ASSIGNMENT-2
INFORMATION RETRIEVAL
REPORT ON VECTOR-SPACED BASED INFORMATION RETRIEVAL SYSTEM

Keshav Sharma	**2017A7PS0140P**
Burhan Boxwalla	**2017A7PS0097P**
Pranav Panchumarthi	**2017A7PS0153P**

OBJECTIVES

1) Building a Ranked Information Retrieval System based on the Vector Space Model.

2) Building proper Indexes which will take a large collection of text corpus and output a structured data structure.

3) Improve upon the system by solving the limitations caused by the Vector Space Model

ASSUMPTIONS

1. Context doesn't matter in the free text query given by the user.
2. The indexing files containing the Posting Lists fit in the memory
3. Terms can appear in the document independent of each other

LIMITATIONS

1. Phrasal queries may be ranked lower depending on frequency of terms in other documents. Eg. For the query 'hello how are you doing', the document containing the phrase exactly may come lower compared to a document containing 2 'hello's, 4 'doing's.
2. The terms in the user query which have probably been spelt wrong are corrected with the most probable suggestion and re-run. The spell check may yield a differently intended spelling. Eg. 'im' is not corrected to 'in', 'muf' is corrected to 'mum' instead of 'mud'.
3. Vector-space model provides a not-so-relevant retrieval model because of the bag of words nature of it.
4. The Spell Check in the improved model doesn't take into consideration the weighted Levenshtein distance, wherein the correction of the wrongly typed letter will be closer on the keyboard. For instance, 'mud' is a better candidate for 'muf' instead of 'mum' since 'd' is closer on the keyboard to 'f' rather than 'm'.

ALGORITHMS

I. INDEX CREATION

The raw text given is parsed and tokenized. Dictionary data structure is used because of its constant look-up time. For each document parsed, the tokens are populated in the dictionary along with their frequency. Terms are populated as and when they appear in the text corpus.

The dictionary consists of term as the key and the value as another dictionary with document ids where the term appears as the key along with the frequency of the term in that document as the value.

Eg. Posting_List[term] = {docid1: freq1, docid2: freq2, …….}

Another Posting List containing the terms as the lemmatized tokens of the previous List is made.

After all the documents are parsed, the indexes are dumped in the pickle files which will be used by query parser. It is to note that the punctuation symbols are removed.

II. LNC.LTC MODEL

Vector Space Model is a mathematical way of representation of documents and queries as vectors. Each dimension corresponds to a separate token. If a token occurs in a document, its value in the vector is non-zero. In this assignment, we have used **tf-idf** weighting scheme. **tf** refers to the value of 1 plus logarithm of frequency of occurrence of the term in the document. **idf** refers to the logarithm of ratio of total number of documents indexed to the number of documents containing the term.

Mathematically, $w_{t,d} = 1 + \log_{10} tf_{t,d}$ and $idf_t = \log_{10}(N/df_t)$

Where

$\mathbf{tf_{t,d}}$ = number of occurrences of term t in document d,

$\mathbf{df_t}$ = number of documents containing the term t and

$\mathbf{N}$ = total number of indexed

After creating the vectors for query and documents using the lnc.ltc SMART notation (using normalization), scalar multiplication of vectors is done and top 10 documents are retrieved. In lnc.ltc, the documents are weighed using logarithmic term frequency along with cosine normalization and query terms are weighed using **tf-idf** weighing scheme with cosine normalization.

III. SPELL CHECK

The SpellChecker library from pySpellChecker is invoked. For a query term wrongly spelled, it first checks the Part-Of-Speech tag. If it is a proper Noun, it is ignored. Or else, it corrects the spelling based on the minimum Levenshtein distance to a probable candidate.

Levenshtein distance tells us about the minimum number of moves (adding a letter/deleting a letter/replacing a letter) to convert a string of characters to other.

IV. LEMMATIZATION

Lemmatization is the process of converting a word to its morphological base. This process is done using a simple lookup on the WordNet database. Lemmatization increases the number of documents retrieved as it looks at the bases of the word rather than its simple usage. Hence the query and posting list both are lemmatized.

V. QUERY PARSING

The lemmatization and spell check optimizations are only applied if we are not able to retrieve the specified number of documents.
The order is as follows:

1. The raw query is used for retrieval.
   If <10 documents are retrieved, then
2. Lemmatize the query and repeat the retrieval
   If still <10 documents are retrieved, then
3. Apply spell check on the query and repeat the retrieval.

EVALUATION RESULTS

1) Part-I Results

Query - 1	Top 10 Documents	Score	Is the document relevant to the query?
IP Addressing	Samuel Dumoulin, ID: 6088573	0.7443	No
	Washington University School of Law, ID: 6091448	0.7443	No
	Motorola Canopy, ID: 6098171	0.7443	Yes
	Megaupload, ID: 6100089	0.7443	Yes
	Web template system, ID: 6093430	0.6679	No
	Tango Magazine, ID: 6095757	0.6679	No
	Ellis Carver, ID: 6106597	0.6679	No
	Jill Derby, ID: 6108009	0.6679	No
	Sonnet 154, ID: 6109814	0.6679	No
	Campaign bus, ID: 6111759	0.6679	No

Query - 2	Top 10 Documents	Score	Is the document relevant to the query?
Nikolai Klyuev Documentary	Nikolai Klyuev, ID: 6103332	0.9970	Yes
	Eurasianism, ID: 6090994	0.5372	No
	Nikolay Govorun, ID: 6096497	0.5372	No
	Dmitrii Menshov, ID: 6096579	0.5372	No
	Lyudmila Smirnova, ID: 6096681	0.5372	No
	Nikolai Chudakov, ID: 6096754	0.5372	No
	Nikolai Kochin, ID: 6097057	0.5372	No
	The Goblin Mirror, ID: 6106753	0.5372	No
	Battle of Tashihchiao, ID: 6108562	0.5372	No
	Ivan Privalov, ID: 6111554	0.5372	No

Query - 3	Top 10 Documents	Score	Is the document relevant to the query?
Clive Uptton Death	Clive Uptton, ID: 6103385	0.95752	Yes
	Harris Savides, ID: 6102888	0.57045	No
	The Nutt House, ID: 6089645	0.56544	No
	William Paterson (Canadian politician), ID: 6092972	0.56544	No
	Stephen Pope, ID: 6106676	0.56544	No
	Eddie Burns, ID: 6107454	0.56544	No
	Tim Pickup, ID: 6108194	0.56544	No
	Ron Willey, ID: 6108389	0.56544	No
	Clive Spong, ID: 6111030	0.56544	No
	Tom Hunting, ID: 6116528	0.56544	No

Query - 4	Top 10 Documents	Score	Is the document relevant to the query?
bands from canada	Jay Ryan (artist), ID: 6089865	0.81081	No
	Nelson Miller, ID: 6096053	0.81081	No
	West Florence High School, ID: 6102103	0.81081	No
	Bass Festival, ID: 6103944	0.81081	No
	Sean McNabb, ID: 6107570	0.81081	No
	Richland High School (Texas), ID: 6110253	0.81081	No
	Norway in the Eurovision Song Contest 2005, ID: 6115142	0.81081	No
	Defence Medal (United Kingdom), ID: 6115415	0.73238	No
	History of the Northwest Territories, ID: 6088904	0.71025	No
	Jarle Vespestad, ID: 6087714	0.70101	No

Query - 5	Top 10 Documents	Score	Is the document relevant to the query?
Black actress	Jacqui Gordon-Lawrence, ID: 6111521	0.99999	Yes
	Malinda Williams, ID: 6103957	0.99813	Yes
	Yoon Eun-hye, ID: 6115380	0.99312	No
	A Tale of Two Cities (Lost), ID: 6096324	0.99166	No
	Aurelio Zen, ID: 6101457	0.99166	No
	Martin Olson, ID: 6103150	0.99166	No
	D. Woods, ID: 6104003	0.99166	Yes
	Cinema of Belgium, ID: 6088014	0.79234	No
	Edward Michael Law-Yone, ID: 6088845	0.79234	No
	Anna Devane, ID: 6089010	0.79234	No

Query - 6	Top 10 Documents	Score	Is the document
food available at mary macs	Sunkist Fun Fruits, ID: 6095812	0.77501	No
	Ireland–United Kingdom relations, ID: 6087470	0.74612	No
	Neglected tropical diseases, ID: 6090525	0.69387	No
	Shameless (magazine), ID: 6094264	0.68166	No
	Biodegradable plastic, ID: 6104177	0.65525	No
	St. Mark's Episcopal Church (Philadelphia, Pennsylvania), ID: 6113883	0.63625	No
	List of 7th Heaven characters, ID: 6096140	0.63556	No
	Howells (department store), ID: 6088116	0.62552	No
	The People Next Door (TV series), ID: 6089212	0.62552	No
	Abram Trigg, ID: 6089573	0.62552	No

Query - 7	Top 10 Documents	Score	Is the document relevant to the query?
french baguette	Rado (watchmaker), ID: 6106436	0.94278	Yes
	Ursenbach, ID: 6087427	0.33341	No
	The Survivalist (novel series), ID: 6087437	0.33341	No
	Thomas de Foix-Lescun, ID: 6087439	0.33341	No
	Wynau, ID: 6087446	0.33341	No
	Akita International University, ID: 6087570	0.33341	No
	GER Class T26, ID: 6087808	0.33341	No
	Cinema of Belgium, ID: 6088014	0.33341	No
	Francesco II Sforza, ID: 6088068	0.33341	No
	Raphaël Géminiani, ID: 6088261	0.33341	No

Query - 8	Top 10 Documents	Score	Is the document relevant to the query?
role of neural networks in image processing	Pulse-coupled networks, ID: 6107563	0.73060	Yes
	Subjective consciousness, ID: 6095477	0.41926	No
	Neocognitron, ID: 6092601	0.41477	Yes
	K. R. Rao, ID: 6106579	0.31024	Yes
	Electronic filter topology, ID: 6099726	0.29219	No
	Volumetric Imaging and Processing of Integrated Radar, ID: 6101856	0.28466	No
	Escient, ID: 6105318	0.27381	No
	Israel Project, ID: 6104877	0.27340	No
	Prefix sum, ID: 6109308	0.26734	No
	FireViewer, ID: 6100809	0.26481	No

Query - 9	Top 10 Documents	Score	Is the document relevant to the query?
rugby leagues	Frank Stewart, ID: 6108338	0.93454	Yes
	Friends of Labatt Park, ID: 6087668	0.79670	No
	Luke Fitzpatrick, ID: 6088025	0.79670	No
	Richard R. Nacy, ID: 6090984	0.79670	No
	Chris Pelekoudas, ID: 6091308	0.79670	No
	Dmitri Radchenko, ID: 6093211	0.79670	No
	Nori Aoki, ID: 6094497	0.79670	No
	List of SANFL premiers, ID: 6094692	0.79670	No
	Pete Appleton, ID: 6096577	0.79670	No
	Willis Otáñez, ID: 6098332	0.79670	No

Query - 10	Top 10 Documents	Score	Is the document relevant to the query?
drama and theatre	Staatstheater Stuttgart, ID: 6102287	0.74595	Yes
	Diego Jiménez de Enciso, ID: 6113413	0.72703	Yes
	Adrià Gual, ID: 6112314	0.71994	Yes
	Joann Condon, ID: 609798	0.70631	Yes
	Juan Ignacio González del Castillo, ID: 6109318	0.70450	Yes
	Louis Aldrich, ID: 6101096	0.69994	Yes
	Louise Brealey, ID: 6094779	0.69573	Yes
	Jill Hyem, ID: 6097749	0.68302	Yes
	Robert Westenberg, ID: 6099331	0.68072	Yes
	Hochschule für Musik und Theater Hamburg, ID: 6116183	0.65877	Yes

Query - 11	Top 10 Documents	Score	Is the document relevant to the query?
kraft mac and cheese receipe	Mary Mac's Tea Room, ID: 6089190	0.47789	Yes
	United Basketball League, ID: 6115431	0.46992	No
	Energetic Disassembly, ID: 6101847	0.32635	No
	Bilateral key exchange, ID: 6096872	0.30530	No
	Mac O'Grady, ID: 6114186	0.28362	No
	The Gingerbread Man, ID: 6104176	0.28194	No
	Escient, ID: 6105318	0.27310	No
	I'm from Barcelona, ID: 6114398	0.24742	No
	Clinton Foundation, ID: 6115598	0.23963	No
	The News (Mexico City), ID: 6108871	0.22086	No

2) Part-II Results

Improvement I: Lemmatizing query and raw corpus

Ans 1) The IR System built matches the query term exactly to the term present in the document. It doesn't take any morphological analysis or the root word of the term into consideration.

Ans 2) We are proposing to perform lemmatization (a pre-processing) on the query terms and posting list.

Ans 3) The proposed improvement will help when there are very less documents retrieved. Lemmatization will convert the word into its base form and hence will match with more documents. For example, 'Doing' is lemmatized to 'Do' and hence for a query containing 'Doing', the system will output document consisting of 'Doing' and 'Do'.

Ans 4) The terms which are not lemmatized properly will lead to incorrect results such as 'Does' lemmatized to 'Doe' instead of 'Do'.

Ans 5) We demonstrate the impact using these queries as reference:

- 'Abbreviations': 2 Documents retrieved in part 1 versus 7 documents retrieved using lemmatization.
- 'Abnormalities': 3 documents retrieved versus 5 documents retrieved using lemmatized query and document posting list.
- 'Abortions': 3 documents retrieved versus 11 documents retrieved with lemmatization.

Improvement II: Performing a Spell Check on User Query

Ans 1) The IR System built does not work well when the user enters a query term which is misspelled.

Ans 2) We are proposing to use a spelling corrector on the query.

Ans 3) The proposed improvement will help because a wrongly typed term may not be present in the corpus or may retrieve irrelevant documents. For example, typing the term 'muf' may not retrieve any document. But correcting it to 'mum' or 'mud' can help the case.

Ans 4) Sometimes, the names in the query may not be recognized as proper nouns and may be subjected to spelling correction leading to incorrect results. For example, 'Doland Trump' is not corrected to 'Donald Trump' but instead corrected to 'Poland Trump'.

Ans 5) We demonstrate the impact using these queries as reference:

- 'IP Addriss': 4 documents retrieved versus 33 documents retrieved with spell check('IP Address')
- 'Cliv Upton': 6 documents retrieved versus 15 documents retrieved with spell check.
- 'bois nd gurls': 8 documents retrieved versus 70 documents retrieved with spell check.