11-693 Software Methods for Biotechnology

Homework 1 report

Xuwei Zou
xuweiz

Architecture of models

TypeSystem:

The type system I create containing 4 features:

1)Mark, which store the id for every sentence;

2)Gene, which store the gene tag retrieved by Annotator;

3)Start, which store the start offset of the gene tag in the sentence.

4)end, which store the end offset of the gene tag in the sentence.

Collection Reader:

The Reader is responsible for read document, separate ID and text and create CAS for Annotator.

Analysis Engine:

The Engine contains one Annotator. The Annotator will analyze CAS created by the Reader and produce gene tags and offsets. Then the Annotator will update CAS with gene tags and offsets.

CAS Consumer:

The Consumer is designed to format the output line and output final result. The Consumer will receive sentence ID, gene tags and offsets in the CAS and calculate the gene tags' offsets (exclude white space) in the formation of given sample.

LingPipe corpus trainer:

The trainer retrieve document in the certain pattern and produce database used as base corpus pool for LingPipe analysis tool.

Algorithms and tools used

In the Annotator, the LingPipe tool is used to analyze the gene tag in the sentence. LingPipe tool provides sophisticated methods to retrieve words and phrase from sentences by using certain database. Also LingPipe provides tool to create corpus for analysis tool. In the trainer, a base data provided by LingPipe example(provided by NCBI) is used to create corpus for LingPipe analysis tool. However, current version(4.10) LingPipe package does not contain the parser used in trainer, the trainer method is conducted under a lower version LingPipe package.