# 11-791 : Design and Engineering of Intelligent Information Systems Homework 1

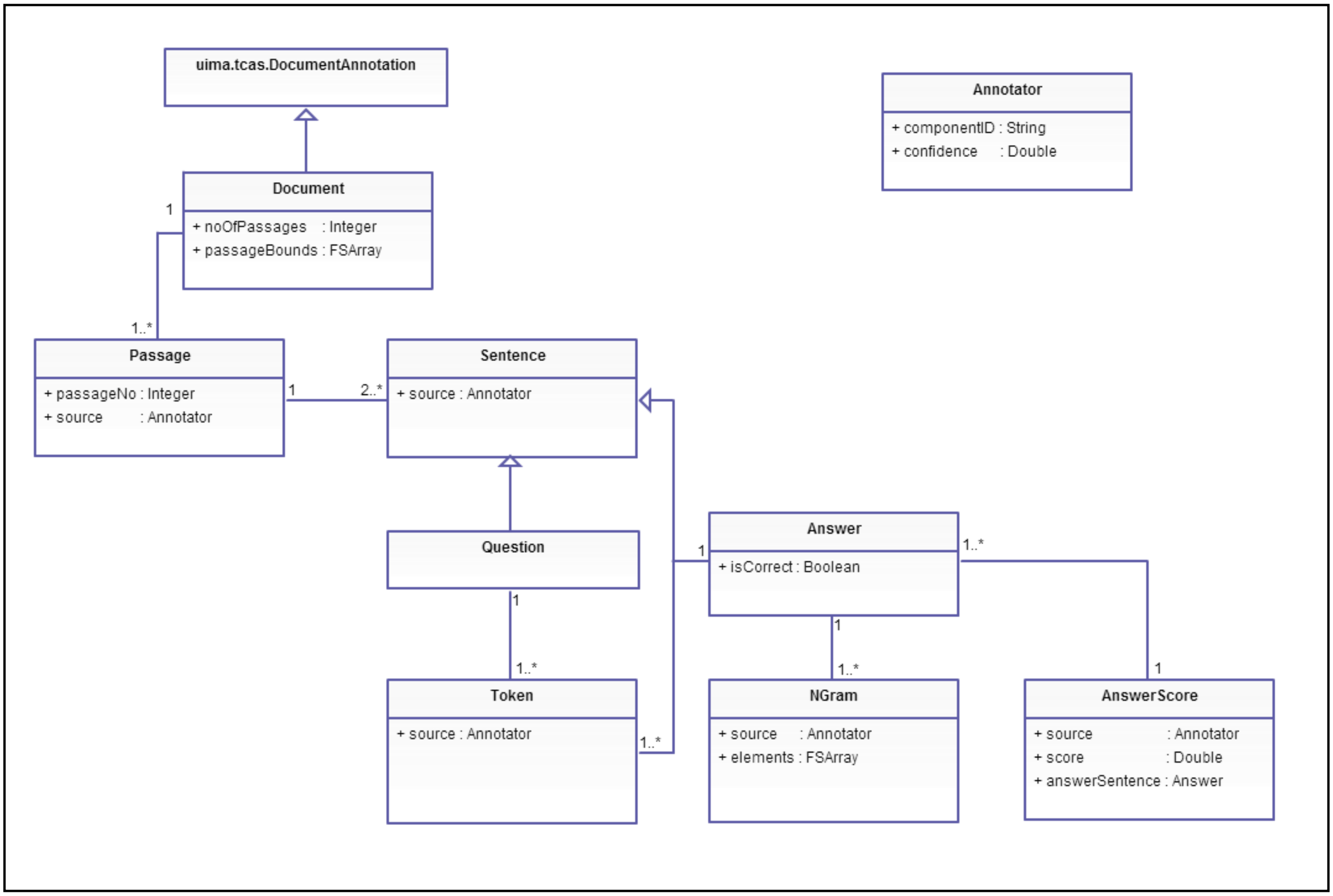
**Name** : Soumya Batra
**Andrew ID** : soumyab

## Type System Design for the Problem Statement :

We have designed a type system for a series of input document files with multiple passages, each passage having 1 question and multiple answers.

**Goal:** The goal of this system design is to enable test element annotation for the given input, allow 1-, 2- and 3- gram annotators to interact with answer candidates and tokens smoothly and finally facilitate easy evaluation of the annotator precision.;

**UML Class Diagram:** Let us have a look at the following UML class diagram to understand the system designed:



**Type System Description:** Let us look at the system design in more detail:

1. Pre-existing classes in the UIMA library – We have used 2 classes from the UIMA library as base classes to derive our custom classes:
   - **uima.tcas.Annotation** class (not shown in diagram) is used as a Base type for types Passage, Sentence, Token, NGram and AnswerScore. This class contains the **begin** and **end** indices of a test element as character offsets as well as methods to return and manipulate these values. This is a particularly useful type as it helps to know the extent of the test element within the scope and is thus the reason for being inherited by all the test element classes.
   - **uima.tcas.DocumentAnnotation** type is used as a Base type for the type Document. It consists of **language** information of the source input document alongwith **begin** and **end** indices used for measuring the extent of the input document in terms of total number of characters.

2. We have defined an **Annotator** type which consists of the following two elements:
   - **componentID** – This gives the name and description of the Annotator that annotated a test element. For eg. 1-gram annotator, answer candidate annotator, etc.
   - **confidence** – This gives the confidence value with which the Annotator has annotated a test element.

     The Annotator type is used as a feature in all test element types.

3. **Document** - This type is at the highest level of hierarchy and its objects are the one from which processing start. This type represents the input source document and extends the DocumentAnnotation type. It has the following two additional attributes:
   - **noOfPassages** - This contains the total number of passages within a document.
   - **passageBounds –** This contains an array of UIMA Annotation objects, each object containing the begin and end indices of the ith passage in the array.

4. **Passage** – This type is associated with the Document type and its elements are derived from it as follows:
   - **passageNo** – The current passage number in the document being read
   - **source** – This contains the source Annotator information alongwith the corresponding confidence value

5. **Sentence** – This type is extended from the UIMA Annotation type and its begin and end indices are set from the current passage being read. It also contains Annotator **source** information and the corresponding confidence value.

6. **Question** – This type is inherited from the Sentence type and can be seen as a redefinition for the same. This contains information of the ‘Question’ sentence in the passage.

7. **Answer** – This type is inherited from the /sentence type and consists information of the ‘Answer Candidate’ sentences in the passage. It has an additional feature **isCorrect** that determines whether an Answer Candidate is correct. A value of TRUE corresponds to an answer candidate being correct and vice-versa.

8. **Token** – This type is inherited from the UIMA Annotation type and consists the token information in a question/ answer candidate sentence. Tokens are separated at blank space or punctuation. It also contains the Annotator **source** information and the corresponding confidence value.

9. **N-Gram** – This type is inherited from the UIMA Annotation type and derives its values from the answer candidates. It consists of the following additional features:
   - **source** – This contains the source Annotator information (in this case either 1-gram, 2-gram or 3-gram annotator) alongwith corresponding confidence value for a sequence of N-gram tokens.
   - **elements** – This contains the list of all tokens encompassed by this N-Gram type. This is an FSArray of type Token.

10. **AnswerScore** – This type contains information about the final confidence score for a particular candidate answer of being the correct answer. It is inherited from the UIMA Annotation tyoe and derives its values from the answer candidates. Its additional features are:
    - **source** – This contains the source Analysis Engine information that combines the scores of all 1-gram, 2-gram and 3-gram annotators to arrive at a final score. It also consist of the corresponding confidence value.
    - **score** – This is the final score for an answer candidate. Higher the score, more the Analysis Engine estimates it to be closer to the correct answer.
    - **answerSentence** – This is the answer sentence for which the score is calculated. We need to store this information since we need to rank all the answer candidates in decreasing order of their scores after processing of all answer candidates is done for a particular question (or all sentences in a passage have been amalysed).