ISMLA 17/18: UIMA Exercises

1 CollectionReader

1. Set up a collection reader descriptor file with two String-valued configuration parameters: language and inputFile.
2. Write a collection reader that gives back a JCas for each line of an input file described by the configuration parameter.

2 CASConsumer

1. Create a CASConsmuer descriptor and a Java class extending `org.apache.uima.collection.CasConsumer_ImplBase`, configure like other descriptors.
2. In the Java file write a CASConsmuer that writes statistics about the document to be processed to a single output file (set via configuration parameter). The output should look similar to the following:

```
/tmp/input-data/doc2.txt
language: en
#sentences: 11
#tokens: 258
#token types: 173
ttr: 0.6705426356589147

/tmp/input-data/doc3.txt
language: en
...
```

3 Assembling Modules

Assemble all the modules:

- Collection Reader

- Aggregate Analysis Engine
- CASConsmuer

via the CPMFrame `org.apache.uima.tools.cpm.CpmFrame`) and test your complete pipeline on input data (on Moodle).