Hw3 Documentation, LTI 11791

Abhimanu Kumar
Andrew Id - ABHIMANK

November 4, 2013

1 Task 1.

I did the following for task 1:

- A FileSystemCollectionReader.xml descriptor and its corressponding class org.apache.uima.tools.components.FileSystemCollectionReader.java was written.
- A CasConsumer.xml was written and included in the pipeline.
- A CPE descriptor, hw3-abhimank-CPE.xml, was written.

All of the above were written under src/main/resources/.

2 Task 2.

I read the tutorials from Apache UIMA before doing this task The basic concepts of UIMA-AS was read from the Apache UIMA documentation.

I first create a UIMA-AS client descriptor xml scnlp-abhimank-client.xml. This is used to connnect remote Stanford-Core-NLP service. I then imported the cleartk-stanfordcorenlp and uimaj-as-activemq dependencies to my hw3 maven project. Besides that, the client was pointed to the appropriate broker URL and end point. Then I imported the type system of the remote cleartk service and named-entity to strengthen the Answer Score annotator.

Next I Combine Named Entity annotations from the remote service. The remote service helps in annotating each word with the relevant named entity. The answer score annotator created computes an additional score by the named entities found in the question and answer. The scoring scheme proceeds as :

- It counts the named entities and its types for each question.
- It counts only those named entities in the answer that are present in the present in the question as well.

	Doc 1	Doc 2
N-gram Score	0.5	1
Named Entity Score	0.5	0.66

	Doc 1	Doc 2
w=0.6	0.25	1
w=0.4	0.5	1

- Finally it normalizes the counts of named entities to obtain the final score.

Named Entity Score vs N-gram score of HW2:

The score in hw2 was based on n-gram overlaps. The new named-entity scores are based on name-entity overlaps. The table below provides a comparison of the average precision of two documents in the two scoring systems:

We devise a final system which is a linear combination of the scores of the 2 systems:

$$final\ score = w * named\_entity\_score + (1 - w) * n\_gram\_score \quad (1)$$

For w=0.4 and w=0.6 gives us the following scores:

It appears from above that system works best when more weight is given to N-gram system.

After this I deply UIMA-AS service in my laptop. I wrote the relevant client and deployment descriptor xmls along with starting the broker. I created the hw3-abhimank-aae-as-client CPE client for testing the local service. I was able to run the CPE client and the generate the output.