

# 11-791 Homework 1

Wenyi Wang (Andrew ID: wenyiwan)

## 1. Introduction

My design of the logical data model is based on the task 1.3 “The Processing Pipeline”. From bottom up, the data structures are: tokens, N-gram (bigram and trigram), sentences, questions/answers.

## 2. Design

### 2.1 Tokens

Tokens are the basic element of a text file after tokenization. Its super type is “uima.tcas.Annotation”. The features are shown below.

Feature name	Range	Description
Postag	uima.cas.String	POS tagging information
WordCount	uima.cas.Integer	# of words in the token
Begin	uima.cas.Integer	The beginning of the token
End	uima.cas.Integer	The end of the token

Table 1. Features of tokens

### 2.2 N-grams

N-grams are a higher level than tokens. Only bigrams and trigrams are included for the purpose of being concise. The super types of them both “uima.tcas.Annotation”. The features of bigrams and trigrams are shown below.

Feature name	Range	Element Type	Description
TokenList	uima.cas.FSArray	Token	List of tokens in the bigram
Begin	uima.cas.Integer	N/A	The beginning of the bigram
End	uima.cas.Integer	N/A	The end of the bigram

Table 2. Features of bigrams

Feature name	Range	Element Type	Description
--------------	-------	--------------	-------------

TokenList	uima.cas.FSArray	Token	List of tokens in the trigram
Begin	uima.cas.Integer	N/A	The beginning of the trigram
End	uima.cas.Integer	N/A	The end of the trigram

Table 3. Features of trigrams

### 2.3 Sentence

Sentence is a higher level than n-grams. In addition to beginning and end of the sentence, the features also include lists of tokens and n-grams. Moreover, any other words or elements that don't belong to tokens or n-grams are included as well. The features of sentence are shown below. The super type if sentence is “uima.tcas.Annotation”.

Feature name	Range	Element Type	Description
TokenList	uima.cas.FSArray	Token	List of tokens in the sentence
BigramList	uima.cas.FSArray	Bigram	List of bigrams in the sentence
TrigramList	uima.cas.FSArray	Trigram	List of trigrams in the sentence
Other	uima.cas.String	N/A	Any other elements in the sentence
Begin	uima.cas.Integer	N/A	The beginning of the sentence
End	uima.cas.Integer	N/A	The end of the sentence

Table 4. Features of sentence

### 2.4 Question and Answer

Question and answer are higher than sentence. The super types of them are both “sentence” defined previously. The features of them are shown below.

Feature name	Range	Description
TypeQ	uima.cas.String	Type of questions
Begin	uima.cas.Integer	The beginning of the question
End	uima.cas.Integer	The end of the question

Table 5. Features of question

Feature name	Range	Description
--------------	-------	-------------

Confidence	uima.cas.Double	The confidence/score of the answer
IsCorrect	uima.cas.Boolean	The tag to tell if the answer is correct
Begin	uima.cas.Integer	The beginning of the answer
End	uima.cas.Integer	The end of the answer

Table 6. Features of answer

### 3. Summary

The design is concise and kind of naïve.