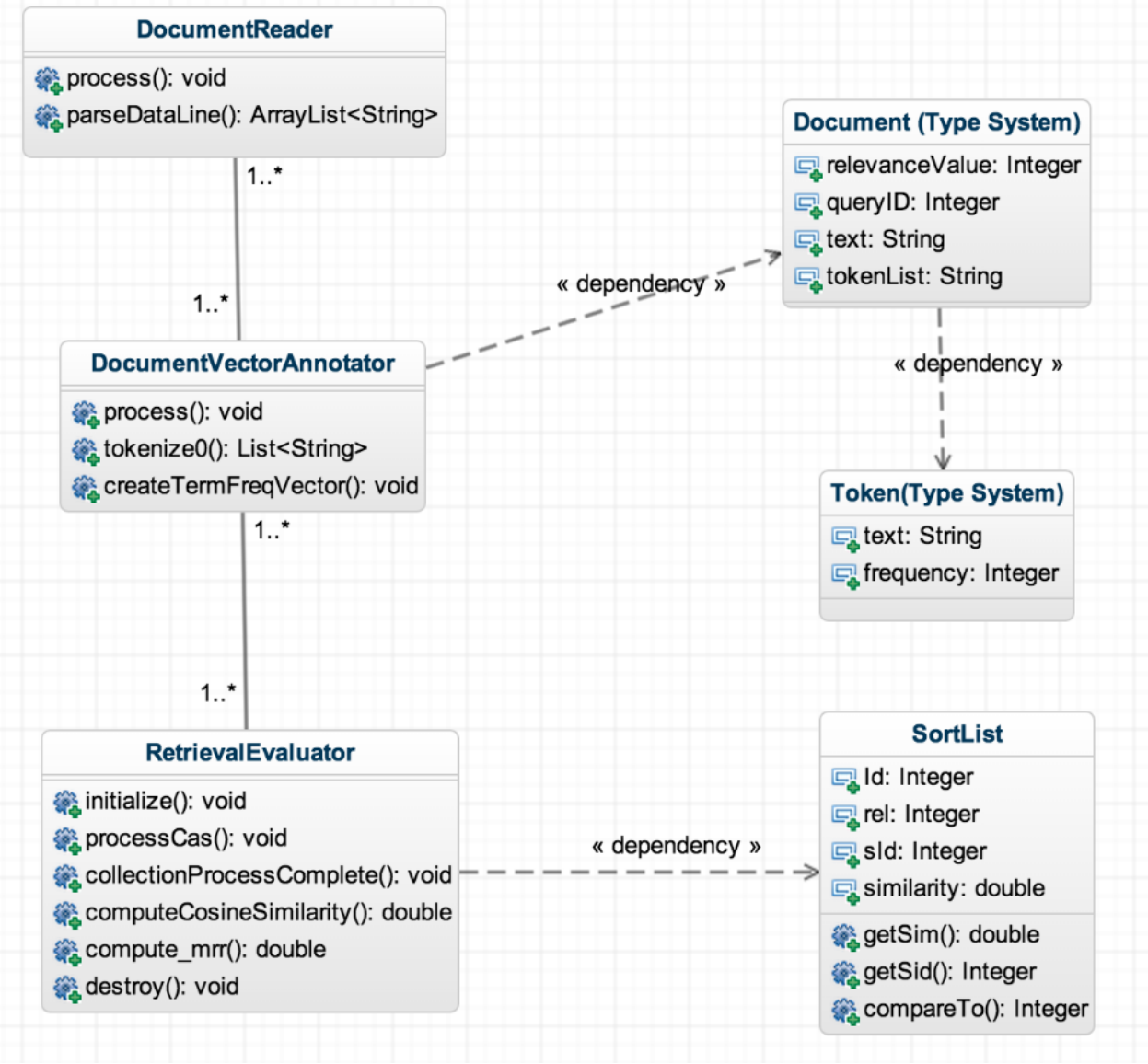
Hongfei Zhang
Andrew ID: hongfeiz
October 21, 2014

HW3 Report

TASK 1

1. System Design



I used most of the archetype and made some modification based on that.

The Document Reader extracts the query id and relevance, and put them with sentence together into a Document type.

The DocumentVectorAnnotator computes word and frequency in each sentence, and sends them into a Document type.

The RetrievalEvaluator extracts features from JCas's Document type, and computes similarities and mean reciprocal rank.

2. Document Vector Annotator Design

In DocumentVectorAnnotator, I used the given white-space based tokenize0. It splits every sentence to make up a word vector and update the tokenList in the Document Type. Method Utils.fromCollectionToFSList() is used in the process.

3. Retrieval Evaluator Annotator Design

3.1 initialize()

In the method, the parameter OutputFile defined in Cas consumer descriptor is assigned to locate the output file.

3.2 processCas()

It extracts Document type from Cas and get features.

3.3 collectionProcessComplete()

It is the major function in Cas Consumer. It extracts all sentences in document and divides them into two parts: query and answer, where I use hash map to store these data.

computeCosineSimilarity() and compute_mrr() are called to compute similarities and MMR value.

To sort the similarities, I used Collections.sort() method, and overrode the compareTo() rule for sorting only similarity. To fulfill it, I created a class SortList and declared these information in it.

3.4 computeCosineSimilarity()

Using definition of cosine similarity, i implemented the computation.

3.5 compute_mrr()

Implemented the mean reciprocal rank computation.

4. Result

Result of the task 1 pipeline that processing the document.txt of 20 queries is MMR=0.4375. It is given in ~/report.txt

TASK 2

1. Error Analysis

There are several different error types that I defined, which are vocabulary mismatch, redundancy and irrelevant vocabulary.

Error Types	Query Id	Count
vocabulary mismatch	1,2,3,4,6,8,11,12,15,16,17,18	12
redundancy	1,2,4,5,7,8,12,13,14,15,17,18,19,20	14
irrelevant vocabulary	1,2,3,4,5,6,7,8,11,13,14,15,16,17,18,20	16

1.1 Vocabulary mismatch happens when words of the answer has different form with the question, such as singularity vs. plurality, different tense (past, present or future), different voice (active and passive), possessive case, punctuation (which is caused by tokenize0 only splitting on space), uppercase letter and lowercase letter.

1.2 Redundancy happens when the answer matches the question, but contains too many words, which decreases its cosine similarity.

1.3 Irrelevant vocabulary happens when the answer contains irrelevant words, such as quantifier, preposition, link verb.

2. Performance Improvement

2.1 To Vocabulary Mismatch

2.1.1 Stemming Algorithm

Using the StanfordLemmatizer.java in package /edu/cmu/lti/f14/hw3/hw3_hongfeiz/utils. A class Morphology() is applied in the method stemWord() and changes forms of words. It extracts, as the testing result shows, stems of words. Using the stemming tool into my pipeline, MMR increased to 0.5500.

2.1.2 Tokenization Algorithm

In DocumentVectorAnnotator.java, I made some modification with the tokenize0() method, which deletes punctuations in sentences. MMR increased to 0.4923.

2.2 To Irrelevant Vocabulary

There is a file stopwords.txt in package /src/main/resources/. It removed all stop words like me,more,most,mustn't,my,myself,no,etc. in sentences, which are mostly irrelevant with intelligent Q&A system. After using the file, MMR increased to 0.4750.

3. Other Similarity Functions

3.1 Jaccard coefficient method

The Jaccard index, also known as the Jaccard similarity coefficient (originally coined coefficient de communauté by Paul Jaccard), is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

Using the algorithm, the MRR increased to 0.4417, as follows.

```
cosine=0.3333 rank=1 qid=14 rel=1 Lionel Richie was lead singer and songwriter for Commodores.
cosine=0.0385 rank=3 qid=15 rel=1 A new look at NASA satellite data revealed that Earth set a new record for coldest temperature r
cosine=0.1538 rank=3 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
cosine=0.0690 rank=3 qid=17 rel=1 Corn futures found support from forecasts for above-normal   temperatures in major growing areas
cosine=0.0278 rank=2 qid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that
cosine=0.0556 rank=3 qid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and crashed in Lakehurs
cosine=0.1429 rank=2 qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living
(MRR) Mean Reciprocal Rank ::0.4417
```

3.2 Dice coefficient method

It can be viewed as a similarity measure over sets:

$$s = \frac{2|X \cap Y|}{|X| + |Y|}$$

Similarly to Jaccard, the set operations can be expressed in terms of vector operations over binary vectors A and B:

$$s_v = \frac{2|A \cdot B|}{|A|^2 + |B|^2}$$

which gives the same outcome over binary vectors and also gives a more general similarity metric over vectors in general terms.

Using the algorithm, the MRR increased to 0.4417, as follows.

```
cosine=0.0385 rank=3 qid=15 rel=1 A new look at NASA satellite data revealed that Earth set a new record for coldest temperature r
cosine=0.1538 rank=3 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
cosine=0.0690 rank=3 qid=17 rel=1 Corn futures found support from forecasts for above-normal   temperatures in major growing areas
cosine=0.0278 rank=2 qid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that
cosine=0.0556 rank=3 qid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and crashed in Lakehurs
cosine=0.1429 rank=2 qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living
(MRR) Mean Reciprocal Rank ::0.4417
```