A REPORT
ON

VECTOR-SPACE BASED INFORMATION RETRIEVAL SYSTEM

BY

Khimani AbdulKadir Salim	2017B4A70696P
Amisha Kothari	2017A3PS0194P
Rithik Dilip Rathi	2017A3PS0266P
Srishti Gupta	2017A3PS0293P
Abhinava Arasada	2017A7PS0028P

in partial fulfilment of the course

CS F469 - Information Retrieval
(November, 2020)

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 title weighting.The second uses Bigram Indexing.

PART 1 : Implementation & Results

A working ranked retrieval based IR system was built using the text corpus provided which consisted of 8392 documents and 231217 unique tokens. The working can be briefly explained as:

Text was extracted from the corpus and was stored document wise, including the document ID and the Title. Vocabulary for the whole corpus was created by word tokenizing the text contained in all the documents. Term frequency for all the documents were done and the normalized logarithmic tf weight of all the documents were stored in a dictionary variable. For all the terms in the vocabulary, the IDF was calculated and stored. The tf-document weight, term idfs, document dictionary having the doc ID and title, and the vocabulary was stored as a byte stream file using the ‘pickle’ library of Python. These were stored so that the index creation process need not be repeated and the files saved can directly be used for searching a query in the second script. For the query, punctuations were removed and tf-query weights were computed for the query terms. Using this and the already saved idf weights, the normalized tf-idf weights of the query terms were calculated and finally the score of a particular document is computed as the product of this and the tf-document weight. The top 10 documents with their respective ID,title and scores were returned for each query.

The results for 10 free text queries are shown in the tables below. These queries cover several cases including common nouns, rare terms, etc.

Query	Top 10 documents	Score	Relevant?
dangerous diseases	Lists of diseases	0.16848720498866096	Yes
	Mariamman Temple, Pretoria	0.11466771130069804	No
	Biomedical Primate Research Centre	0.07891832656730208	No
	Gestational hypertension	0.06955127989989755	Yes
	Hasinai	0.06569344772907612	No
	Winchester Magnum	0.0575771955281371	No
	Amin al-Majaj	0.057476045482328764	No
	Alexander disease	0.05707385371997221	Yes
	Romanian School of Neurology	0.056179404714300045	No
	Antiserum	0.055442919365774376	No

Query	Top 10 documents	Score	Relevant?
famous movie actors	La risa en vacaciones	0.14261999977196438	No
	Kleine Freiheit	0.1334493244931549	No
	Arts Vision	0.11937180444686413	No
	Battle of Chaeronea	0.11936744896725532	No
	Otto Waalkes	0.09340571968040434	Yes
	Harry Reichenbach	0.08423881267569733	Yes

	Mary Dresselhuys	0.08241886226492931	Yes
	Songs from the Last Century	0.0818128872307335	No
	Fernando Allende	0.08100234734779946	Yes
	Benoît Poelvoorde	0.0775783700420913	Yes

Query	Top 10 documents	Score	Relevant?
snooker rankings	Snooker world rankings 1987/1988	0.35313615938292325	Yes
	Snooker world rankings 1986/1987	0.35313615938292325	Yes
	Snooker world rankings 1989/1990	0.3148085005918162	Yes
	Snooker world rankings 1988/1989	0.3148085005918162	Yes
	Tony Jones (snooker player)	0.17477803834936378	No
	Gary Wilkinson (snooker player)	0.13151190451827133	No
	Malta Grand Prix	0.13073014710290926	No
	Martin Clark (snooker player)	0.11531056520789544	No
	Irish Open (snooker)	0.11240266762377325	No
	Troy Shaw	0.09918666390078278	No

Query	Top 10 documents	Score	Relevant?
	Granite Mountains (Alaska)	0.28097181642094926	Yes

mountain range	Granite Mountains	0.19262381524013686	Yes
	West Humboldt Range	0.17811185430735138	Yes
	Sepree River	0.17183306969096207	No
	Grundarfjörður	0.14705699663745558	No
	Truong Son muntjac	0.14321242209165147	No
	Pangaion Hills	0.13878295080866215	Yes
	Kerlingarfjöll	0.13764980600219862	Yes
	Meander Valley Council, Tasmania	0.13356099854253634	No
	Granite Mountains (California)	0.13206832063882334	Yes

Query	Top 10 documents	Score	Relevant?
popular novel series	The Twins (novel)	0.13922734585459667	Yes
	Dalziel and Pascoe	0.1272914944797203	No
	Eric Pierpoint	0.12253775281339108	No
	List of Puerto Rican television series	0.1096769001735634	No
	Lincoln Child	0.10847636865870688	No
	Ting Hai effect	0.10804593944330901	No
	Nanosite	0.10762832355507755	No
	Gossip Girl (novel series)	0.1062942161492046	Yes
	New Spring	0.10404507901823899	Yes

	Eleftheria i thanatos	0.10048956362701694	No

Query	Top 10 documents	Score	Relevant?
types of wrestling matches	SummerSlam	0.08792511214845493	Yes
	Tag team	0.08605204819082168	Yes
	Frontier Wrestling Alliance	0.08022451201217694	No
	Hell in a Cell	0.07960249745333442	Yes
	Jonny Storm	0.07818785758090258	No
	Archery at the 1996 Summer Olympics	0.07732788091374787	No
	Tables, Ladders, and Chairs match	0.07672585438561377	Yes
	"I Quit" match	0.07166831423315281	Yes
	WWE 2K	0.06858031499668037	No
	Kevin Thorn	0.06856036430904278	No

Query	Top 10 documents	Score	Relevant?
political figure	Sándor Rónai	0.20935870935191891	Yes
	Geoffrey Robert Gardner	0.18541999682094512	Yes
	White Horse	0.14581671931788404	No
	John Wilbur Dwight	0.12939343884181817	Yes
	Émile Eddé	0.12503565378421297	Yes
	Mohammad Hashim Khan	0.1248710689177048	Yes

	Harry Lane Englebright	0.12424801389416174	Yes
	Andrew Cunningham (politician)	0.12364683470810717	Yes
	Jumblatt family	0.11327360307876783	No
	Juan Antonio Lavalleja	0.10976015716945421	Yes

Query	Top 10 documents	Score	Relevant?
what is virtual router redundancy protocol (rare terms)	Virtual Router Redundancy Protocol	0.13382872310445162	Yes
	Route distinguisher	0.08206376690364178	No
	Virtual sit-in	0.06517498226566716	No
	XIO	0.05175950473707774	No
	Noesis Cultural Society	0.0461468646577635	No
	H.245	0.04537782356643494	No
	/dev/random	0.04502460963552299	No
	Dreamscape (chat)	0.04270648902328446	No
	TTCN	0.04134172061306803	No
	Client-to-client protocol	0.040891941381650865	No

Query	Top 10 documents	Score	Relevant?
different programming languages	List of audio programming languages	0.3713618631281824	Yes
	Fifth-generation programming language	0.15175354739205066	Yes
	DataFlex	0.14602394128066482	Yes
	Joy (programming language)	0.12698810389359366	Yes
	DirectSetup	0.11964266696833264	No
	Perl Data Language	0.11155016184901034	Yes
	Deterministic algorithm	0.1091826670407435	No
	TTCN	0.1085180325686191	Yes
	Lispkit Lisp	0.10435673740941635	No
	Euler (programming language)	0.1009588811467306	Yes

Query	Top 10 documents	Score	Relevant?
types of penguin genus	Megadyptes	0.2920270382686878	Yes
	Eudyptula	0.15738158781747427	Yes
	Aptenodytes	0.1477643919641482	Yes
	Crested penguin	0.11417940527398993	Yes
	Raven (disambiguation)	0.10817787510844402	No

	Pygoscelis	0.1004977110780258	Yes
	Nama (plant)	0.09783324399999013	No
	Rutilus	0.09238085691879813	No
	Arapaima gigas	0.09228127472413741	No
	Penguin sweater	0.09011833250325238	No

PART 2 :

Improvement 1 - Title weighting to the query-document scores

1. The current implementation follows the basic vector space model, considering the title and body of a doc as an equal entity and ranking a document based on the product of the tf-idf weights of term in the query and the tf weights in the documents. The consideration of title and body as an equal entity while searching for a query is a potential drawback, as the title of a document plays a significant role in giving us a basic idea about the content of the document and should be given more weightage. This will work very well as we are working on HTML pages of web (Wikipedia) corpus and we can easily compare the query with the ‘title’ metadata of the page and give more weightage to that particular document if the term matches. So, instead of the basic zone weighting, we use the parameterized weighting to increase weight in case of title-match.
2. While searching for the query in the second script, we increase the scores of documents which have query terms present in its title(after converting it to lowercase, as it has in general first character capitalised) by a value of 0.10 for each term, thereby giving it a more preference over those which have the query terms only in the content. For this, the title of each document was converted into lowercase and then the query terms were compared with the title.
3. A drawback of this improvement could be that it would also retrieve documents which have misleading titles, i.e. a very vague title which does not relate well to the content and that could mislead the user. Also, unnecessary increasing weighting due to stop words being present in the title is wrong, so the query tokens was processed for removing the stop words and then passed for further process.

Here, we demonstrate the benefits from these improvements when compared to the basic vector space based ranked retrieval model by taking examples of 3 queries and comparing the retrieved documents and their scores.

Let's say the user wants to search everything available about the query **'New York'**, the results from the different models are given as:

Without Improvement:

Query	Top 10 documents	Score	Relevant?
new york	1912 in architecture	0.2493925558989286	No
	Neochori	0.24475804654544173	No
	Pieksämäen maalaiskunta	0.19081955836751405	No
	Virtasalmi	0.18743757126850413	No
	AZS Częstochowa	0.16011089608001441	No
	Irish Open (snooker)	0.1500173268072722	No
	Source upgrade	0.13474514145117114	No
	Jäppilä	0.13082221093811516	No
	Burgães	0.12953629385705356	No
	Ilinka Mitreva	0.12706721449706201	No

With Improvement:

Query	Top 10 documents	Score	Relevant?
	1912 in architecture	0.2493925558989286	No
	St. Joseph's College (New York)	0.2478227496026887	Yes
	New York-New York Hotel and Casino	0.24772392650798264	Yes

new york	New York City Opera	0.24669491886554534	Yes
	New York, Rio, and Buenos Aires Line	0.24510310597457616	Yes
	Neochori	0.24475804654544173	No
	New York Medical College	0.22934380345492197	Yes
	Peterboro, New York	0.2	Yes
	W (New York City Subway service)	0.2	Yes
	United States District Court for the Southern District of New York	0.2	Yes

Let's say the user wants to search everything available about the query **'various cricket rules'**, the results from the different models are given as:

Without Improvement:

Query	Top 10 documents	Score	Relevant?
various cricket rules	Pro Cricket	0.10874131823462803	No
	Structural rule	0.10260719316029882	No
	Wide (cricket)	0.09034134816869807	Yes
	Dean Headley	0.08706511397712714	No
	Tim Curtis	0.08394376192474927	No
	Tuart Hill, Western Australia	0.08318808974327248	No
	The Twelfth Man	0.08041790109890397	No
	Cricket Australia	0.07858913349214164	No
	Yūko Miyamura	0.07820836671622201	No

	Frederick Wills (Guyana)	0.07512814572463304	No

With Improvement:

Query	Top 10 documents	Score	Relevant?
various cricket rules	Pro Cricket	0.20874131823462805	No
	Wide (cricket)	0.19034134816869808	Yes
	Cricket Australia	0.17858913349214167	No
	Bye (cricket)	0.16065528569087484	Yes
	Toss (cricket)	0.16037775942253968	Yes
	Appeal (cricket)	0.14971097702026673	Yes
	Road Rules: The Quest	0.1476669581327694	No
	Nottinghamshire County Cricket Club	0.14603644794628645	No
	Warwickshire County Cricket Club	0.14269975343179217	No
	Road Rules	0.1359142819779103	No

Let's say the user wants to search everything available about the query **'famous comics to read'**, the results from the different models are given as:

Without Improvement:

Query	Top 10 documents	Score	Relevant?
	Cat (comics)	0.294762881395818	Yes
	1965 in comics	0.21075966070451857	No
	1964 in comics	0.21075966070451857	No
	1950s in comics	0.2093488523881364	No
	Battle of	0.10303537866501279	No

famous comics to read	Chaeronea		
	1940s in comics	0.09417554951669142	No
	Evan Dorkin	0.08697737352783333	No
	S.A.M.	0.08441194837447241	No
	Juan Díaz Canales	0.08228350699904297	No
	Comic Book Legal Defense Fund	0.07526909974701776	No

With Improvement:

Query	Top 10 documents	Score	Relevant?
famous comics to read	Cat (comics)	0.3901232866513842	Yes
	1965 in comics	0.31078751863428	No
	1964 in comics	0.31078751863428	No
	1950s in comics	0.3093765238391436	No
	1940s in comics	0.19312789602769914	No
	Fawcett Comics	0.1575444239248944	Yes
	Archie Goodwin (comics)	0.14203256724740707	Yes
	Bone (comics)	0.13599861372506972	Yes
	Ghost Rider (comics)	0.13012585445666497	Yes
	Warlord (comics)	0.12807168903167593	Yes

Improvement 2 - Bigram Indexing

1. The current vector space system considers each word separately to calculate tf-idf score. It does not account for queries with two tokens constituting a single semantic unit (such as United States,Notre Dame etc.). The current model also retrieves documents containing individual matching words,which may not be useful to the user.
2. Our approach: We compute bigram weights for all the terms in the dictionary and bigram weights for most frequent combinations i.e top K bigrams (where K=1000). We've used a combination of two approaches namely, naive unigram weighing model and calculating top K most frequent combinations using chi-square scores after case folding. Thus our score formula is as follows, score= w1*unigramWt + w2*bigramWt, where w1 +w2=1 and w1= 0.6, w2= 0.4.
3. With the current improvement our model can also handle the cases where bigrams are not present, as we have taken unigram weight into account.
4. The proposed implementation may not be efficient for large corpus as the time taken to compute indexes increases. Also, some bigrams may not be considered as this approach takes only top K collocations in account.

   Here, we demonstrate the benefits from these improvements when compared to the basic vector space based ranked retrieval model by taking examples of 3 queries and comparing the retrieved documents and their scores.
   We noticed that even in cases where the basic query returns mostly relevant documents, this improvement not only helps us filter out irrelevant documents, but additionally refines the order of the documents retrieved.

Query1 - 'Sorting Hat'

Without Improvement:

Query	Top 10 documents	Score	Relevant?
Sorting Hat	British Columbia Highway 12	0.061793193950346784	No
	Harry Potter and the Chamber of Secrets (film)	0.05662176846324691	Yes
	Meadow Fresh	0.05314391912758644	No
	Poodle Hat	0.05060632410401693	No

	Eazel	0.05030328246282882	No
	Magical objects in Harry Potter	0.04084807345089124	Yes
	Harry Potter and the Philosopher's Stone (film)	0.037789557804376026	Yes
	Marzullo's algorithm	0.03707272364274042	No
	Sean Hayes (actor)	0.03430678135653547	No
	Free as in Freedom	0.03409390687937386	No

With Improvement:

Query	Top 10 documents	Score	Relevant?
Sorting Hat	Harry Potter and the Chamber of Secrets (film)	0.2544467906678632	Yes
	Magical objects in Harry Potter	0.19425322378378376	Yes
	Harry Potter and the Philosopher's Stone (film)	0.1566257610433203	Yes
	Harry Potter and the Deathly Hallows	0.10276955486797032	Yes
	British Columbia Highway 12	0.037075916370208066	No
	Meadow Fresh	0.03188635147655186	No
	Poodle Hat	0.030363794462410155	No
	Eazel	0.03018196947769729	No
	Marzullo's algorithm	0.02224363418564425	No
	Sean Hayes (actor)	0.02058406881392128	No

Query2- 'Climate Change'

Without Improvement:

Query	Top 10 documents	Score	Relevant?
Climate Change	Climate Change Levy	0.14938543626272327	Yes
	Christopher Lee (historian)	0.09627135679228266	No
	Craig D. Idso	0.09030877809650013	Yes
	United Kingdom Climate Change Programme	0.0886703833525894	Yes
	Queen Elizabeth Islands	0.06886864101885942	No
	Karl Kruszelnicki	0.059271174679319444	Yes
	Moorish Delta 7	0.05863846922049858	No
	Malawian general election, 2004	0.050753987711827486	No
	Keith E. Idso	0.04719380046980173	Yes
	Sherwood B. Idso	0.04583887131224746	Yes

With Improvement:

Query	Top 10 documents	Score	Relevant?
Climate Change	Climate Change Levy	0.489631261757634	Yes
	Craig D. Idso	0.4541852668579001	Yes
	United Kingdom Climate Change Programme	0.4532022300115537	Yes
	Karl Kruszelnicki	0.4355627048075917	Yes
	Moorish Delta 7	0.4351830815322992	No

	Christopher Lee (historian)	0.38899592908305397	Yes
	ConAgra Foods	0.30364982250619466	Yes
	Queen Elizabeth Islands	0.28451556321502797	No
	Hydrogen economy	0.18388066376977394	Yes
	National Museum of Natural History	0.1669716438648528	No

Query3 - 'Prime Minister'

Without Improvement:

Query	Top 10 documents	Score	Relevant?
Prime Minister	Fredrik Gyllenborg	0.2704338677113459	Yes
	Minister of Civil Aviation	0.2559226558444806	No
	Prime Minister of Zambia	0.25552817064498534	Yes
	List of fictional Prime Ministers of the United Kingdom	0.18774067775403594	Yes
	Prime Minister of Afghanistan	0.16295873945981898	Yes
	Nelson Oduber	0.16158267535437013	Yes
	Poul Schlüter	0.16044928963929145	Yes
	Adamantios Androutsopoulos	0.15958808398430935	Yes
	Jóannes Eidesgaard	0.15918698289638955	Yes
	Prime Minister of Turkey	0.15869044903399865	Ye

With Improvement:

Query	Top 10 documents	Score	Relevant?
Prime Minister	Fredrik Gyllenborg	0.5622603206268075	Yes
	Prime Minister of Zambia	0.5533169023869913	Yes
	List of fictional Prime Ministers of the United Kingdom	0.5126444066524216	Yes
	Prime Minister of Afghanistan	0.49777524367589143	Yes
	Nelson Oduber	0.4969496052126221	Yes
	Poul Schlüter	0.4962695737835749	Yes
	Adamantios Androutsopoulos	0.49575285039058564	Yes
	Prime Minister of Turkey	0.4952142694203992	Yes
	Hari Kostov	0.4915902811871694	Yes
	Deutsche Akademie	0.4902912710426258	Yes