HW-3 REPORT

Name: Prajwal Yadapadithaya
AndrewID: pyadapad

Vector Space Retrieval Model using UIMA:

The basic implementation was done using the basic tokenizer provided and the cosine similarity function. A cosine similarity function was implemented to compare the similarity values. Based on the cosine similarity values, ranks of the relevant documents were calculated and also the Mean Reciprocal Rank. In this report, I have described the error analysis done on this implementation, and some of the improvements (Including Stanford lemmatizer for tokenization, and similarity functions such as Dice coefficient and Jaccard coefficient)
The results of the run are saved on report.txt as mentioned in the handout.

Error Analysis:

The table lists the various errors for which the relevant tokens were not considered into the calculation for cosine similarity between the query and document strings.

Query-1: Give us the name of the volcano that destroyed the ancient city of Pompeii

Error type	String in query string	String in document string
Tokenization error	Pompeii	Pompeii;
Tokenization error	Pompeii	Pompei;
Synonym match error	Destroyed	Buried

Query-2: What has been the largest crowd to ever come see Michael Jordan

Error Type	String in query string	String in document string
Synonyms error	largest	greatest
Vocabulary mismatch	Crowd	Fans

Query-3: In which year did a purchase of Alaska happen?

Error Type	String in query string	String in document string
Grammar error	purchase	purchased
Vocabulary mismatch	happen	negotiated

Query-4: What year did Wilt Chamberlain score 100 points?

Error Type	String in query string	String in document string
Grammar error	score	scored
Tokenization error	points?	points

Query-5: What river is called China's Sorrow?

Error Type	String in query string	String in document string
Tokenization error	China’s	China
Grammar error	called	call
Tokenization error	Sorrow	“sorrow

Query-6: Who was the first person to run the mile in less than four minutes

Error Type	String in query string	String in document string
Tokenization error	four minutes	four-minute
Tokenization error	four minutes	4-minute

Query-7: What year did Alaska become a state?

Error Type	String in query string	String in document string
Grammar error	become	became
Spelling error	Alaska	Aaska
Tokenization error	state?	state

Query-8: When did Mike Tyson bite Holyfield's ear?

Error Type	String in query string	String in document string
Grammar error	bite	bit
Stemming error	bite	biting
Tokenization error	ear?	ear.

Query-9: What was the first spaceship on the moon

Error Type	String in query string	String in document string
Synonyms error	spaceship	spacecraft
Tokenization error	moon	moon,

Query-10: Who won the Nobel Peace Prize in 1992?

Error Type	String in query string	String in document string
Tokenization error	1992?	1992.
Vocabulary mismatch	won	honored

Query-11: Where is Devil's Tower

Error Type	String in query string	String in document string
Tokenization error	Devil’s	Devils

Query-12: What is the height of the tallest redwood?

Error Type	String in query string	String in document string
Tokenization error	redwood?	redwood

Query-13: How deep is Crater Lake?

Error Type	String in query sentence	String in document sentence
Tokenization error	Lake?	Lake
Stemming error	deep	depth

Query-14: Who was the lead singer for the Commodores

Error Type	String in query sentence	String in document sentence
Tokenization error	Commodores	Commodores.

Query-15: What is the coldest place on earth?

Error Type	String in query sentence	String in document sentence
Tokenization error	earth?	Earth

Query-16: When did Bob Marley die

Error Type	String in query sentence	String in document sentence
Grammar error	die	died

Query-17: Which U.S. state is the leading corn producer?

Error Type	String in query sentence	String in document sentence
Tokenization error	producer?	producer
Vocabulary mismatch	U.S.	United States
Synonyms detection error	state	area
Synonyms detection error	leading	largest

Query-18: Where was the first McDonald's built?

Error Type	String in query sentence	String in document sentence
Tokenization error	McDonald’s	McDonald
Vocabulary mismatch	first	single

Query-19: The Hindenburg disaster took place in 1937 in which New Jersey town?

Error Type	String in query sentence	String in document sentence
Vocabulary mismatch	disaster	burned and crashed
Tokenization error	New Jersey	N.J.,

Query-20: What is the Keystone State?

Error Type	String in query sentence	String in document sentence
Tokenization error	State?	State,
Spelling error	Keystone	Keytstone

Summary of error analysis:

Error Type	Number of Queries
Tokenization Error	17
Grammar Error	6
Vocabulary Mismatch	6
Spelling Error	2
Stemming Error	2

Implementation of better tokenizer algorithms and similarity functions:

In this section, I've documented the results of implementations of various other similarity functions using the given dumb tokenizer and stanford lemmatizer. I have also computed the p-values using the t-test method for the two cases where the maximum MRR was found, i.e using Stanford Lemmatizer for tokenization and Dice coefficient and Jaccard distance for similarity functions.
The results of the implementations are as follows:

Using (Dumb Tokenizer + Cosine Similarity):
(MRR) Mean Reciprocal Rank :: 0.4375

Using (Stanford Lemmatizer + Cosine Similarity):
(MRR) Mean Reciprocal Rank :: 0.55

Using (Dumb Tokenizer + Sørensen–Dice_coefficient):
(MRR) Mean Reciprocal Rank :: 0.4625

Using (Stanford Lemmatizer + Sørensen–Dice_coefficient) **[p-value: 0.1159]**
(MRR) Mean Reciprocal Rank :: 0.5667

Using (Dumb Tokenizer + Jaccard Distance)
(MRR) Mean Reciprocal Rank :: 0.4625

Using (Stanford Lemmatizer + Jaccard Distance) **[p-value: 0.2151]**
(MRR) Mean Reciprocal Rank :: 0.5667