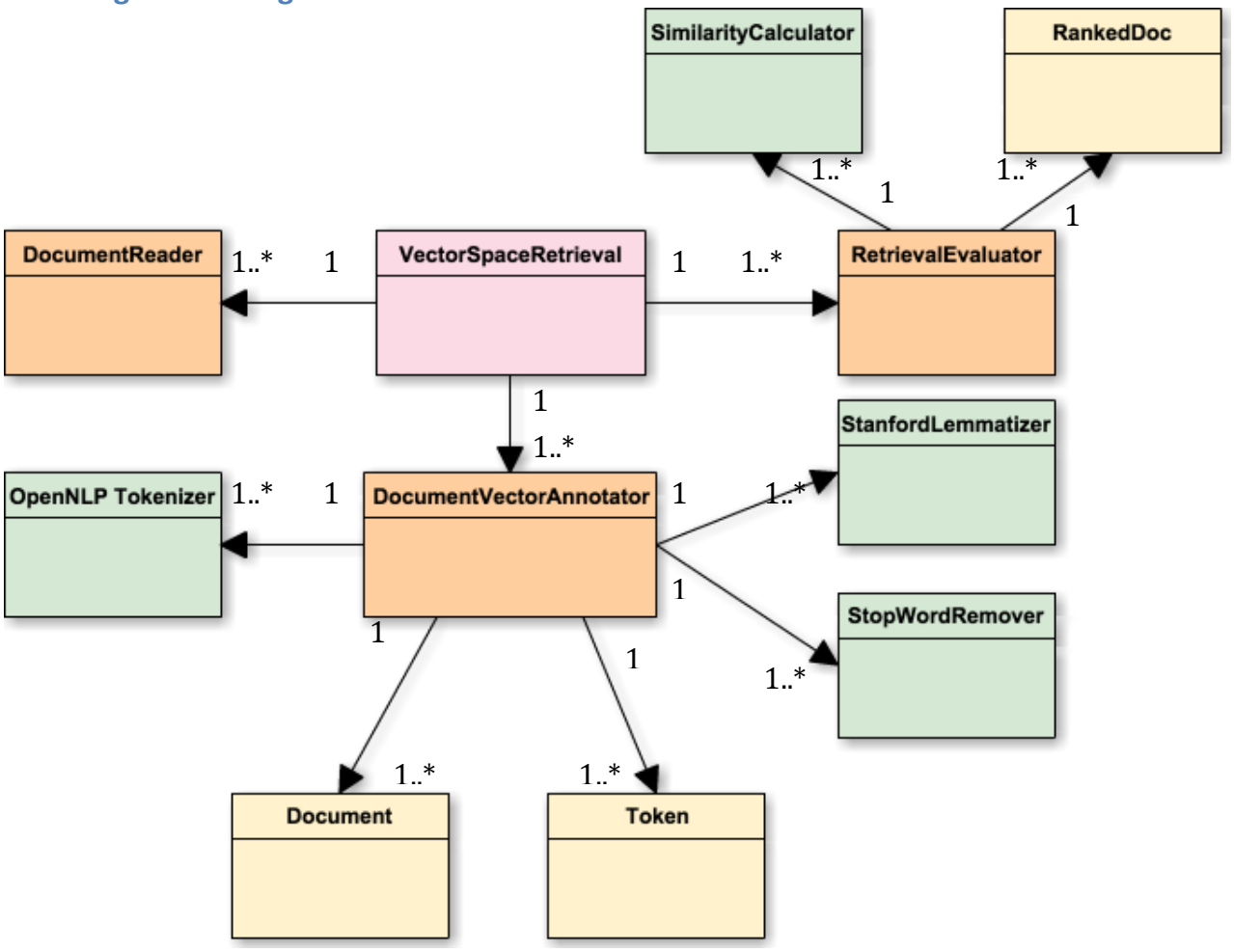
Homework 4
Report

Name: LIU Xi
AndrewID: xiliu1
Date: 19 Oct 2014

1. System Overview

1.1 Design Class Diagram



The UML diagram above demonstrates an overview of the system. The VectorSpaceRetrieval acts as a central manager, coordinating among the DocumentReader, DocumentVectorAnnotator, and RetrievalEvalutator, which are the controllers. The DocumentVectorAnnotator manipulates Documents and Tokens, and the RetrievedEvaluator manipulates RankedDocs. The OpenNLP Tokenizer, StanfordLemmatizer, and SimilarityCalculator serve as utilities.

I kept the original design for UIMA pipeline, and implemented several new features to the system.

Firstly, I designed RankedDoc to facilitate ranking the answers by similarity.
Secondly, I implemented SimilarityCalculators as an API for similarity computation, such as Cosine Similarity, Dice Similarity, and Jaccard Similarity.

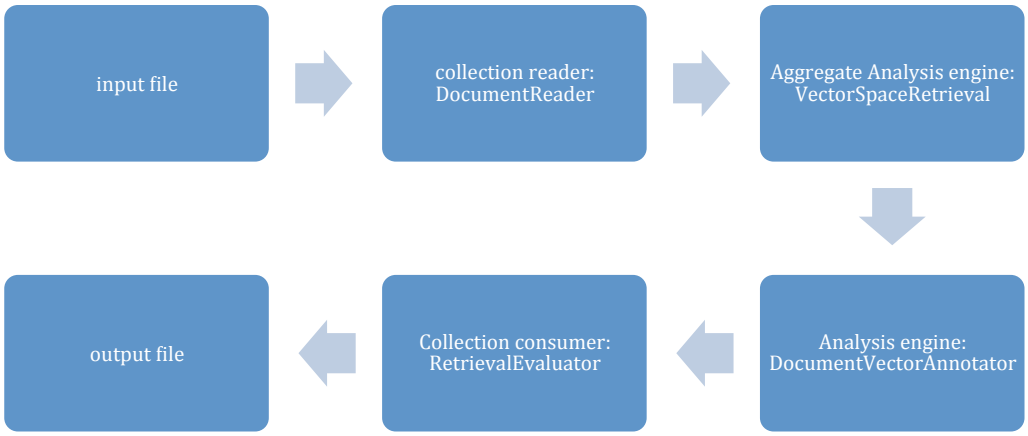
Thirdly, I built StopWordRemover to remove the punctuations and stopwords.

Fourthly, I used OpenNLPTokenizer and OpenNLP tokenization model as a new way to tokenize the sentences.

Finally, I designed several experiments, which will be described in section 2. The code implementations could be found in DocumentVectorAnnotatorForTask2 and RetrievalEvalutatorForTask2.

To reproduce the experiment result, you could change the java class path in descriptors, then remove the /**/ for each experiment.

1.2 Information Flow



2. Error Analysis

2.1 Mismatch Types

Mismatch Type	Problematic Instances
Variation	Q: In which year did a purchase of Alaska happen? A: William Seward negotiated a purchase of Alaska for $7.2 million. Actual Answer: Alaska was purchased from Russia in year 1867.
Conjunctions	Q: Where was the first McDonald's built? A: McDonald's Corporation is the world's largest chain of hamburger fast food restaurants. Actual Answer: From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that the McDonald brothers developed to offer customers reasonably priced food at a rapid pace formed the cornerstone of the fast-food business.

Special Term Mismatch	Question: Who was the lead singer for the Commodores Answers: The Commodores originally came together from groups the Mystics and the Jays. Actual Answers: Lionel Richiewas was lead singer and songwriter for Commodores.
Question-answer Mismatch	Q: What is the Keystone State? A: Keystone Resort is the largest ski resort in Summit County located in Keystone Colorado. Actual Answer: They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to its name.

2.2.1 Vocabulary Mismatch

Variations

Verb "visit" could appear as "visits", "visited", "was visited", and "visiting", while noun "apple" and "tomato" could appear as "apples" and "tomatoes". However, during the tokenization process, these words were separated by spaces and stop words, so that the variations could not be recognized as one term. The naïve tokenization method affects the frequency and further reduced the value of cosine similarity, therefore other documents ranked higher.

Conjunctions

Some words were joint in some structures, such as "isn't", "I'm", "doesn't", and "David's", etc. These words could not be recognized as the "is not", "I am", "does not", and "sth of David" due to the limitation of tokenization, so that the similarity is lower than the actual value.
For example, a sentence talking about "China's mother river" was selected to answer the question on "China's Sorrow".

2.2.2 Special Term Mismatch

Some phrases have only special meanings, such as "United States", "Michael Jordan", "Michael Jackson", and "Hidden Markov Model". These phrases were broken into pieces instead of being recognized as one entity. If candidate documents contains single words, not the phrases, those single words were counted for the similarity computation, and lead to selecting an irrelevant document.
For example, a sentence talking about "Michael Jackson" was selected to answer the question on "Michael Jordan".

2.2.3 Question –answer mismatch

Some documents have very high similarity with the questions, however it does not answer the questions. For example, a document talking about the date of Hindenburg disaster was selected to answer the question on the location of the disaster.

3. Solution

3.1 Statistical Significant Improvement

Stop-word Remover	Stanford Lemmatizer	OpenNLP Tokenizer	Jaccard Similarity	Dice Similarity	MRR Increase
X	X	O	X	X	11.43%
X	O	X	X	X	22.86%
X	X	X	O	X	25.71%
O	X	X	X	X	27.61%
O	O	X	X	X	27.61%
O	X	O	X	O	27.61%
X	O	O	X	O	33.33%
X	O	X	O	X	36.18%
X	**O**	**O**	**O**	**X**	**40.00%**

3.2 Solution Implementation

3.2.1 Vocabulary Mismatch

To reduce vocabulary mismatch, I implemented the following features to firstly convert the conjunctions and variations to the original words, secondly remove the punctuations, and finally lowercase the words. This process was achieved by the following approach.

Document Pre-processing

Regular Expression and Lowercase

I implemented regular expressions to remove the punctuations and convert all tokens to lowercases. However, the MRR remains unchanged.

Stop-word Remover

Then I tried to add the punctuations to stopword.txt and implement a stop-word remover to remove the stop words as well as punctuations. This time the MRR increased from *0.4375* to *0.5583*.

Also, the problem caused by conjunctions is solved.

Mismatch Type	Problematic Instances	Problem Solved?
Conjuction	Q: Where was the first McDonald's built? A: From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that the McDonald brothers developed to offer customers reasonably priced food at a rapid pace formed the cornerstone of the fast-food business.	YES

Stanford Lemmatizer

Stanford Lemmatizer is for converting the varied words to its original form. I implemented the StemText before tokenize the sentences. The MRR increased from *0.4375* to *0.5375*.
Also, the problem for variation mismatch is solved.

Mismatch Type	Problematic Instances	Problem Solved?
Variation	Q: In which year did a purchase of Alaska happen? A: Alaska was purchased from Russia in year 1867.	YES

Combining with the stop-word remover, the MRR is *0.5583*.

3.2.2 Special Term Mismatch

OpenNLP Tokenizer

I downloaded OpenNLP Tokenizer and related model from OpenNLP official website. OpenNLP Tokenizer could perfectly replace the conjunctions with phrases, for example, "won't" is replaced with "will not". Besides, OpenNLP deals with misspellings and special terms.
Implementing OpenNLP Tokenizer improved the MRR to *0.4875*.
Combining with stop-word remover the MRR is *0.5583*, while combining with StanfordLemmatizer the MRR is *0.5833*.
What is more, the chance for mismatch caused by special terms is reduced.

Mismatch Type	Problematic Instances	Problem Solved?
Special Term Mismatch	Question: Who was the lead singer for the Commodores Answers: The Commodores originally came together from groups the Mystics and the Jays. Actual Answers: Lionel Richiewas was lead singer and songwriter for Commodores.	YES

3.2.3 Question-answer Mismatch

Jaccard Similarity

The definition of Jaccard similarity is based on the set theory in discrete math.

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

Replacing Cosine Similarity with Jaccard Similarity improved the MRR to *0.55*.
Combining Jaccard with StanfordLemmatizer, the MRR is *0.5958*.
Combining Jaccard with StanfordLemmatizer and OpenNLPTokenizer, the MRR increased to *0.6125*.
What is more, the chance for Question-Answer Mismatch is reduced.

Mismatch Type	Problematic Instances	Problem Solved?
Question-answer Mismatch	Q: What is the Keystone State? A: They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to its name.	YES

Dice Similarity

The definition of Jaccard similarity is based on the set theory in discrete math too.

$$QS = \frac{2C}{A+B} = \frac{2|A \cap B|}{|A| + |B|}$$

Replacing Cosine Similarity with Dice Similarity, the MRR is *0.4417*.
Combining Dice with StanfordLemmatizer, the MRR is *0.55*.
Combining Dice with StanfordLemmatizer and OpenNLPTokenizer, the MRR increased to *0.5667*.

4. Future Enhancements

The project is constrained by time and resource.
First of all, due to the small dataset, the evaluations may not fully reflect the performance of the system. Secondly, with larger datasets, more time and resources given, we could introduce more advanced algorithms and machine learning models to this system.

I have some sketches for future enhancements.

Spelling Corrector

Some mismatches are caused by misspellings, for example, the tokenizer regards “banana” and “banna” as two different words, so that the similarity might be lower than expected.
A spelling corrector could be implemented to correct the misspellings in the document. With all correctly spelled words, the precision could be improved. A simple spelling corrector could be implemented with a dictionary built on the data structure Tries.
Lingpipe also provides resources for building a spelling corrector, with domain sensitivity and context-sensitive corrections.

BM25

$$score(q,d) = \sum_{i=1}^{|q|} idf(q_i) \cdot \frac{tf(q_i,d) \cdot (k_1+1)}{tf(q_i,d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{avgdl}}$$

$$idf(q_i) = \log \frac{N - df(q_i) + 0.5}{df(q_i) + 0.5}$$

Based on a retrieval framework, BM25 is a ranking function used by search engines to rank matching documents according to relevance to a given query. Given larger datasets, BM25 could hopefully produce better predictions on the relevance than the simple similarity computations implemented in the current system.

OpenNLP NameFinder

To reduce the question-answer mismatch, we could firstly look into the questions to understand the what kind of answer the query is asking for. Then we explore the candidate for answers, and verify if it contains such types. For example, “In which year did a purchase of Alaska happen?” is

asking for a 4-digit numerical answer to represent year. We explore the answers, and if it contains tokens like “1954”, it is more likely to be the correct answer. OpenNLP NameFinder could solve the problem.
However, 1954 is also possible a room number, student id, or street address, so this idea could be further developed combining with Lingpipe, to achieve context-sensitive matching.