HW #4 Report: Vector Space Retrieval Model

Version: 0.1

Fall 2013 11-791

10/27/2013

Jeffrey Gee

Contents

1 Introduction

1.1 Purpose

The purpose of this document is to describe the contents of Assignment #4.

1.2 Terminology

The following is a list of unobvious terms/acronyms that may be used throughout this document.

Word/Acronym	Definition
AAE	Aggregate Analysis Engine
BOW	Bag of Words
CPE	Collection Processing Engine
MRR	Mean Reciprocal Rank
TSV	Tab Separated Values

1.3 Assignment Tasks

The following tasks were accomplished in Assignment #4.

1. Extract bag of words feature vectors from input documents.
2. Compute cosine similarity between query and answer pairs.
3. Compute the Mean Reciprocal Rank (MRR) for all queries in the text collection.
4. Use Error Analysis to improve the MRR performance measure.
5. Improve the efficiency of the program by doing error analysis of the retrieval system.

1.4 Assumptions

The following assumptions were made based on the input data.

```
qid=1   rel=99 Classical music is dying
qid=1   rel=0  Pop music has absorbed influences from most other genres of popular music
qid=1   rel=1  Classical music may never be the most popular music
qid=1   rel=0  Everybody knows classical music when they hear it
qid=2   rel=99 Energy plays an important role in climate change
qid=2   rel=0  Old wine and friends improve with age
qid=2   rel=0  With clothes the new are the best, with friends the old are the best
qid=2   rel=1  Climate change and energy use are two sides of the same coin.
qid=3   rel=99 One's best friend is oneself
qid=3   rel=1  The best mirror is an old friend
qid=3   rel=0  My best friend is the one who brings out the best in me
qid=3   rel=0  The best antiques are old friends
qid=4   rel=99 The shortest distance between new friends is a smile
qid=4   rel=0  Wear a smile and have friends; wear a scowl and have wrinkles
qid=4   rel=1  If you see a friend without a smile, give him one of yours
qid=4   rel=0  Behind every girls smile is a best friend who put it there
qid=5   rel=99 It takes a long time to grow an old friend
qid=5   rel=0  Old wine and friends improve with age
qid=5   rel=0  With clothes the new are the best, with friends the old are the best
qid=5   rel=1  Old friends are best
```

1. Proper order is not guaranteed. Although the test data is pre-ordered, the solution to the problem should be able to run whether a query comes after corresponding answers or before corresponding answers.
2. Although our test data has a single correct answer for each query, we should build a solution that works in cases where there may be multiple correct answers. In these situations, the MRR score will be based on the correct answer for queries with the highest rank.

2 Vector Space Retrieval Model

2.1 Introduction

The Vector Space Retrieval Model uses vector-based similarity measures to rank documents. A template Aggregate Analysis Engine was provided as part of the assignment and consists of three basic parts:

- Document Reader
- Document Vector Annotator
- Retrieval Evaluator

2.2 Type System

The following type system was provided as part of the assignment and was sufficient for completing all tasks:

Annotation Type	Members	Types	Comment
Document	relevanceValue	Integer	Denotes whether a sentence is a query, correct answer, or wrong answer.
	queryID	Integer	Denotes what query a sentence belongs to.
	text	String	Stores the text of the sentence.
	tokenList	FSList<Token>	Stores the token list for the sentence
Token	text	String	Stores the text for a token in a sentence
	frequency	Integer	Stores the frequency of the token in a sentence

2.3 Pre-Analysis

In this assignment, a single input document contains multiple sets of queries and answers. However, for the sake of our AAE, each sentence is processed individually. Therefore each sentence is annotated as an individual “Document.”

2.4 Document Reader

The Document Reader code was unmodified during the assignment. Document Reader reads in documents, which are lines of tab-separated values (TSV). Each line is parsed by the reader and Document annotations containing the query ID, the relevance, and the text content are created.

2.5 Document Vector Annotator

The Document Vector Annotator, reads in each document annotations output in the Document Reader, and converts each document into a bag-of-words (BOW) vector. When using BOW, each document contains only a small subset of the entire dictionary, therefore BOW vectors can be considered *sparse vectors*. As such, it is more efficient to store the vectors as maps, as opposed to arrays.

The *createTermFreqVector()* method tokenizes a line of text, and creates a map of tokens and token frequencies. These token-frequency pairs are then stored as Token annotations in an FSList, and added to each Document Annotation as its BOW vector.

2.6 Document Evaluator

The Document Evaluator, reads in all Document Annotations and calculates answer scores based on the cosine similarity between each query-answer pair. Upon completion of the calculations, we compute a MRR based on the rank of the highest scoring correct answer. As explained in our assumptions, although the test data does not show it, we expect any number of correct answers.

3 Error Analysis

3.1 Introduction

This section will outline the steps taken to improve the Vector Space Retrieval Model.

3.1.1 Baseline

The baseline was as follows:

- Sentences tokenized by white-spaces and punctuations
- No stop-words
- Cosine similarity distance

3.1.2 Ideas for MRR Improvement

Some ideas for improvement include:

- Ignore case during tokenization
- Utilize stop-word list
- Try other similarity measures

3.1.3 Other Goals

The following are a set of goals that were considered during implementation:

- For correct sentences that were ranked at the top, maximize the margin between its score and the score of the closest wrong sentence.
- For correct sentences that were not ranked at the top, minimize the margin between its score and the top score.
- Minimize execution times.

3.2 Results

Improvement	Component	MRR	Execution Time (s)	Worst Rank	Comment
Baseline		**0.77**	**0.838**	**3**	**For qid 5, cosine similarity results in a zero when case-sensitive. However, it still receives a rank 2 due to the zero for another answer.**
Remove Case-sensitivity	Document Vector Annotator	0.87	0.791	3	Improved performance. However, qid 3 had a low rank. This is likely due to the length of the sentence, which affects the cosine similarity calculation (increases the norm).
Implement Stop-word List	Document Vector Annotator	0.8	0.797	2	Lowered the MRR score. However, it improved performance under qid 3. Likely to improve performance over larger data sets.
Baseline #2		**0.8**	**0.797**	**2**	**Based on stop-word list implementation**

Dice Coefficient Similarity	Dice Coefficient Similarity	0.8	0.793	2	MRR and ranks similar to cosine similarity.
Euclidean Distance Similarity	Similarity Measure	0.9	0.824	2	MRR higher, and ranks completely different (qid 2 is worst-case).
Jaccard Coefficient Similarity	Similarity Measure	0.9	0.804	2	Same MRR as Euclidean, but qid 4 is worst case.
Tanimoto Similarity	Similarity Measure	0.8	0.794	2	Same MRR as cosine similarity.

In general, the best MRR results came from remove case-sensitivity, implementing the stop-word list, and using the Euclidean Distance to score the answers. In terms of execution times, since the data set is so small, there were no noticeable differences.

Results are based on the MRR and Execution Time scores for the following test data:

```
qid=1   rel=99 Classical music is dying
qid=1   rel=0  Pop music has absorbed influences from most other genres of popular music
qid=1   rel=1  Classical music may never be the most popular music
qid=1   rel=0  Everybody knows classical music when they hear it
qid=2   rel=99 Energy plays an important role in climate change
qid=2   rel=0  Old wine and friends improve with age
qid=2   rel=0  With clothes the new are the best, with friends the old are the best
qid=2   rel=1  Climate change and energy use are two sides of the same coin.
qid=3   rel=99 One's best friend is oneself
qid=3   rel=1  The best mirror is an old friend
qid=3   rel=0  My best friend is the one who brings out the best in me
qid=3   rel=0  The best antiques are old friends
qid=4   rel=99 The shortest distance between new friends is a smile
qid=4   rel=0  Wear a smile and have friends; wear a scowl and have wrinkles
qid=4   rel=1  If you see a friend without a smile, give him one of yours
qid=4   rel=0  Behind every girls smile is a best friend who put it there
qid=5   rel=99 It takes a long time to grow an old friend
qid=5   rel=0  Old wine and friends improve with age
qid=5   rel=0  With clothes the new are the best, with friends the old are the best
qid=5   rel=1  Old friends are best
```

4 Bonus

4.1 Introduction

In addition to the Cosine Similarity measure, a handful of other similarity measures were used on the test data compute answer scores:

- Dice Coefficient
- Jaccaard Coefficient
- Euclidean Distance
- Tanimoto Distance

In the Document Evaluator class, a similarity measure setting was added so that classes implementing an ISimilarityMeasure interface could easily be substituted.

4.2 Results

Measure	Rank qid=1	Rank qid=2	Rank qid=3	Rank qid=4	Rank qid=5	MRR	Avg Correct Answer Score	Avg. Correct Answer Margin
Cosine Similarity	1	1	2	2	1	0.8	0.399	0.082
Dice Coefficient	1	1	2	2	1	0.8	0.081	0.009
Jacaard Coefficent	1	1	1	2	1	0.9	0.238	0.035
Euclidean Distance	2	1	1	1	1	0.9	0.288	0.013
Tanimoto Distance	1	1	2	1	1	0.8	0.271	0.041

We used the alternative distance measures and ran it against the test data. As shown in the Error Analysis Section, the MRR can be improved by using a Jaccard Coefficient or Euclidean Distance-based similarity measure. However, is it safe to assume that these were the better similarity measures based on MRR score alone?

In addition to the ranks of the correct answers, the mean answer scores and margins were calculated for each of the queries. For the margin, **we calculated the difference between the correct answer and the highest scoring incorrect answer**. For queries in which the correct answer was not ranked highest, this resulted in a negative margin. As you can see, although the Euclidean Distance and Jaccard Coefficient similarities resulted in the best MRR, the average correct answer score and margin were significantly lower than that of the Cosine Similarity. This means they did a poorer job of distinguishing correct answers from incorrect answers. This fact should be considered, when judging which similarity measure is most effective.