriptor.xml