A Project Report on

Assignment II

Building an Query Retrieval IR System

Submitted By

Akash Kabra 2016B3A70562P
Kumar Anant Raj 2016B4A70520P

Submitted to

Dr. Abhishek

For partial fulfillment of the course

Information Retrieval

On March 7, 2020

PART 1

In part 1 , we have implemented lnc.ltc tf-idf vectorization to retrieve relevant documents from the corpus for a query. We have cleaned the corpus by splitting it into different documents by </doc> html tag. The underlying algorithm is similar to the one discussed in class.

This algorithm has a few limitations:
Completely misspelled query will not retrieve any relevant documents.

It is assumed that a user writes English queries in 8-bit ASCII format.

Also, words in different morphological form from the one in the relevant corpus, may not lead to a correct retrieval.

	Document Title	Score	Is the Document Relevant
Emperor were friends before Gaozong became an Emperor	MUYEOL OF SILLA	14.498	Yes
	BẢO VÀNG	13.1867	No
	BRITISH EMPEROR	8.9199	No
	NERO (2004 FILM)	8.7961	No
	THIERRY AMAR	6.5730	No
	NGUYỄN LORDS	6.5434	No
	EMPEROR DAOWU OF NORTHERN WEI	6.5108	No
	MOSCOW PASSAZHIRSKAYA RAILWAY STATION	6.1680	No
	JOHN OF MONTECORVINO	6.1680	No
	PATRIMONIUM SANCTI PETRI	5.9262	No

	Document Title	Score	Is the Document Relevant
NCAA BASKETBALL TOURNAMENT	1940 NCAA MEN'S DIVISION I BASKETBALL TOURNAMENT	32.4228	Yes
	1942 NCAA MEN'S DIVISION I BASKETBALL TOURNAMENT	31.6669	Yes
	1939 NCAA MEN'S DIVISION I BASKETBALL TOURNAMENT	31.5942	Yes
	1943 NCAA MEN'S DIVISION I BASKETBALL TOURNAMENT	31.5731	Yes
	1953 NCAA MEN'S DIVISION I BASKETBALL TOURNAMENT	31.4002	Yes
	1941 NCAA MEN'S DIVISION I BASKETBALL TOURNAMENT	31.3868	Yes
	1944 NCAA MEN'S DIVISION I BASKETBALL TOURNAMENT	31.0649	Yes
	1984 NCAA MEN'S DIVISION I BASKETBALL TOURNAMENT	30.9687	Yes
	1948 NCAA MEN'S DIVISION I	30.4732	Yes

	BASKETBALL TOURNAMENT		
	1970 NCAA MEN'S DIVISION I BASKETBALL TOURNAMENT	30.4155	Yes

	Document Title	Score	Is the Document Relevant
Behennah aspired to study medicine and become a plastic surgeon	MICHELLE BEHENNAH	15.3956	Yes
	FISH MEDICINE	9.5078	No
	RICHARD CRUESS	6.9152	No
	HERMANN LEBERT	4.5576	No
	HEAT DEFLECTION TEMPERATURE	4.2079	No
	DUNEDIN MULTIDISCIPLINARY HEALTH AND DEVELOPMENT STUDY	4.2079	No
	SPUDGER	3.9634	No
	MARIE COLINET	3.7358	No
	ACCRA SPORTS STADIUM DISASTER	3.5212	No
	HARRY ROWSELL	3.4269	No

MTB Endia (Misspelled from MTV India)	**Document Title**	**Score**	**Is the Document Relevant**
	No Documents Retrieved.		

Instrument United Kingtom (Misspelled Kingdom)	**Document Title**	**Score**	**Is the Document Relevant**
	DOTARA	12.72688	No
	BORE (WIND INSTRUMENTS)	11.8462	No
	LIST OF STATUTORY INSTRUMENTS OF THE UNITED KINGDOM, 2003	10.5516	Yes
	LIST OF STATUTORY INSTRUMENTS OF THE UNITED KINGDOM, 2004	10.5516	Yes
	LIST OF STATUTORY INSTRUMENTS OF THE UNITED KINGDOM, 1998	10.5516	Yes
	LIST OF STATUTORY INSTRUMENTS OF THE UNITED KINGDOM, 2001	9.4875	Yes
	SALLANEH (LUTE)	9.0035	No
	SCALE LENGTH (STRING INSTRUMENTS)	8.1985	No

	FODAY MUSA SUSO	7.9069	No
	INDIAN PINK	7.6936	No

	Document Title	Score	Is the Document Relevant
Kim Chunhcu	MUYEOL OF SILLA	17.6141	Yes
	NIKKI CLEARY	8.4273	No
	KIM LEWIS	7.3097	No
	KIM VAN KOOTEN	4.4183	No
	KIM CHRISTOFTE	3.8145	No
	WHPC	3.5795	No
	CHARLOTTE SOMETIMES (FILM)	2.9602	No
	PUKYONG NATIONAL UNIVERSITY	2.9584	No
	NIKKI CLEARY	2.7353	No
	STEVEN MCEWAN	2.5361	No

	Document Title	Score	Is the Document Relevant
Seven Fleets large invasion	LIST OF SHIPS PRESENT AT INTERNATIONAL FLEET REVIEW, 2005	8.8637	No

	USS BURKE (DE-215)	7.8441	No
	PHILIPPINES CAMPAIGN (1944–45)	7.7102	Yes
	USS PC-1137	7.6433	No
	USS ENRIGHT (DE-216)	6.7308	No
	YI EOKGI	6.5601	No
	BATTENBERG CUP	6.5011	No
	USS PC-1138	5.6463	No
	USS COOLBAUGH (DE-217)	5.5818	No
	USS NORRIS (DD-859)	5.3471	No
	PATRIMONIUM SANCTI PETRI	5.9262	No

	Document Title	**Score**	**Is the Document Relevant**
	AB MOTORFABRIKEN I GÖTEBORG	9.0126	No
	LIST OF LUNAR CRATERS NAMED FOR SPACE EXPLORERS	8.8566	No
	SUFFREN-CLASS CRUISER	7.6451	No
	AB NYKÖPINGS	7.3729	No

Four Hundreds named automobile (Hundreds here is a misspelled proper noun)	AUTOMOBILFABRIK		
	PACKARD FOUR HUNDRED	5.9802	Yes
	1998 NCAA MEN'S DIVISION I BASKETBALL TOURNAMENT	5.6854	No
	1956 NCAA MEN'S DIVISION I BASKETBALL TOURNAMENT	5.6386	No
	1944 NCAA MEN'S DIVISION I BASKETBALL TOURNAMENT	5.4779	No
	1952 NCAA MEN'S DIVISION I BASKETBALL TOURNAMENT	5.2660	No
	GROSSE ILE TOLL BRIDGE	5.2353	No

microbiolog	**Document Title**	**Score**	**Is the document relevant**
	No documents retrieved.		

	Document Title	Score	Is the Document Relevant
	NAM-GU	20.3111	No
	LIST OF STATUTORY	17.0585	No

Southern Korean Kingdom	INSTRUMENTS OF THE UNITED KINGDOM, 2003		
	LIST OF STATUTORY INSTRUMENTS OF THE UNITED KINGDOM, 2004	17.0585	No
	LIST OF STATUTORY INSTRUMENTS OF THE UNITED KINGDOM, 2004	17.0585	No
	LIST OF STATUTORY INSTRUMENTS OF THE UNITED KINGDOM, 1998	15.3382	No
	LIST OF STATUTORY INSTRUMENTS OF THE UNITED KINGDOM, 2001	11.7754	No
	DANDA KINGDOM	11.3277	No
	LIST OF STATUTORY INSTRUMENTS OF THE UNITED KINGDOM, 2000	11.2577	No

PART II

Q1. What is the issue with the IR system built in part 1?

The tf-idf *lnt.ltc* system built in part 1 faces issues for the following cases:

- When the spelling of all the tokens in the query is wrong, no documents are retrieved
- When the word is not presented in it's base form. Like "*Hundreds*" instead of "*Hundred*", the relevant documents might not be retrieved.
- When the words are presented in a different morphological order, like "*working*" or "*worked*" for "*work*".

Example:

"MTB Endia" retrieves no relevant document by part 1 tf-idf vectorization, as none of the word appears in the corpus. Whereas, if we are able to Identify that this spelling is misspelled as "Endia" instead of "India" and "MTB" instead of "MTV", then relevant documents can be retrieved.

Q2. What improvement are you proposing?

We propose to do the following 2 improvements:

1. **Stem** the document as well as the query to increase recall of the IR system. **Lemmatize** the document to improve the precision of the IR system. Obviously, using both of them together will not lead to a really significant change in the retrieval as compared to using just one of them. So, we propose to use both of them individually and not together.

2. **Spell Check**: Spelling is checked by comparing the input word with all unique tokens in the corpus by the distance metric D, defined as:

$$D = \lambda * (Levenshtein\ Distance) + (1 - \lambda) * (Soundex\ Distance) \quad ...(1)$$

where Levenshtein Distance is the edit distance of inserting, deleting or replacing the input word to get the output word, as discussed in class. Soundex distance is an algorithm that matches phonetic representation of the input word with the output word, rather than the actual syntactic form. It is a heuristic to get the phonetic representation(soundex form) of an input word, then it's levenshtein distance is calculated the soundex form of the entire corpus.

Algorithm to get Soundex Representation:

Change all occurrence of the following letters to '0' (zero):
'A', E', 'I', 'O', 'U', 'H', 'W', 'Y'.
Change letters from the following sets into the digit given:
1 = 'B', 'F', 'P', 'V'
2 = 'C', 'G', 'J', 'K', 'Q', 'S', 'X', 'Z'
3 = 'D','T'
4 = 'L'
5 = 'M','N'
6 = 'R'
Remove all pairs of digits which occur beside each other from the string that resulted after step (4).
Remove all zeros from the string that results from step 5.0 (placed there in step 3)
Pad the string that resulted from step (6) with trailing zeros and return only the first four positions, which will be of the form <uppercase letter> <digit> <digit> <digit>.

We deploy the following steps to get top K documents:

1. Preprocess Query(remove punctuations, convert to lower, stem/lemmatize if needed)
2. For each token in query:
   Calculate D^* as defined in (1) with the entire corpus.
   Retrieve top K^* tokens closest to word in the query.
   Amongst these K tokens, keep those with D < $Threshold^*$
   Add all these tokens to the new query.
3. Search in the corpus by new spell checked query to retrieve top 10 documents.

* K, λ, Threshold are hyperparameters adjusted as K = 1, λ = 0.8, Threshold= 0.35 in the implementation.

Overall Model

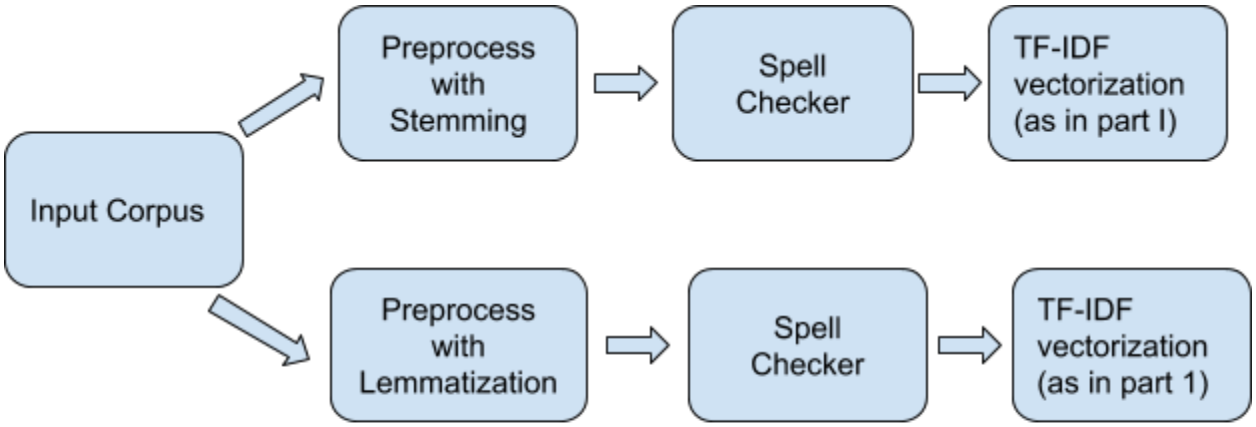

Fig: Two approaches Followed

3. How will the proposed improvement address that issue?

- We are deploying spell checker by a weighted distance of Levenshtein Distance and Soundex Distance. This would enable us to identify that user has committed a mistake in typing the query, and thus suggest the closest word.
- For query written in incorrect morphological form, like writing “thousands” instead of “thousand”, would retrieve wrong documents, if “thousand” is used as a relevant noun form in the query, wrong documents would have been retrieved if there was no stemming. This would lead to an increase in recall of the system.
- Lemmatization may lead to an increase in precision of the system.

4. A corner case (if any) where this improvement might not work or can have an adverse effect.

The IR system in the implementation designed has used K = 1, i.e. it would give the closest term in the corpus to the term in the query. In other words, if the word in query matches with any word in the corpus, then spell check will not modify the word, else, it will find the closest word in the corpus.

So, if the closest word found is actually an irrelevant word, which is possible as finding the closest term might give an irrelevant word. This will increase False Positives in the result.

Also, if the spelling of the word was misspelled, but is a well-spelled word making a completely different meaning, like writing “Indiana” in place of “Indian”, then the spell checker will not correct the spelling and would retrieve for the wrong word “Indiana”

5. Demonstrate the actual impact of the improvement. Give three queries, where the improvement yields better results compared to the part 1 implementation.

Query	Part I	Stemming + Spell Check	Lemmatization + Spell Check	Comments
MTB Endia	No Documents retrieved	MTV INDIA	MTV INDIA	Both our implementations are able to retrieve relevant documents.
Microbiolog	No Document retrieved	PUKYONG NATIONAL UNIVERSITY	PUKYONG NATIONAL UNIVERSITY	Both our implementations are able to retrieve relevant documents.
Shigeki Segusa (misspelled from)	No Document retrieved	SHIGEAKI SAEGUSA	SHIGEAKI SAEGUSA	Both our implementations are able to retrieve relevant documents.