A report on

Vector Space Based Information Retrieval System

Submitted by

Kalit Inani	2018A7PS0207P
Nisarg Vora	2018A7PS0254P
Harsh Sharma	2018A7PS0230P
Rahil Jain	2017B4A70541P
Prateek Grover	2018A3PS0338P

In partial fulfillment of the course
Information Retrieval
(CS F469)

Date: 4th May, 2021

Acknowledgements

The successful completion of this project required a lot of support and guidance. We were extremely privileged to have obtained this all through the process. We would like to express our appreciation to all those who helped us in making this project reach its completion.

A special thanks to Prof. Sudeept Mohan, Head of Department, Computer Science and Information Systems, BITS Pilani for providing the opportunity to work on this topic. We would like to acknowledge Dr. Vinti Agarwal, Assistant Professor, Computer Science and Information Systems, BITS Pilani for formulating the problem statement, and mentoring us throughout the process.

We would also like to thank the course teaching assistants, Aditya Deshmukh and Raksha for coordinating issues faced by us with the professors.

At last, we express our gratitude to our family and friends for a constant source of motivation that they consistently provided over the course of this project.

Abstract

This is a report to the assignment “Vector-space based information retrieval system”. It covers an overview of implementation, results, improvements and constraints on making an information retrieval system.

This assignment gave us a better understanding of how actual IR systems work, how the scoring system works, and made us understand the limitations of vector-space based systems.

In this assignment we learnt how to implement a few really good research articles, explore very useful packages of python, and gave us industry level experience of building systems.

We came across a method called query relaxation wherein the scores of synonyms of the tokens of the query are added to the scores of each document calculated by the vanilla model. If insufficient words are retrieved by this method, then co-hyponyms of the tokens are also used. This is particularly helpful when the tokens of the query are rare words, or when the user is not confident about the query they intend to use.

We came across a very interesting topic Latent Semantic Indexing where with constant use of memory we can add a lot of documents into the model. We found that this is advantageous when the size of the corpus is very large and if the data is streamed continuously. We found that this method is really fast and efficiently understands the semantics of the query.

The naive implementation of the vector space model was evaluated on 10 multi term queries. Moreover, 3 multi-term queries were chosen to show the improvements in the model.

The project is restricted to vector space ranking techniques, however more advanced techniques like Page Rank could be explored in the future.

Table of Contents

Introduction

Information Retrieval refers to the process of dealing with storage, retrieval and evaluation of information from a large number of documents. The information exchange can be in the form of text, images, sounds or a combination of all of them. Information retrieval is particularly useful in obtaining required relevant material from an otherwise unstructured source of data. This procedure using textual information is often implemented by software systems based on IR Models that have access to books, journals and articles. Web search engines are one of the best examples of IR Software Systems.

The IR Model would rank the documents based on a query input by the user, this is possible as the documents and queries would be represented in similar manner, and the ranking would be made possible based on a matching function which would help retrieve a relevance score for each document in the collection.

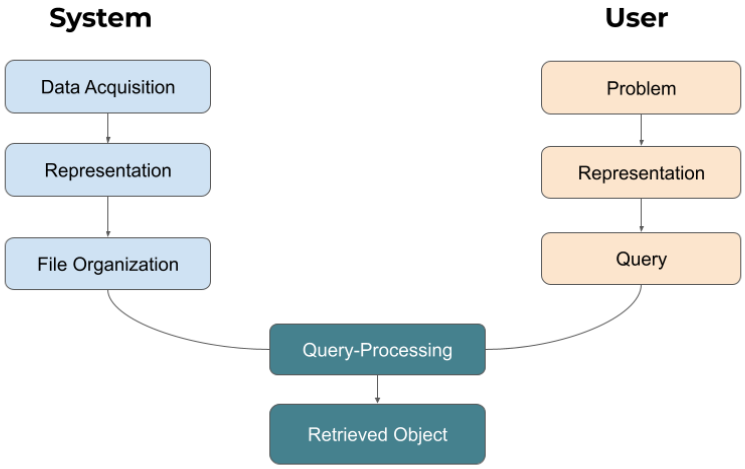

***Figure 1**: Flowchart depicting working of an IR Model*

An IR Model has several different steps, the first of which is **Data Acquisition**: this would involve collection of documents from web sources and their storage in a database. The second is **Representation** : this consists of classification and indexing of documents based on their content and bibliographic information. Next step is choosing a mode of **File Organization**, this can either be sequential : containing documents by their data, Inverted: containing term by term records of documents under each term, or a combination of them. Finally **Query-Processing**: queries, which are statements of information needs are passed into the IR System, after which several objects matching the query with varying degrees of relevance are retrieved.

Our research in this project shall be limited to creating a vector-space information retrieval model that allows users to search for relevant Wikipedia pages using queries input by them into a large database of Wikipedia Pages.

Methodology

In a basic vector-space model, documents and the queries are converted into n-dimensional vectors (n being the number of distinct terms over all documents and queries) where the value of the i^{th} term is based on the weights assigned to it in the document or query, taking into account its frequency.

At retrieval, cosine of the angle between each of the document vector and query vector is calculated, and the documents are ranked on the order of decreasing matching scores.

The documents are then ranked on the basis of a matching function between the document vector and the query vector, this matching function is a cosine function here.

A common way to represent term frequency in large documents is through logarithmic term frequency score which is $tf_{score} = 1 + log\,(tf_{t,d})$, where $tf_{t,d}$ is the occurrences of term t in document d.

A common way to represent document frequency in large documents is through inverse document frequency score, $idf_{score} = log(N/df_t)$, where $N$is the number of documents and $df_{t,}$is the number of documents that contain term t.

The weights of document and query vector are often normalized by multiplying them with cosine normalization factor $k = 1/(w_1^2 + w_2^2 + \ldots + w_n^2)^{1/2}$**,** where $w_i$is the weight of i^{th}term in the n-dimensional vector.

We calculate the weight values of document and query vectors using a common $lnc.\ ltc$ scoring scheme based on SMART Notation, where the document vector values are calculated using lnc (logarithmic term frequency calculation and cosine normalization) scheme, while query vector values are calculated using ltc(logarithmic term frequency calculations, inverse document frequency calculation and cosine normalization).

After the final values of document and query vector are obtained,

$Cosine\ Similarity_{q,d} = \sum_{i=1}^{n} q_i d_i \ / \ ((\sum_{i=1}^{n} q_i^2)^{1/2}(\sum_{i=1}^{n} d_i^2)^{1/2})$, here q and d are query vector and document vector respectively, and are calculated between the query and each of the documents. After which, the top 10 results are obtained based on the highest cosine similarity values.

PART 1

Implementation

Dataset and Preprocessing

The dataset taken for the project was a large text corpus of Wikipedia containing 6362 pages with more than a million tokens, and about thousands of unique tokens, of Wikipedia's pages in HTML format. The pages were first preprocessed using HTML parser from BeautifulSoup Library, using which textual data was extracted from them, removing the tags, the data was cleaned and stored in data structure all_docs, tokens were generated using nltk.word_tokenizer() function, and their document vectors are stored.

Data Structures Used

A python dictionary $all\_documents$ was created, containing the textual information for each document. A dictionary $vocabulary$ was generated containing the token IDs as keys and the token objects as values. Another dictionary $reverse\_vocabulary$ was generated, which contained token values as the keys and token objects as the values. An $inverted\_index$ was generated as a python dictionary to store the terms as the keys and the list of documents that they occur in as the values.

Query Processing

Queries taken as input from the user are preprocessed and tokenized similar to the way the documents were preprocessed. Query vectors were generated using $get\_query\_vector()$**,** after which using the $lnc.ltc$ scheme described above scores are generated using $calc\_lnc\_ltc()$ function. Finally the top 10 results were retrieved using the $retrieve\_ranked\_docs()$ function.

Evaluation Metrics

The IR Model was evaluated by searching for a number of multi-term queries, which resulted in top 10 documents having the greatest similarity score out of the corpus. To check the accuracy of the results obtained, ideally we would need to calculate the relevance score of the retrieved document, however, here we manually check the relevance of documents, awarding them a score of 1 if relevant, and a score of 0, if irrelevant.

Binary Relevance Accuracy was measured by:

$$Accuracy = No.\ of\ Relevant\ Documents\ Retrieved\ /\ No.\ of\ Documents\ Retrieved$$

Discounted Cumulative Gain(DCG) was employed to measure performance of IR Model taking into account the rank of retrieved documents as well as their relative relevance. It is measured as:

$DCG_{query} = r_{score,1} + \sum_{i=2}^{n} (r_{score,i})/log_2 i$, where $r_{score,i}$ is relevance score of i^{th} retrieved result.

Results

Following are the retrieved results based on multi-term queries on a basic vector space model, and their relevance score along with binary relevance as described in the evaluation metrics.

1. Query : "Presidential Campaign of Barack Obama"

Query	Top 10 Documents	Score	Binary Relevance
Presidential Campaign of Barack Obama	Nebraska Republican Presidential Primary Results	0.147	Yes
	Eric L. Beach	0.112	No
	Quad Cities (Amtrak train)	0.087	No
	2016 Presidential Primaries in Manatee County, FL	0.085	Yes
	Oleg Savelyev	0.084	No
	United States presidential election, 2020 timeline	0.078	Yes
	United States presidential visits to East Asia	0.067	No
	Jenna Weiss-Berman	0.065	No
	United States presidential visits to Northern Europe	0.064	No
	Haile Thomas	0.061	No

$$Accuracy_{query\ 1} = 3/10 = 0.3$$
$$DCG_{query\ 1} = 1 + 1/log_2 5 + 1/log_2 7 = 1.786$$

2. Query : ‘Greenhouse Gas Emissions’

Query	Top 10 Documents	Score	Binary Relevance
Greenhouse Gas Emissions	Richmond Gas Company Building	0.075	Yes
	Offshore Oil Engineering	0.069	Yes
	Tucuruí transmission line	0.059	No
	Faysal Ahmad Ali al-Zahrani	0.054	No
	LEDA 89996	0.052	No
	John W. Stephenson	0.046	No
	Knoxville Veterans Administration Hospital Historic District	0.040	No
	Jack K. Horton	0.037	No
	Waimiri Atroari Indigenous Territory	0.037	No
	Sophie Warny	0.036	No

$$Accuracy_{query\ 2} = 2/10 = 0.2$$
$$DCG_{query\ 2} = 1 + 1/log_2 3 = 1.6309$$

3. Query : ‘Software Development Industry’

Query	Top 10 Documents	Score	Binary Relevance
Software Development Industry	Andscacs	0.190	No
	Intermediate Data Format	0.158	Yes
	Software Heritage	0.127	Yes
	Noah Glass (Olo Online Ordering)	0.124	No
	Zona (streaming video software)	0.110	No
	EWorkexperience	0.107	Yes

	Janos Sztipanovits	0.107	No
	Stuart Arnold	0.103	No
	Ras Al Khaimah Tourism Development Authority	0.103	No
	Development Financial Institutions Act 2002	0.103	No

$$Accuracy_{query\ 3} = 3/10 = 0.3$$
$$DCG_{query\ 3} = 1/log_2 3 + 1/log_2 4 + 1/log_2 7 = 1.487$$

4. Query : 'Highest Grossing Films'

Query	Top 10 Documents	Score	Binary Relevance
Highest-grossing Films	List of highest grossing Nepali films	0.286	Yes
	List of most expensive Bangladeshi films	0.178	Yes
	Fairwater Films	0.168	No
	List of Albanian documentary film	0.153	No
	List of Albanian animated films	0.153	No
	List of Albanian films of the 2010s	0.138	No
	List of Albanian films of the 2000s	0.138	No
	List of Albanian films of the 1990s	0.138	No
	List of Albanian films of the 1980s	0.138	No
	List of Albanian films of the 1970s	0.138	No

$$Accuracy_{query\ 4} = 2/10 = 0.2$$
$$DCG_{query\ 4} = 1 + 1/log_2 3 = 1.6309$$

5. Query : 'Places in India'

Query	Top 10 Documents	Score	Binary Relevance

Places in India	Places of worship in Mavelikkara	0.320	Yes
	Places of worship in Ernakulam	0.185	Yes
	Aircraft Act	0.156	No
	List of moths of India (Arctiidae)	0.1459	No
	Emmanuel Shearith Israel Chapel	0.145	No
	Carlos Ospina	0.144	No
	Institutes of Technology Act	0.141	No
	KEPT-LP	0.139	No
	2016 North American heat wave	0.138	No
	Kathura	0.137	Yes

$$Accuracy_{query\,5} = 3/10 = 0.3$$
$$DCG_{query\,5} = 1 + 1/log_2 3 + 1/log_2 11 = 1.919$$

6. Query : ‘teaching institutions of the world’

Query	Top 10 Documents	Score	Binary Relevance
teaching institutions of the world	Development Financial Institutions Act 2002	0.229	No
	Instituto de Estudos Medievais	0.184	Yes
	Fareed Town Islamabad	0.148	No
	List of educational institutions closed in the 2016 Turkish purges	0.130	Yes
	Mary Cannell	0.129	No
	2016 Steglitz shooting	0.127	No
	Governance without government	0.124	No
	Female empowerment in Nigeria	0.122	No
	Watson Gordon Chair of Fine Art	0.120	No
	Sharmin and Bijan Mossavar-Rahmani Center	0.120	Yes

	for Iran and Persian Gulf Studies		

$$Accuracy_{query\ 6} = 3/10 = 0.3$$
$$DCG_{query\ 6} = 1/log_2 3 + 1/log_2 5 + 1/log_2 11 = 1.353$$

7. Query : 'Nuclear Energy Policy of United States'

Query	Top 10 Documents	Score	Binary Relevance
Nuclear Energy Policy of United States	Look East policy	0.229	No
	Runit Island	0.184	Yes
	List of Mayors of Gulfport, Mississippi	0.147	No
	Snowball, Minnesota	0.130	No
	List of United States tornadoes from June to August 2016	0.128	No
	2016 Speedway World Cup Event 2	0.126	No
	Five Country Conference	0.123	Yes
	Harold Quinton	0.122	Yes
	Benjamin Sommers	0.120	No
	List of ambassadors of Afghanistan to the United States	0.120	No

$$Accuracy_{query\ 7} = 3/10 = 0.3$$
$$DCG_{query\ 7} = 1/log_2 3 + 1/log_2 8 + 1/log_2 9 = 1.28$$

8. Query : 'Salubrious Nourishment'

Query	Top 10 Documents	Score	Binary Relevance
Salubrious Nourishment	The Nashville Sessions (Townes Van Zandt album)	0.000	No
	Eve Marie Shahoian	0.000	No

	Gospodor Monument Park	0.000	No
	2016 FIVB Volleyball Men's Club World Championship	0.000	No
	Fox hunting (film)	0.000	No
	Latin American and Caribbean Economic Association (LACEA)	0.000	No
	Charles Douglas Carpendale	0.000	No
	Drumheller, Washington	0.000	No
	Earl A. Fitzpatrick	0.000	No
	Douglas Mountain (Washington)	0.000	No

$$Accuracy_{query\,8} = 0/10 = 0$$
$$DCG_{query\,8} = 0$$

9. Query : ‘The Indian cricket team’

Query	Top 10 Documents	Score	Binary Relevance
The Indian cricket team	Indian cricket team in the West Indies in 2017	0.358	Yes
	Indian cricket team in Sri Lanka in 2017	0.347	Yes
	Zimbabwean cricket team in Sri Lanka in 2017	0.274	No
	Pakistani cricket team in Bangladesh in 2017	0.264	No
	Afghan cricket team in the West Indies in 2017	0.263	No
	List of Bangladesh Under-23 international cricketers	0.246	No
	Indian Navy (football club)	0.244	No
	Bangladesh national under-23 cricket team	0.236	No
	List of Bangladesh A international cricketers	0.215	No

	Irish cricket team against Afghanistan in India in 2016–17	0.207	No

$$Accuracy_{query\ 9} = 2/10 = 0.2$$

$$DCG_{query\ 9} = 1 + 1/log_2 3 = 1.63$$

10. Query : ‘Action flick’

Query	Top 10 Documents	Score	Binary Relevance
Action flick	Democratic Action Party leadership election, 2012	0.290	No
	Sebastian Rödl	0.166	No
	Radical Action to Unseat the Hold of Monkey Mind	0.147	No
	High & Low: The Movie	0.143	Yes
	S. P. Bhargavi	0.129	Yes
	Rezin Beall	0.124	No
	Lukamari (2016 film)	0.116	Yes
	Bill Healy	0.107	No
	V. Suresh Thampanoor	0.107	No
	Bleeding Steel	0.102	Yes

$$Accuracy_{query\ 10} = 4/10 = 0.4$$

$$DCG_{query\ 10} = 1/log_2 5 + 1/log_2 6 + 1/log_2 8 + 1/log_2 11 = 1.44$$

Cumulative Results of the Basic Vector - Space Model

Average accuracy obtained on the set of 10 Multi-term Queries above is:

$$Accuracy_{q1-10} = 1/10 \cdot \sum_{n=1}^{10} Accuracy_{query\ i} = 0.25$$

Average DCG Score obtained on the set of 4 Multi-term Queries above is:

$$DCG_{q1-10} = 1/10 \cdot \sum_{n=1}^{10} DCG_{query\ i} = 1.4156$$

Part 2

Improvement 1:

1. **What are issues with the vector space model in part 1?**

   User query may not always be a very accurate representation of information need. Vector Space Model looks only for the terms and mentioned by the user in the query. Often, the documents relevant to the user may not exactly contain the query terms but contain its synonyms or hypernyms. For some queries the vector space model retrieves very less or no documents when the query term words are very rare in the document corpus.

2. **What improvement are you proposing?**

   We propose Query Relaxation using Synonyms and Co-hyponyms which would enable the system to identify documents which are relevant to documents but may not exactly contain the terms in the query.

3. **How does Query Relaxation using synonyms and hypernyms improve upon the previous issues?**

   Synonyms are words that are similar to another word or have a related meaning, **hypernyms** are words with a broad meaning constituting a category into which words with more specific meanings fall; a superordinate and **co-hyponyms** are words which have the same hypernym.

   In query relaxation we form multiple queries by replacing each term in the query with its synonyms and hypernyms. The query score is then the sum of scores of all the queries constructed by replacing the words with their synonyms. In case the number of documents retrieved are very less then we replace the words with their hypernyms and retrieve the corresponding documents with non zero scores as well.

   Query relaxation enables us to identify the relevant documents even when the documents do not contain the exact query words but represent relevant information using synonyms and co-hyponyms. Using co-hyponyms is also useful for the cases when query search contains very few results.

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

For some cases the synonyms may have different meaning for different contexts which may lead to retrieval of irrelevant documents. Co-hyponyms are only a broad representation of the query term and can lead to retrieval of irrelevant documents.

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 1	Relevance	Query Relaxation	Relevance
teaching institutions of the world	Development Financial Institutions Act 2002	No	Development Financial Institutions Act 2002	No
	Instituto de Estudos Medievais	Yes	Instituto de Estudos Medievais	Yes
	Fareed Town Islamabad	No	Mary Cannell	No
	List of educational institutions closed in the 2016 Turkish purges	Yes	Rajapruk University	Yes
	Mary Cannell	No	Mount Sinai Hospital (Minneapolis)	No
	2016 Steglitz shooting	No	Fareed Town Islamabad	No
	Governance without government	No	Female empowerment in Nigeria	No
	Female empowerment in Nigeria	No	Watson Gordon Chair of Fine Art	No
	Watson Gordon Chair of Fine Art	No	List of educational institutions closed in the 2016 Turkish purges	Yes
	Sharmin and Bijan Mossavar-Rahmani Center for Iran and Persian Gulf Studies	Yes	2016 Steglitz shooting	No

Part 1:

$Accuracy_{query\ 2} = 3/10 = 0.3$

$DCG_{query\ 2} = 1/log_2 3 + 1/log_2 5 + 1/log_2 11 = 1.35064551621$

Model with Query Relaxation:

$Accuracy_{query\ 2} = 3/10 = 0.3$

$DCG_{query\ 2} = 1/log_2 3 + 1/log_2 5 + 1/log_2 10 = 1.3626015276$

Query	Part 1	Relevance	Query Relaxation	Relevance
Action flick	Democratic Action Party leadership election, 2012	No	Democratic Action Party leadership election, 2012	No
	Sebastian Rödl	No	High & Low: The Movie	Yes
	Radical Action to Unseat the Hold of Monkey Mind	No	Sebastian Rödl	No
	High & Low: The Movie	Yes	S. P. Bhargavi	Yes
	S. P. Bhargavi	Yes	Radical Action to Unseat the Hold of Monkey Mind	No
	Rezin Beall	No	Lukamari (2016 film)	Yes
	Lukamari (2016 film)	Yes	V. Suresh Thampanoor	No
	Bill Healy	No	Rezin Beall	No
	V. Suresh Thampanoor	No	Bleeding Steel	Yes
	Bleeding Steel	Yes	Jaguar (2016 film)	Yes

Part1:

$Accuracy_{query\ 2} = 4/10 = 0.4$

$DCG_{query\ 2} = 1/log_2 5 + 1/log_2 6 + 1/log_2 8 + 1/log_2 11 = 1.4399112184$

Model with Query Relaxation:

$Accuracy_{query\ 2} = 5/10 = 0.5$

$DCG_{query\ 2} = 1/log_2 3 + 1/log_2 5 + 1/log_2 7 + 1/log_2 10 + 1/log_2 11 = 2.00787046359$

Query	Part 1	Relevance	Query Relaxation	Relevance
Salubrious Nourishment	The Nashville Sessions (Townes Van Zandt album)	No	Sai bhaji	Yes
	Eve Marie Shahoian	No	Intake Two	No
	Gospodor Monument Park	No	Blue Dot for Diabetes	No
	2016 FIVB Volleyball Men's Club World Championship	No	Maha Koraiem	Yes
	Fox hunting (film)	No	Haile Thomas	Yes
	Latin American and Caribbean Economic Association (LACEA)	No	Helga Schultze	No
	Charles Douglas Carpendale	No	Kelly Murumets	No
	Drumheller, Washington	No	Mandy's	Yes
	Earl A. Fitzpatrick	No	Maurice Guest (novel)	No
	Douglas Mountain (Washington)	No	Narinder Singh Randhawa	No

Part 1:

$$Accuracy_{query\ 2} = 0/10 = 0.0$$

$$DCG_{query\ 2} = 0$$

Model with Query Relaxation:

$$Accuracy_{query\ 2} = 5/10 = 0.5$$

$$DCG_{query\ 2} = 1 + 1/log_2 4 + 1/log_2 5 + 1/log_2 6 + 1/log_2 9 = 2.63296782997$$

Improvement 2:

1. **What are the issues with the vector space model built in part 1?**

   The basic vector space model built in Part 1 does not attach any semantic meaning to the words, and the documents are retrieved only on the basis of measures like term frequency, and inverse document frequency. As a result, this model may miss some important documents or retrieve some of the irrelevant information.

2. **What improvement are you proposing?**

   We propose to improve our model by introducing Latent Semantic Indexing.

3. **How does Latent Semantic Indexing improve upon the previous issues?**

   Addressing the above mentioned issue, latent semantic indexing retrieves documents considering the semantic meaning of the query, even if the query and the retrieved documents do not share words(or synonyms). This approach uses the principles of singular value decomposition(SVD) to identify patterns relationships between query terms and document content. Moreover, words used in same contexts tend to have similar meanings.

   Some of the issues with this approach are:
   a. LSI requires relatively high computational performance and memory in comparison to other information retrieval techniques. However, using a scalable python package solves this.
   b. Choosing the right number of dimensions for the application of SVD. Increasing the number of dimensions provides us with more specific results, but the curse of dimensionality also comes into the play.

4. **A corner case (if any) where this improvement might not work or can have an adverse effect?**
   The above suggested approach may fail if the query's semantic meaning is not relevant for the documents to be retrieved. In such a situation, the vector space model would yield better results, as they would exactly look for the query terms in the documents.

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

   Below are the example queries where an LSI-based model would yield more relevant documents than a basic vector space model.

Query	Part 1	Relevance	LSI	Relevance
Constitution of Honduras	2016–17 in Honduran football	No	Guerra de los Padres	Yes
	Carlos Zúñiga Figueroa	No	Kheri (Ludhiana West)	No
	Ladian Khurd	No	Khark (Ludhiana West)	No
	Ladian Kalan	No	Ladian Khurd	No
	Kheri (Ludhiana West)	No	Ladian Kalan	No
	Khark (Ludhiana West)	No	Feminism in Honduras	Yes
	Secretariat of public works, transport and housing (Honduras)	No	Talwandi Khurd	No
	1827 Honduran coup d'état	Yes	Qutbewal Gujjran	No
	Secretariat of Public Education (Honduras)	No	Secretariat of public works, transport and housing (Honduras)	No
	Thakarwal	No	Mannewal	No

$$Accuracy_{query\ 1-basic} = 1/10 = 0.1$$
$$DCG_{query\ 1-basic} = 1/log_2 9 = 0.315$$
$$Accuracy_{query\ 1-LSI} = 2/10 = 0.2$$
$$DCG_{query\ 1-LSI} = 1 + 1/log_2 7 = 1.356$$

Query	Part 1	Relevance	LSI	Relevance
space exploration	Lust In Space – Live At The National	No	Ekspress-AMU1	Yes
	Andrés Ruzo	No	Ekspress MD1	Yes
	BodyHack with Todd Sampson	No	JCSAT-16	Yes
	Splendor & Misery	No	Yamal-601	Yes
	Metric Systems Corporation	No	Next Generation Launcher	Yes
	International Electric Propulsion Conference	Yes	Ekspress (satellite constellation)	No
	Elizabeth Street Garden	No	Himawari 9	Yes
	Albert Socin	No	DS2000	Yes
	Eloy Urroz	No	JCSAT-4B	No
	Charles Watson Boise	No	Ekspress (satellite bus)	No

$$Accuracy_{query\ 2-basic} = 1/10 = 0.1$$
$$DCG_{query\ 2-basic} = 1/log_2 7 = 0.356$$
$$Accuracy_{query\ 2-LSI} = 7/10 = 0.7$$
$$DCG_{query\ 2-LSI} = 2.595$$

Query	Part 1	Relevance	LSI	Relevance
Who is the prime minister of australia?	Lifespan timeline of Prime Ministers of Singapore	No	Juvenile detention in the Northern Territory	Yes
	Costin Borc	No	Stella Maris, Darwin	No
	Electoral history of Bill Rowling	No	Royal Commission into the Protection and Detention of Children in the Northern Territory	Yes

	Electoral history of Norman Kirk	No	Mainland Australia	No
	Siber cabinet	No	Australian National Travel Association	Yes
	Medea (TV serial)	No	Australia's Shame	Yes
	Carbasea elegans	No	Mosaic of Rehob	No
	Governments of Mohammad Mosaddegh	No	Babungo (village)	No
	Secretary of State for International Trade	No	Left Unity (South Australia)	No
	Khurshid Gohar Qalam	No	Australian Law Librarians' Association	No

$$Accuracy_{query\ 3-basic} = 0/10 = 0.0$$
$$DCG_{query\ 3-basic} = 0$$
$$Accuracy_{query\ 3-LSI} = 4/10 = 0.4$$
$$DCG_{query\ 3-LSI} = 1 + 1/log_2 4 + 1/log_2 6 + 1/log_2 7 = 2.243$$

Query	Part 1	Relevance	LSI	Relevance
Places in India	Places of worship in Mavelikkara	Yes	Haryau	Yes
	Places of worship in Ernakulam	Yes	Ladian Kalan	Yes
	Aircraft Act	No	Ladian Khurd	Yes
	List of moths of India (Arctiidae)	No	Khark (Ludhiana West)	Yes
	Emmanuel Shearith Israel Chapel	No	Kheri (Ludhiana West)	Yes
	Carlos Ospina	No	Nurpur Bet	Yes
	Institutes of Technology Act	No	Majara Kalan	Yes

	KEPT-LP	No	Lalton Kalan	Yes
	2016 North American heat wave	No	Majra Khurd	Yes
	Kathura	Yes	Qutbewal Gujjran	Yes

$$Accuracy_{query\ 4-basic} = 3/10 = 0.3$$
$$DCG_{query\ 4-basic} = 1 + 1/log_2 3 + 1/log_2 11 = 1.919$$
$$Accuracy_{query\ 4-LSI} = 10/10 = 1.0$$
$$DCG_{query\ 4-LSI} = 4.543$$

Future scope and improvements:

In the above discussion, we implemented a basic vector space model and suggested two improvements to improve upon the relevance of documents retrieved. However, our implementation was tested on a small text corpus from Wikipedia. The internet is growing day by day, and making scalable systems is of future importance. So, to make our model scalable, more robust algorithms such as Page Rank could be tried. Also, the current implementation of our model is static in nature. Thus, improvements could be done to make the model search across the Web. Moreover, machine learning and natural language processing techniques could be employed to improve upon the accuracy.

References

1. C. D. Manning, P. Raghavan and H. Schutze. Introduction to Information Retrieval, Cambridge University Press, 2008.
2. Wargärde, N. (2020). Using WordNet Synonyms and Hypernyms in Automatic Topic Detection (Dissertation). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-97745
3. Latent Semantic Indexing API Reference https://radimrehurek.com/gensim/models/lsimodel.html