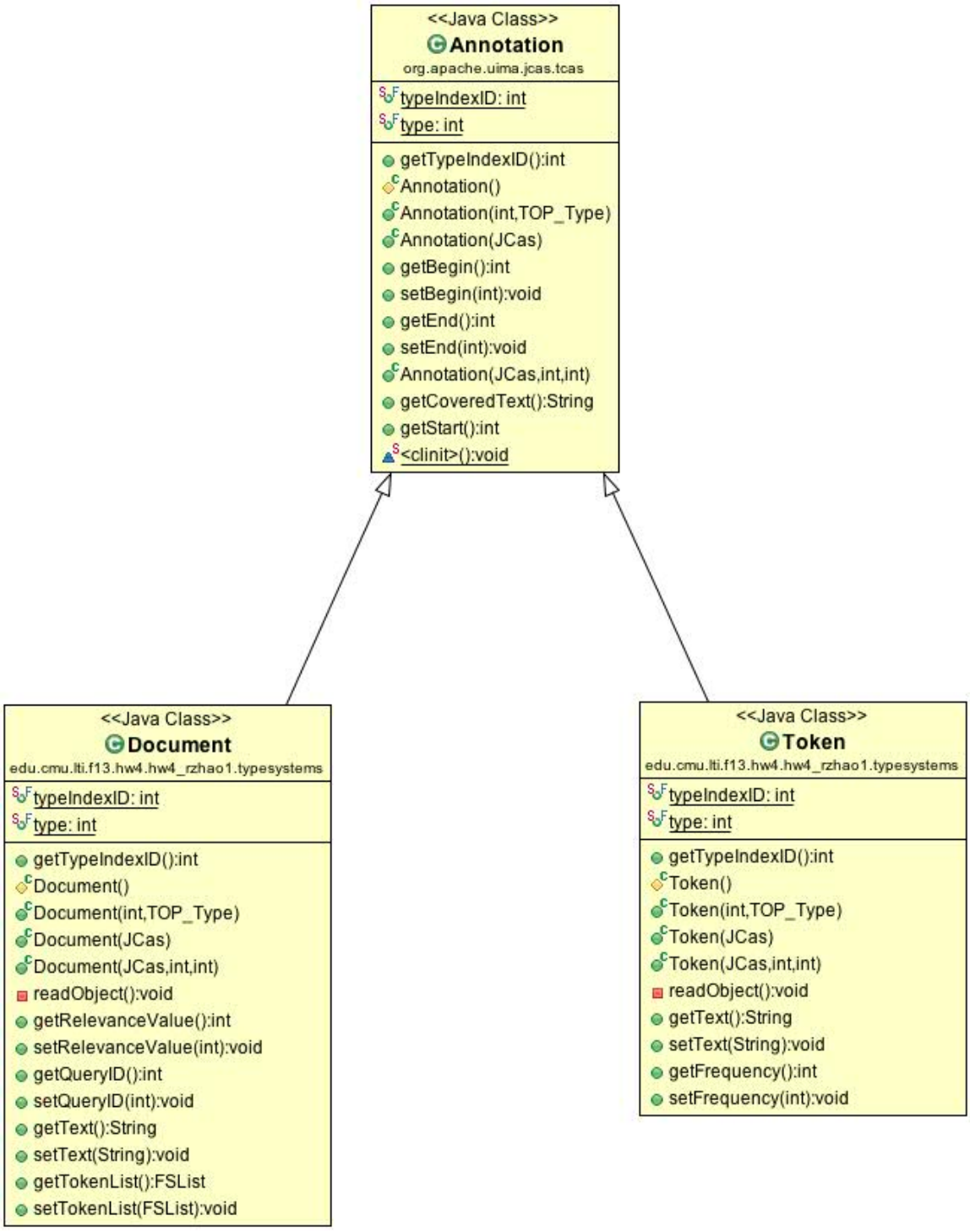
Homework4 Report

Ran Zhao ID: rzhao1

Part I. Type system



I am using the provided type system. Document annotation is used to store the Relevance Value, Query ID, Text String and Token list. Token annotation is used to store the token string and its frequency in the sentence.

Part II. Retrieval System design

Firstly, after documentreader read all the sentences into cas, I deploy lingPipe tokenizer to make the tokenlist for each sentence.

Secondly, I will calculate the frequency for each token in a sentence.

Thirdly, in retrieval evaluator analysis engine, I create four parallel lists, one for question id, one for relevance value, one for text string, the other one for token list. The order of this four lists are fixed and never be changed. Thus, I calculate the similarity of each answer candidate using these four list.

Fourthly, I start to calculate the cosine similarity of each answer candidate. Each answer is a vector, which represents the frequency of each words. The method of calculating the cosine similarity is followed:

$$Similarity=Sum(FQ(word\_i)*FA(word\_i))/(|VQ|*|VA|)$$

FQ(word_i): The frequency of word_i in question.
FA(word_i): The frequency of word_i in answer.
VQ: 1-norm of question vector.
VA:1-norm of answer vector.

Fifthly, after I calculate the cosine similarity of all the answer candidates. I will rank them based on the score of cosine similarity. Then I select the first correct answer based on the its relevance value and store its rank.

Lastly, as I have all the first correct answer ranks, I will calculate the value of MRR based on the equation provided in the homework instruction.

Part III. Experiment Output

Score:0.45226701686664544 rank=1 rel=1 qid=1 Classical music may never be the most popular music

Score:0.30618621784789724 rank=1 rel=1 qid=2 Climate change and energy use are two sides of the same coin.

Score:0.5070925528371099 rank=1 rel=1 qid=3 The best mirror is an old friend

Score:0.2581988897471611 rank=3 rel=1 qid=4 If you see a friend without a smile, give him one of yours

Score:0.15811388300841897 rank=1 rel=1 qid=5 Old friends are best

(MRR) Mean Reciprocal Rank ::0.8
Total time taken: 0.795

Part IV. Error Analysis

Firstly, I did a basic optimization by making the string comparison case insensitive. However, I notice that the result of question 4 is not satisfactory because the system did not recognize friends and friend are from the same word root. In order to further improve the performance of the retrieval system, I consider to stemming and lemmatizing each token in the sentence.