Homework 2 - Report

By
Vinay V Vemuri (Andrew ID: vvv)

Requirements

1. Annotate question from input text: User provides an input file with question, proposed answers and a boolean flag indicating whether each answer is correct. System reads input file and annotates questions.

2. Annotate answers from input text: User provides an input file with question, proposed answers and a boolean flag indicating whether each answer is correct. System reads input file and annotates all proposed answers. System also records whether each proposed answer is correct.

3. Annotate tokens in each question and answer: The system takes in a sentence (question or an answer) as input and annotates all tokens in the sentence.

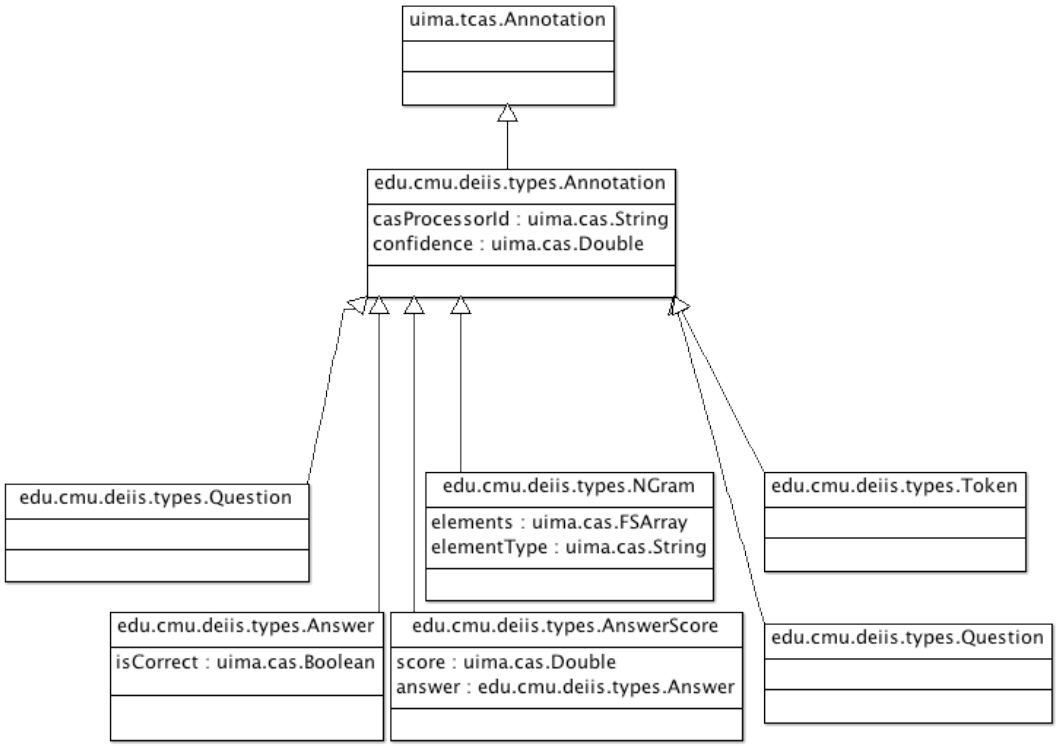
4. Annotate N-grams in each question and answer: The system takes in a sentence (question or an answer) as input and annotates all 1-grams, 2-grams and 3-grams in the sentence.

5. Assign a score to each answer: The system takes in an answer as input and assigns a score to the answer.

6. Sort answers according to scores and calculate precision: The system will take in all answers and the score assigned to each answer as input and sort the answers according to their scores. Uses the total number of correct answers and the number of predicted correct answers to compute precision.

Given Type System

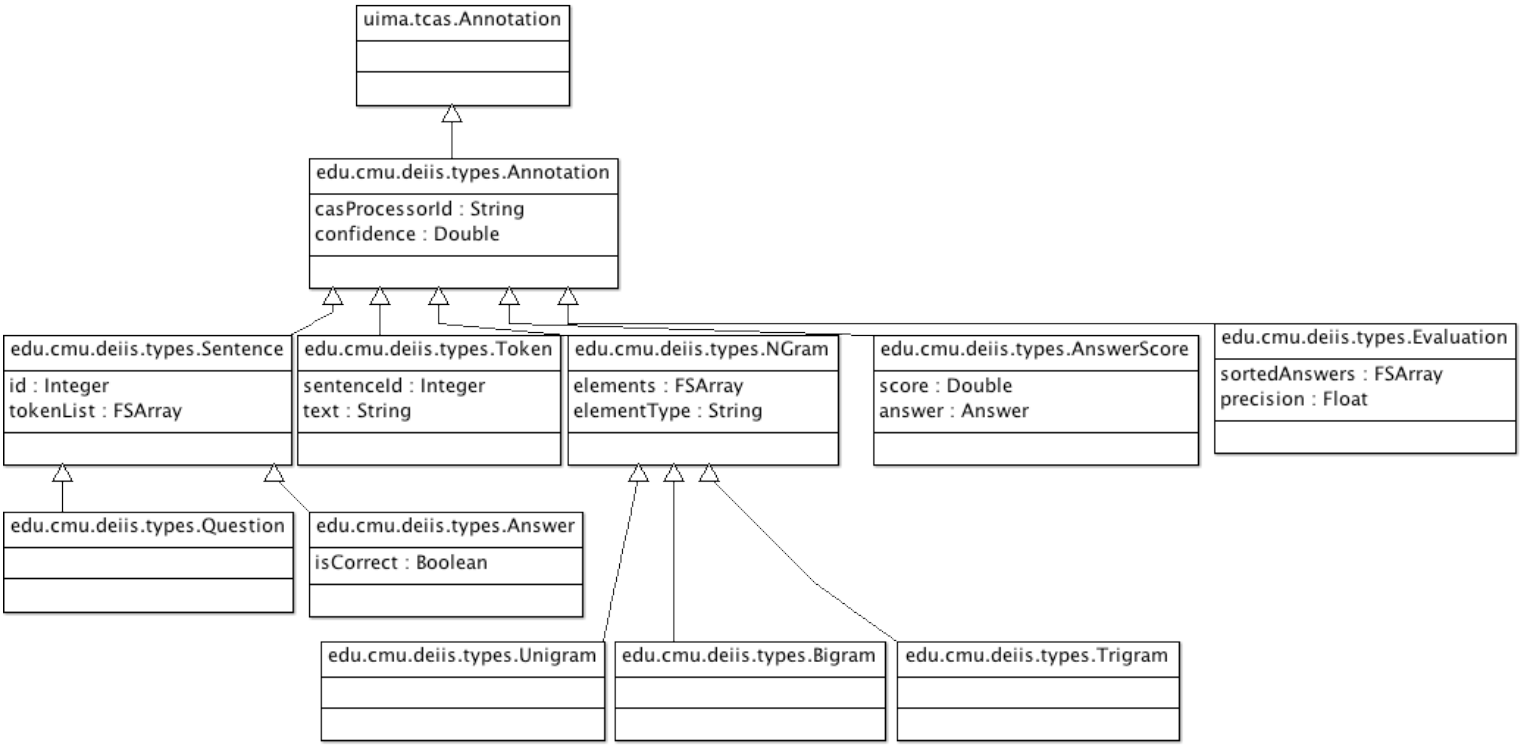
The UML Class diagram shown below provides a visualization of the given type system.



Homework 2 - Report

Improved Type-system

The following diagram shows the extended type system. Notice that no existing types from the given type system were deleted.



- **Annotation**:The Annotation type is the supertype of all the types in the proposed type system. The Annotation type inherits from the uima.tcas.Annotation type.

  Its purpose is to ensure that all subtypes have a CasProcessorId (to indicate which class made the annotation) and a confidence value (to indicate how confident the source of the annotation was).

- **Sentence**: The Sentence type is the supertype of the Question and Answer types. It captures features common to both the Question and Answer types.

  The 'Id' feature is a unique identifier assigned to each sentence in a given input file. The purpose of adding this feature was to allow each 'Token' to easily determine which sentence the Token is part of.

  The TokenList feature of the sentence type is an array of all the Tokens in the sentence. The actual sentence text is captured as a feature of each Token in the TokenList.

- **Question & Answer:** The Question type inherits from the Sentence type. For the purposes of this project, the Question type does not have any more features other than the ones inherited from Sentence. However, more fields can be added as the need arises.

  The Answer type also inherits from the Sentence type. However, unlike Question, the Answer has two other features in addition to the features inherited from Sentence. The 'IsCorrect' feature indicates whether the particular answer correctly answers the question.

Homework 2 - Report

- **AnswerScore:** The AnswerScore keeps track of the score assigned to the answer by the scoring system. It has two fields namely 'score' and 'answer'. The 'score' feature records the score assigned to the answer to allow answers to be sorted according to score during evaluation. The 'answer' feature keeps track of the answer for which the corresponding score has been recorded.

- **Token:** The Token type captures information about tokens in a Sentence (delimited by space and punctuation). It has two features namely SentenceId and Text. As mentioned previously, the 'SentenceId' is a unique identifier assigned to each sentence (and kept track of by each Token) to allow a Token to determine which sentence it is part of. The 'Text' feature captures the text contained in a token.

- **NGram:** The NGram is the supertype of Unigram, Bigram and Trigram and captures information about continuous sequences of tokens in a Sentence (Question or Answer). It consists of two features namely a list of Tokens named 'elements' of length N contained in the NGram and the type of element being stored in the elements array.

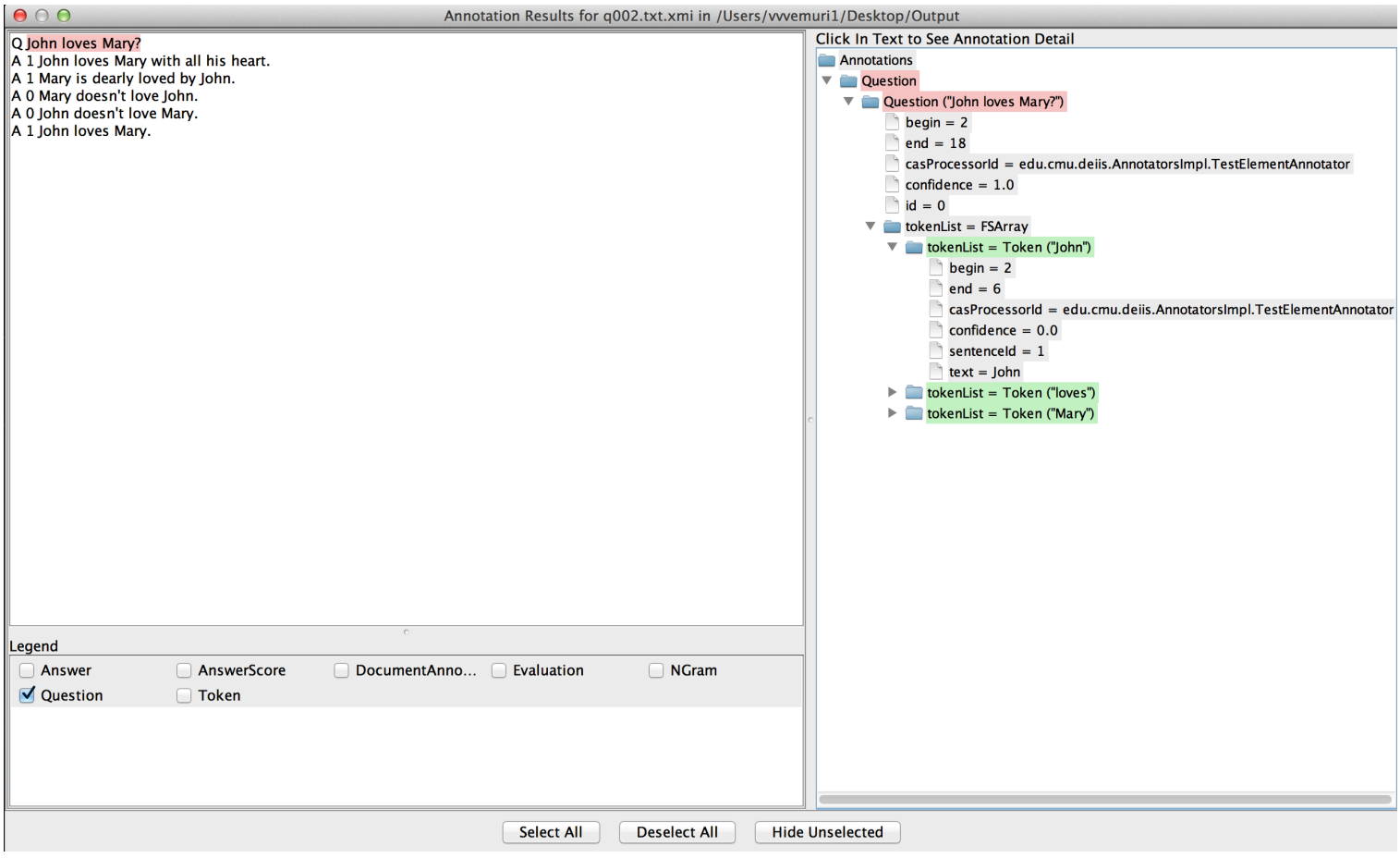
- **Unigram, Bigram & Trigram**: Unigram, Bigram and Trigram inherit from the NGram type. For the purposes of this project, these types don't have any additional fields to the ones they inherit from the NGram type. However, more fields can be added as the need arises.

- **Evaluation:** The Evaluation type captures information necessary to measure the performance of the entire system. This type consists of an array of Answers sorted by score to allow users to see which answers from input file best answer the question. This type also consists of a precision value to allow users to see the ratio of the actual number of correct answers to the predicted number of correct answers.

Screenshots

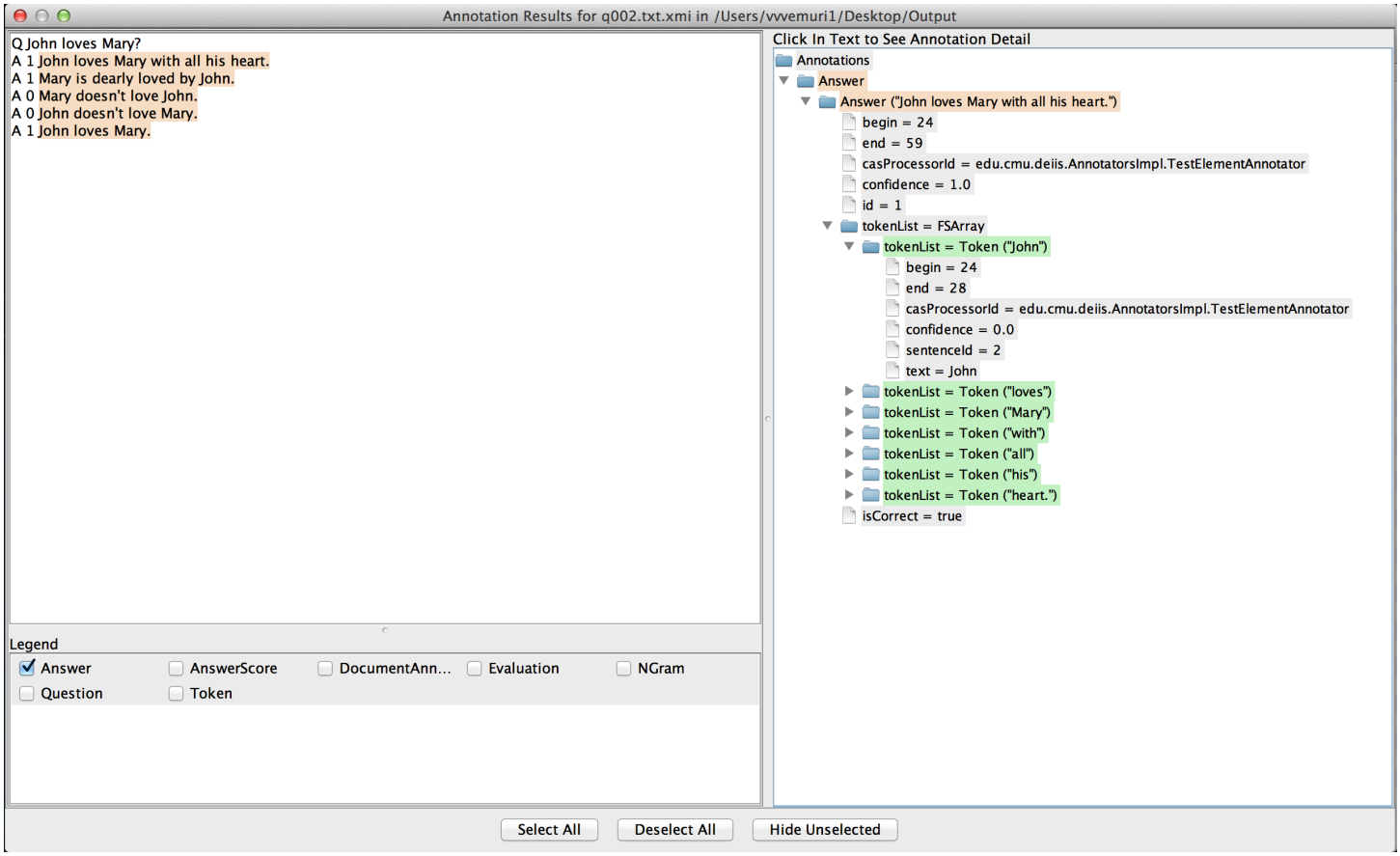
The following are screenshots of the UIMA Document Analyzer output with FullAnalysisEngine.xml.

Question

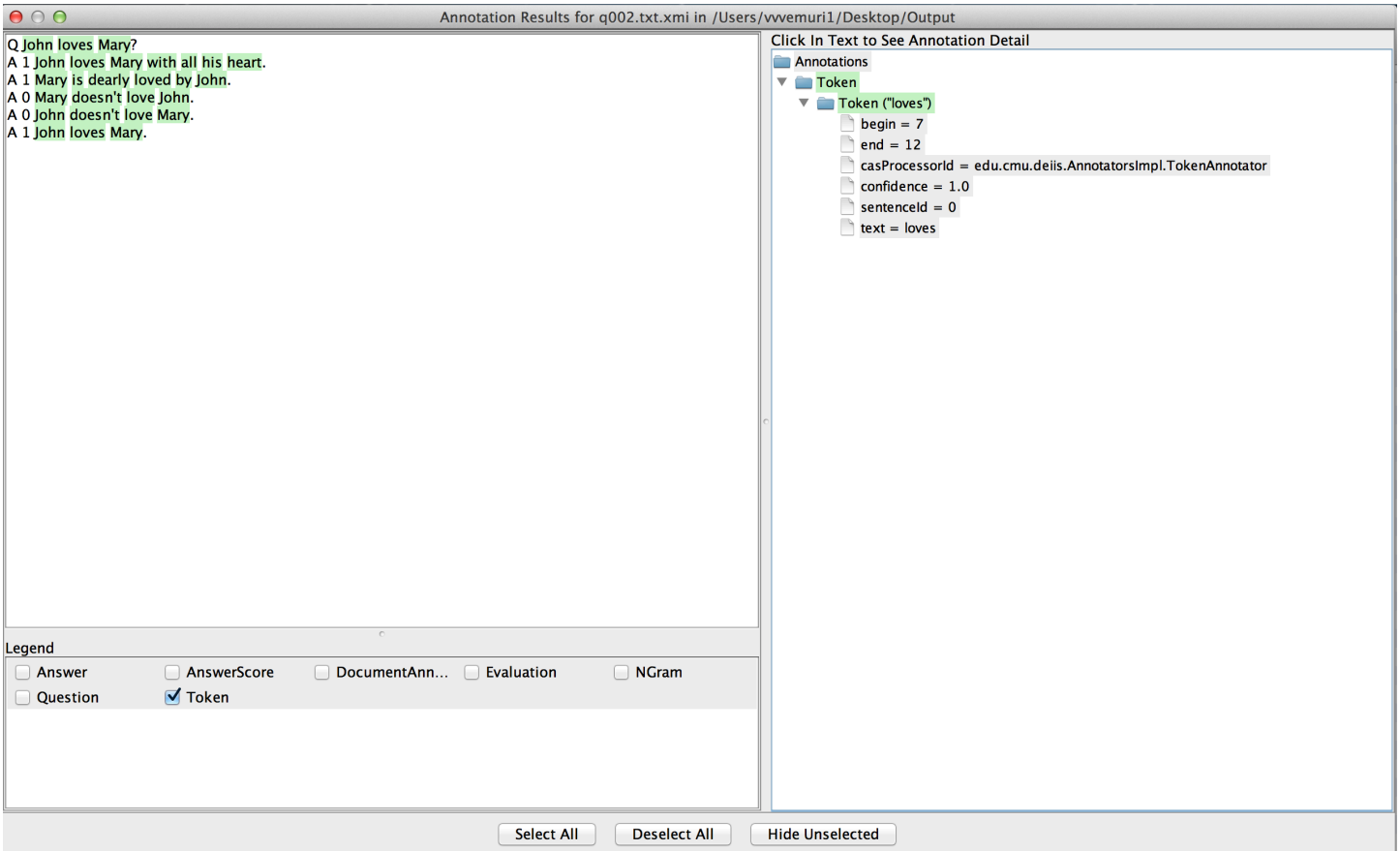


Homework 2 - Report

Answer

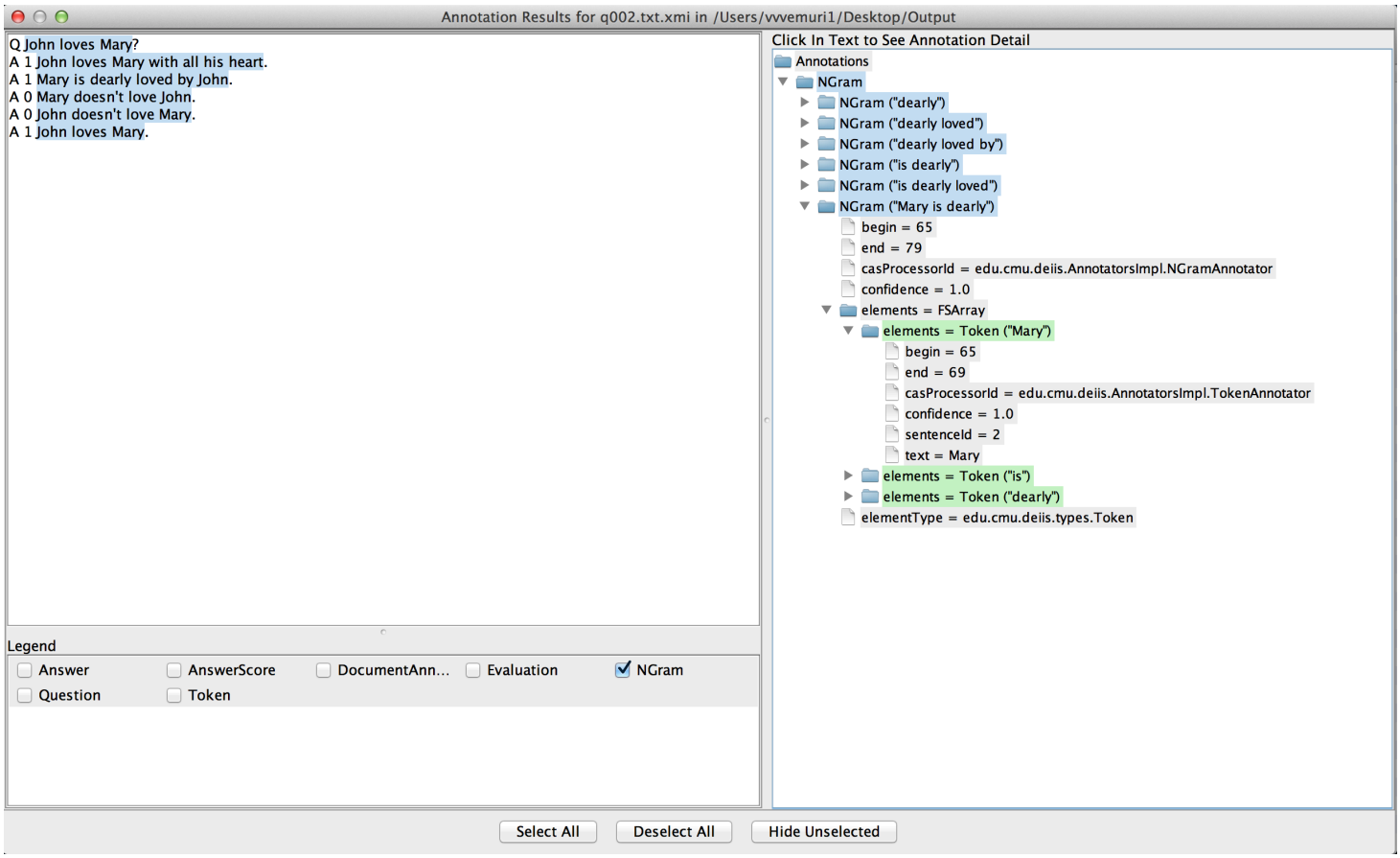


Token

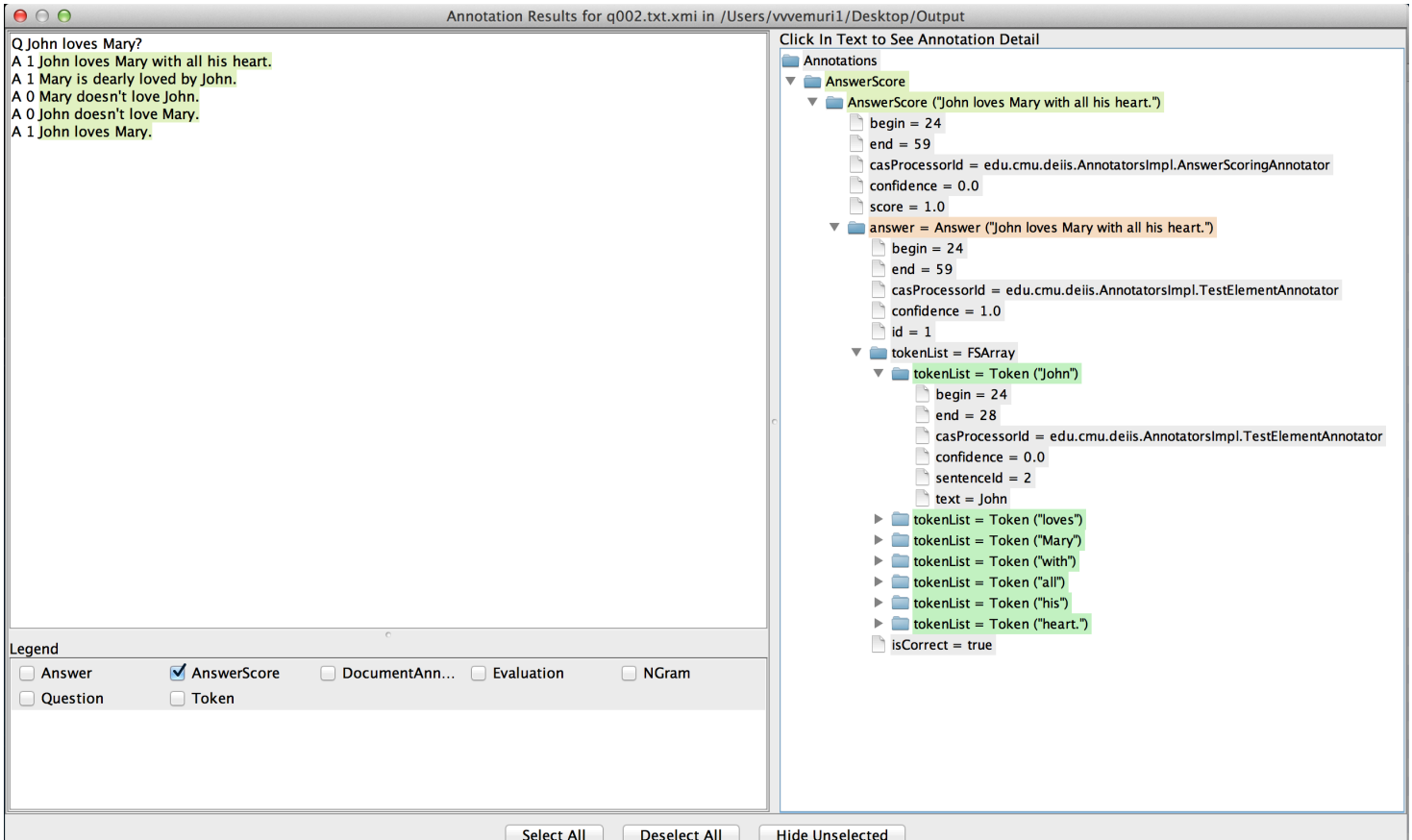


Homework 2 - Report

NGram



AnswerScore



Homework 2 - Report

Evaluation