Report

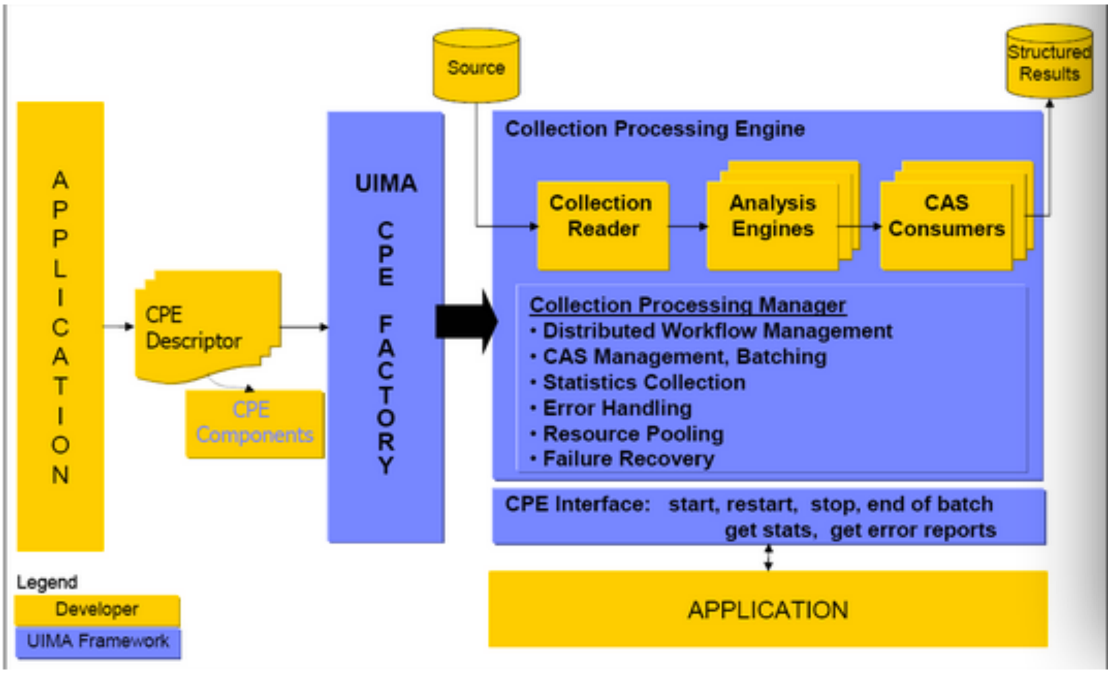
Zhu MENG mpzhu

beginning

In this homework, first I felt confused about how to use eclipse to build a maven project which can realize the function as a CPE. How to read a text? How to analyse the NameEntity? How to get the answers and how to organize all this part? With the question, I try to find out the answer.

Build the CPE:

As far as I know from the tutorial, the whole project should contain these parts:



`Collection Reader：`

A Collection Reader is responsible for obtaining documents from the collection and returning each document as a CAS. Like all UIMA components, a Collection Reader consists of two parts — the code and an XML descriptor.

I use the File System Collection Reader, which simply reads documents from files in a specified directory. It contains a .xml and a .java. there are some specific function in this part such as hasNext(), getNext(),etc…which will do the work as "read" the text.

Analysis Engine:

An Analysis Engine consists of two parts - Java classes (typically packaged as one or more JAR files) and *AE descriptors* (one or more XML files). This is the part that I should figure out how to recognize the gene Named Entity.

Cas Consumer:

A CAS Consumer receives each CAS after it has been analyzed by the Analysis Engine. CAS Consumers typically do not update the CAS; they typically extract data from the CAS and persist selected information to aggregate data structures such as search engine indexes or databases. I use the AnnotationPriinter as the Cas Consumer which will use annot. to find the indexes in the AE and print them out in hw1-mpzhu. out text.

CPE:

After these three part been realized, I use genetypeannotationsystem as the CPE which combine all the three parts together

Thinking about the methods:

There are some kinds of methods that can be used to mark the Name Entity.

Dictionary: find a database and find out whether the words can be found. However, I do not know how to resolve whether the Name Entity contains one word or more words that I can not simply cut them into piece of word.

The rule of gene name:

Then I want to use pattern() and match() to set the rules of gene names. But there are some words are also the knowledge of gene but do not need to follow the rule of genes' or proteins' names.

Machine Learning:

Use the simple.in and simple. Out as train datas. Also can use 0.632bootstrape to achieve a better model. While I have limited time so I do not try this method.

NLP:

Since there are many NLP resources in the website that finally I choose this method.

Stanford NER (also known as CRFClassifier) is a Java implementation of a Named Entity Recognizer. Named Entity Recognition (NER) labels sequences of words in a text which are the names of things, such as person and company names, or gene and protein names. The software provides a general (arbitrary order) implementation of linear chain Conditional Random Field (CRF) sequence models, coupled with well-engineered feature extractors for Named Entity Recognition.

The imports are:

```
edu.stanford.nlp.ling.CoreAnnotations.PartOfSpeechAnnotation;
edu.stanford.nlp.ling.CoreAnnotations.SentencesAnnotation;
edu.stanford.nlp.ling.CoreAnnotations.TokensAnnotation;
edu.stanford.nlp.ling.CoreLabel;
edu.stanford.nlp.pipeline.Annotation;
edu.stanford.nlp.pipeline.StanfordCoreNLP;
edu.stanford.nlp.util.CoreMap;
```

since there are a lot other nlp package such as lingpipe, etc, I just choose the this Stanford NER. It based on calculate the nature language features and reach the result. While it just find out the TokensAnnotation but do not focus on gene name, this is a drawback. If the time is not limited, I hope I can find a way to figure out this problem such as add a dictionary or a reference database.