CS F469
Information Retrieval
Assignment-2

Vector Space-Based Information Retrieval System

Submitted By:
Miloni Mittal----------2017A3PS0243P
Asrita Venkata Mandalam----------2017A7PS1179P
Simran Sehgal----------2017A8PS0405P

March, 2020

Index

Introduction

The objective of this assignment was to build a vector space-based information retrieval system. There were two parts. The first part used a vector space model with lnc.ltc (SMART notation) as its scoring scheme. The second part consists of two attempts to improve the retrieval and ranking of the documents. The first method uses a spell-checker and lemmatization. The second uses LSI (Latent Semantic Indexing).

Part-1 (Implementation and Results)

A working ranked retrieval based IR system was built using the text corpus provided which consisted of 5196 documents and around 1 million tokens and 75971 unique tokens. The steps involved in making the information retrieval system were:

1. Preprocessing: Text was extracted from the corpus and stored document wise. All punctuation was removed.
2. Term-Document Incidence Matrix: Using CountVectorizer from the sklearn open source library, a matrix to record term count in each document was built.
3. Inverted Index: The inverted index was built using the same library as above and stored in a dictionary data structure with terms as keys and posting list as values.

For faster working of queries against the information retrieval system, the posting list, the normalisation factors for the lnc.ltc scoring and the titles were stored in .npy format.

4. Query Preprocessing: The query was preprocessed similar to corpus preprocessing done earlier
5. Tokenization of query: The query was tokenized using nltk and its word count was obtained.
6. Tf-idf weight of query: The tf-idf weight was calculated according to ltc standards i.e logarithmic term frequency weighting, idf into consideration (for this posting list was fetched), and cosine normalisation.
7. Score calculation: The document scores for each query were calculated. Safe ranking was assumed, hence no measures to reduce the number of documents in the sample space were employed.

Note: TF-IDF weights were normalised on the fly, when calculating the scores because it was decided to store the posting list in integer format, and not floats, as it is less memory intensive to do so.

Query	Doc/Score	Score	Relevant?
	The Singleton Argus	0.21029416894364944	yes
	Zamindar (newspaper)	0.20361207255968686	yes
	The Leader (Liverpool, New South Wales, newspaper)	0.20301613251083722	yes
	The Hawkesbury Chronicle and Farmers' Advocate	0.19960774138958182	yes
	Bell's Life in Sydney and Sporting Reviewer	0.19297338459477478	yes
	Hawkesbury Herald	0.1842399175908783	yes
	The Biz (newspaper)	0.18120786781577775	yes
	La Gaceta Mexicana	0.18089634823647371	yes
	The Colonist	0.17467773954683874	yes
	The Liverpool Herald	0.17458113476579865	yes
sadness symptom of depression kids	Encounter: The Killing	0.0579545263131198	no
	Dino Time	0.0563965842774143	no
	La Gaceta Mexicana	0.0521080616154968	no
	Boris McGiver	0.0516738640276069	no
	Bel Air (song)	0.0466369208971157	no
	Emotional granularity	0.0461806666210467	no
	Don Knowlton	0.0412588723955048	no
	Separation anxiety disorder	0.0386357134235659	yes
	Bernard Berkhout	0.0381092472515326	no
	Withrow Minstrels	0.0360090646507815	no

lieutenant colonel george	George Rowell	0.307747505907136	no
	Lorraine Campaign order of battle	0.185249130276384	no
	Penny George Institute for Health and Healing	0.173506907409289	no
	Federico Rauch	0.152870367190977	no
	116th Field Artillery Regiment	0.150552731538114	yes
	Brett King	0.123497285442986	no
	13th Manitoba Legislature	0.112613414530398	no
	Beech Aircraft Corp. v. Rainey	0.111927161296399	no
	George T. Felbeck	0.111927161296399	no
	James Walker (Surveyor General)	0.0975820793936705	no
mass communication	Anand Kumar (director)	0.190207799654943	no
	Golm Metabolome Database	0.178528792585701	no
	Mass song	0.173998756311767	no
	Adam Earnheardt	0.158048064932961	no
	Universidad de la Comunicación (México)	0.157854460760564	yes
	HD 27631 b	0.118770097862044	no
	NGC 1277	0.109628568121065	no
	Arctic Intermediate Water	0.10869671914267	no
	Goran Senjanović	0.105080431515396	no
	Krasimir Anev	0.0998064749783432	no
nuclear energy	Nuclear energy in Saudi Arabia	0.291921113085054	yes
	Shelby Brewer	0.220729403072995	yes
	Lisbon Protocol	0.170927324542277	yes
	Brahm Prakash	0.149676054761071	yes
	Surprisal analysis	0.140733359412483	no
	Flattop (critical assembly)	0.119965019035657	no
	George T. Felbeck	0.0997656452322478	no
	March of Progress (album)	0.0967107392888636	no

	D66 Strain of Chlamydomonas reinhardtii	0.0933641125769774	no
	William Lewis (physical chemist)	0.0900227407788223	no
how to wrestle	This Is How It Feels (album)	0.136839345351504	no
	Physical Data Flow	0.0714159285613651	no
	Abandonment rate	0.0683900480953025	no
	Crossmodal attention	0.0648996354421839	no
	Frankenberg family	0.0639010971434095	no
	Priceless (TV series)	0.057564753339246	no
	Surface and bulk erosion	0.056234367414543	no
	John C. Martin (businessman)	0.0554040494904479	no
	The Shield (professional wrestling)	0.0551623227787404	yes
	Perceptual dialectology	0.054024786106036	no
all time best movies	Gloria Sevilla	0.216892064360654	no
	All Esper Dayo!	0.171770216407979	no
	Saturn Award for Best Editing	0.149236625001752	no
	Nina Rillstone	0.148997466257174	no
	Temalangeni Dlamini	0.141472602767103	no
	Beijing Blues	0.138049906725266	yes
	All Alone (pornographic film series)	0.134110485372917	yes
	Golden Horse Award for Best Feature Film	0.125668615332629	no
	All India Station Masters' Association	0.122650832070501	no
	Paul Brannigan	0.120998661094719	no
gender differences in various countries	Gender, Work and Organization	0.144036696030851	no
	List of countries by titanium production	0.142692983916779	no
	Gender inequality in China	0.124605556311074	yes
	Gender inequality in Thailand	0.121644550487688	yes

	Turn-taking	0.110733967798424	no
	International Journal of Social Welfare	0.104280370753464	no
	International Olympic Committee and gender equality in sports	0.102194888964397	no
	Saloca kulczynskii	0.0925671505658321	no
	Perceptual dialectology	0.0913649467368224	no
	Media of Honduras	0.0827361924149129	no
open source coding	High Efficiency Video Coding tiers and levels	0.153191417480012	no
	Open Hardware and Design Alliance	0.135011880273173	yes
	Service-Oriented Localisation Architecture Solution	0.113421136612281	yes
	DPRK Amateur Golf Open	0.107695808116605	no
	Turtlebot	0.100038124319342	no
	2013 Heineken Open	0.0949112724161303	no
	2013 Ladies European Tour	0.0911371030476091	no
	WebSharper	0.0905289661937867	yes
	Seven Springs, Gloucestershire	0.0900018059121665	no
	ImageNets	0.0880074649839709	yes
night crickeu	Small Town Saturday Night	0.54166592046737	no
	Blackout Wednesday	0.228369633648202	no
	Sirota's Court	0.216409766977747	no
	Mike Schank	0.184894469583753	no
	Télétoon la nuit	0.18426939706933	no
	Werkdiscs	0.15665375745304	no
	Brad Scott (fighter)	0.143893309375075	no
	Devorah Frost	0.117735608582043	no
	Parasola plicatilis	0.115660926968265	no
	Priceless (TV series)	0.113868474654981	no

Part-2 (Improvements)

i. First Improvement

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

When the user enters queries which have a small spelling mistake, the previous IR system would eliminate that word from the query dictionary. We want a more robust system which is relatively more resistant to minor spelling errors.
Moreover, it might so happen that the form of word used in the query might not exist in the corpus or vice versa. That is why, it is necessary to bring the words in both query and document to a common base form for better comparison.

What improvement are you proposing?

To solve the first problem, we added a spelling corrector to the query. This spelling corrector would use Levenshtein distance to find the word which is closest to the given wrongly spelled word. An assumption is made that all the spellings in the documents are correct and they are not to be changed.
To solve the problem of different forms of words, we used lemmatizer on the query and the document.

How will the proposed improvement address that issue?

The proposed improvement solves the problem by correcting wrong spellings and bringing all the words to a common base form thus making more relevant comparisons.

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

A corner case might occur with ambiguous words. For example, “mass” has three meaning in three different contexts as in: “mass communication”, “Christian mass” and “physical mass”. On typing “mass communication” as the query, the results include documents with mass referring to physical mass and other such irrelevant documents.

Demonstrate the actual impact of the improvement. Give three queries, where the improvement yields better results compared to the part 1 implementation.
The following are a few examples of queries where this worked better than the vector space model.

Key: Highlighted documents are relevant

Query: night crickeu

Part 1	Improvement 2
Small Town Saturday Night	Lahore City Cricket Association Ground

Blackout Wednesday	1977 Benson & Hedges Cup
Sirota's Court	Paramount Cricket Promotion Association Stadium
Mike Schank	Small Town Saturday Night
Télétoon la nuit	1976 Benson & Hedges Cup
Werkdiscs	North Yorkshire and South Durham Cricket League
Brad Scott (fighter)	List of international cricket centuries by Mohammad Azharuddin
Devorah Frost	Malnad Gladiators
Parasola plicatilis	Fanchon the Cricket

Priceless (TV series)	Bahawalpur Stags

The vector space model retrieved 0 relevant documents because of the minor spelling error of the word cricket. In contrast, the first improvement corrected the spelling by employing Levenshtein distance and returned relevant documents.

Query: sadness symptom of depression kids

Part 1	Improvement 2
Encounter: The Killing	List of medical triads and pentads
Dino Time	Cerebellar activation
La Gaceta Mexicana	Separation anxiety disorder
Boris McGiver	Autosomal recessive cerebellar ataxia
Bel Air (song)	La Gaceta Mexicana

Emotional granularity	Dino Time
Don Knowlton	Luck Films
Separation anxiety disorder	William D. Steers
Bernard Berkhout	Encounter: The Killing
Withrow Minstrels	Emotional granularity

As the second improvement works by lemmatization, it gave the relevant document a higher ranking since more words in the document were related to the query after lemmatization.

Query: How to wrestle

Part 1	Improvement 1
This Is How It Feels (album)	Sargis Tonoyan
Physical Data Flow	Denis Forov

Abandonment rate	Karine Shadoyan
Crossmodal attention	Arman Geghamyan
Frankenberg family	Vadim Laliev
Priceless (TV series)	Mahmed Aghaev
Surface and bulk erosion	Arayik Gevorgyan
John C. Martin (businessman)	Araik Baghdadyan
The Shield (professional wrestling)	Artur Shahinyan
Perceptual dialectology	This Is How It Feels (album)

Again, the first improvement gave a large number of relevant results as it retrieved everything related to wrestling but the vector space model was stuck because it searched strictly for the query words.

II. Second Improvement

What is the issue with the IR system built in part 1? What improvement are you proposing? How will the proposed improvement address that issue?

Latent Semantic Indexing (LSI) is a technique that helps retrieve documents based on the meaning or topic of the document. It is an extension and improvement on the vector space model. The vector space model created in Part 1 does not take into consideration synonyms of words. For example, if the user wants to find articles about ‘Public Speaking’, he might search for ‘Talking in public’. The LSI model takes into consideration synonyms of words (and words that are close in meaning) and will return documents about ‘Public Speaking’ and ‘Talking in public’ with a better ranking than what the vector space model returns.

One disadvantage that we noticed is that it takes the same amount of runtime as the vector space model. Also, while the gensim library helps with its implementation, the optimal number of dimensions (num_topics parameter) for the model had to be manually tested and changed.

For our implementation, first, we took the document as an input and split it into a list of titles and a list of documents. Next, we removed stopwords and got the frequency of each of the

terms. After making a dictionary out of the text, we used doc2bow to create a sparse vector. Using the gensim library, we created a tf-idf model of the corpus and then ran an LSI model on it. After manually testing it multiple times, we decided to keep the number of dimensions as 4000. We trained the corpus and query vectors on that model. After calculating the similarities, we sorted the output in descending order to obtain the result of the top documents.

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

There are a few cases where this method might return the required document one or two ranks worse than the vector space model. This occurs when the query is vague. The LSI model looks for similar documents based on synonyms as well so this is probably why this corner case occurs.

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

The following are a few examples of queries where LSI worked better than the vector space model.

Query: gender differences in various countries

Part 1	Improvement 2
Gender, Work and Organization	Gender inequality in China
List of countries by titanium production	Gender inequality in Thailand
Gender inequality in China	Gender, Work and Organization
Gender inequality in Thailand	List of countries by titanium production
Turn-taking	International Olympic Committee and gender equality in sports

International Journal of Social Welfare	Turn-taking
International Olympic Committee and gender equality in sports	List of medical triads and pentads
Saloca kulczynskii	International Journal of Social Welfare
Perceptual dialectology	Kenya–Malawi relations
Media of Honduras	Saloca kulczynskii

We observed that the LSI gave the more relevant documents at the top than the vector space model because the LSI preserves and gives importance to context rather than just regular matching word expressions.

Query: mass communication

Part 1	Improvement 2
Anand Kumar (director)	Universidad de la Comunicación (México)
Golm Metabolome Database	Mass song
Mass Song	Golm Metabolome Database
Adam Earnheardt	Adam Earnheardt

Universidad de la Comunicación (México)	Anand Kumar (director)
HD 27631 b	Extreme mass ratio inspiral
NGC 1277	NGC 1277
Arctic Intermediate Water	Goran Senjanović
Goran Senjanovic	Krasimir Anev
Krasimir Anev	Calostomal

The second improvement, based on LSI works on preserving context and solves the problems of synonyms and polysemy and thus gave the relevant result at a higher rank.

Query: MA performance arts

Score	Improvement 2
Sport, Exercise, and Performance Psychology	Master of Performing Arts
Master of Performing Arts	Die Augen der Mumie Ma
Sharon Hayes (artist)	Sport, Exercise, and Performance Psychology

Ulvi Azizov	Secretary of State for Culture, Arts and Sports
Electronic Café International	Sharon Hayes (artist)
Die Augen der Mumie Ma	Withrow Minstrels
Andrew Surmani	Győr-2
Secretary of State for Culture, Arts and Sports	Electronic Café International
Center Theater (Hartsville, South Carolina)	Ipswich Town Hall
Rachel Lachowicz	Giovanni Battista Armenini

Again, LSI worked on synonyms and related data, preserving context, and hence gave better results by relevance as compared to the Vector Space Model.

Conclusion

We have built a working vector space information retrieval system using lnc.ltc SMART notation. It works well for basic searches but fails where we add ambiguous words, spelling errors, synonyms etc. We suggested two improvements for the same. The first improvement works on building a more robust system with spell check in queries and lemmatizing the corpus. The second improvement focuses on getting results even with synonymous words.

References

1. https://nlp.stanford.edu/IR-book/html/htmledition/document-and-query-weighting-schemes-1.html
2. https://www.cse.iitk.ac.in/users/nsrivast/HCC/ranked%20retrieval.pdf
3. https://medium.com/@acrosson/summarize-documents-using-tf-idf-bdee8f60b71
4. https://pypi.org/project/pyspellchecker/
5. https://stackoverflow.com/questions/43288550/iopub-data-rate-exceeded-in-jupyter-notebook-when-viewing-image
6. https://www.geeksforgeeks.org/python-lemmatization-with-nltk/
7. https://github.com/Prakhar0409/Latent-Semantic-Indexing?files=1
8. https://radimrehurek.com/gensim/index.html
9. https://scikit-learn.org/stable/