Name: Niloy Gupta **AndrewID: niloyg**

11-791 Design and Engineering of Intelligent Information System

Homework 3

Engineering and Error Analysis with UIMA

Task1

In task 1, tokenization and cosine similarity was implemented. The documents were sorted based on the similarity measure and the relevant document was ranked. The final Mean Reciprocal Rank (MRR) was calculated.

The intermediate processed data is stored in memory by a singleton class object (DocumentVectorCache.java). The tokens are grouped by documents/queries which are in turn grouped by QIDs. This makes retrieval and ranking faster.

A customized comparator method is written which sorts the documents based on the similarity measure. If there is a tie, the relevant document is ranked higher.

Task 2

Error Analysis

Error Analysis (White-space Tokenizer + Cosine Similarity Measure)

Q: Give us the name of the volcano that destroyed the ancient city of Pompeii.
A:In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volcanic ash the Roman city of Pompeii; an estimated 20,000 people died.
Rank: 2
Analysis:

Query Token	Relevant Document Token Missed	Error Type
Pompeii	Pompeii;	Tokenization Error
volcano	volcanic	No stemming

Q: What has been the largest crowd to ever come see Michael Jordan
A: When Michael Jordan--one of the greatest basketball player of all time--made what was expected to be his last trip to play in Atlanta last March, an NBA record 62,046 fans turned out to see him and the Bulls.
Rank 2:

Query Token	Relevant Document Token Missed	Error Type
Jordan	Jordan--one	Tokenization Error

Q: In which year did a purchase of Alaska happen?
A: Alaska was purchased from Russia in year 1867.
Rank: 3

Query Token	Relevant Document Token Missed	Error Type
purchase	purchased	stemming Error

Q: What year did Wilt Chamberlain score 100 points?
A: On March 2, 1962, Wilt Chamberlain scored a record 100 points in a game against the New York Knicks.
Rank: 2

Query Token	Relevant Document Token Missed	Error Type
points?	points	Tokenization Error
score	scored	No stemming

Q: What river is called China's Sorrow?
A: People of China have mixed feelings about River, which they often call "sorrow of China"
Rank: 3

Query Token	Relevant Document Token Missed	Error Type
China's	(China", China)	Tokenization and stemming Error
river	River,	Vocabulary and stemming error
Sorrow?	"sorrow	Tokenization and stemming Er

Q: Who was the first person to run the mile in less than four minutes
A: Roger Bannister was the first to break the four-minute mile barrier.
Rank: 2

Query Token	Relevant Document Token Missed	Error Type
minutes	4-minute	Tokenization and stemming Error

Q:What year did Alaska become a state?
A: And that's not even to mention the breathtaking beauty of Alaska that became, in 1959, our 49th state.
Rank: 3

Query Token	Relevant Document Token Missed	Error Type
state?	state.	Tokenization Error

Q:When did Mike Tyson bite Holyfield's ear?
A: Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inexplicably bit his opponent's right ear.
Rank: 2

Query Token	Relevant Document Token Missed	Error Type
bite	bit	No stemming
ear?	ear	Tokenization Error

Q: What was the first spaceship on the moon
A: Luna 2 was the first spacecraft to reach the surface of the Moon.
Rank: 2

Query Token	Relevant Document Token Missed	Error Type
moon	Moon	Case Error/Grammar
spaceship	spacecraft	Vocabulary Error (Synonym)

Q: Who won the Nobel Peace Prize in 1992?
A: Menchu won the Nobel peace prize in 1992.
Rank: 1

Query Token	Relevant Document Token Missed	Error Type
1992?	1992	Tokenization Error
Peace	Peace	Grammar/Vocabulary Error
Prize	prize	Grammar/Vocabulary Error

Statistical Analysis of Tokenization algorithms, Stemming strategies and Similarity metrics

The following strategies were implemented and executed on the provided dataset.

Tokenization:
1. White space tokenization: Tokens are extracted between whitespaces in the documents.
2. Stanford NLP Tokenizer: We use the available Stanford NLP tokenizer and remove punctuations from the token set. The tokens are converted to lower case and stemmed.

Stemming:
1. Stanford Lemmatizer
2. Pruning Stop Words: Use the data provided in stopwords.txt to remove stop words from input data.

Similarity Measures:
1. Cosine Similarity
2. Sorensen Dice Coefficient
3. Jaccard Index
4. Okapi BM 25

Combinations of the above methods were combined and reciprocal ranks of the relevant documents were compared against each approach for statistical analysis. The results are summarized below:

Table 1: MRR Results

Method (Tokenization+Stemming+Similarity Measure)	(MRR) Mean Reciprocal Rank
White-space tokenizer + Cosine Similarity	0.4375
White-space tokenizer +Stemmer+ Cosine Similarity	0.5500
White-space tokenizer +Stemmer+ BM25	0.6250
White-space tokenizer +Stemmer+ Jaccard Coefficient	0.8375
White-space tokenizer +Stemmer+ Dice Coefficient	0.5000
Stanford Tokensizer +Stemmer+ Cosine Similarity	0.6625
Stanford Tokensizer +Stemmer+ BM25	0.6042
Stanford Tokensizer +Stemmer+ Jaccard Coefficient	0.8292

Stanford Tokensizer +Stemmer+ Dice Coefficient	0.6292
Stanford Tokensizer +Stemmer+ Stopword-prunning+ BM25	0.6625
Stanford Tokensizer +Stemmer+ Stopword-prunning+ Jaccard coefficient	0.9667
Stanford Tokensizer +Stemmer+ Stopword-prunning+ Dice Coefficient	0.6458

Table 2: T Tests

Initial Method	Experimented Method	P value (Paired T-Test)	Null Hypothesis (Reject/Accept) at alpha = 0.05
White-space tokenizer + Cosine Similarity	White-space tokenizer +Stemmer+ Cosine Similarity	0.0637	Accept
White-space tokenizer + Cosine Similarity	White-space tokenizer +Stemmer+ BM25	0.0583	Accept
White-space tokenizer + Cosine Similarity	White-space tokenizer +Stemmer+ Jaccard Coefficient	0.0583	Accept
White-space tokenizer + Cosine Similarity	White-space tokenizer +Stemmer+ Dice Coefficient	0.2891	Accept
White-space tokenizer + Cosine Similarity	Stanford Tokensizer +Stemmer+ Cosine Similarity	0.0022	Reject
White-space tokenizer + Cosine Similarity	Stanford Tokensizer +Stemmer+ BM25	0.0755	Accept
White-space tokenizer + Cosine Similarity	Stanford Tokensizer +Stemmer+ Jaccard Coefficient	6.0839e-05	Reject
White-space tokenizer + Cosine Similarity	Stanford Tokensizer +Stemmer+ Dice Coefficient	0.0109	Reject
White-space tokenizer + Cosine Similarity	Stanford Tokensizer +Stemmer+ Stopword-prunning + BM25	4.0202e-09	Reject
White-space tokenizer + Cosine Similarity	Stanford Tokensizer +Stemmer+ Stopword-prunning + Jaccard Coefficient	1.3880e-08	Reject
White-space tokenizer + Cosine Similarity	Stanford Tokensizer +Stemmer+ Stopword-prunning + Dice Coefficient	1.1642e-09	Reject

Discussion:

It can be observed that the Jaccard Coefficient proves to be a more efficient similarity measure for ranking the given document set. Moreover combining Jaccard Coefficient with Stanford Tokenizer , stemming and pruning stop words yields statistically improved results.