Information Retrieval: Assignment 2

Group Number 11:
Saurav Virmani 2017A7PS0090P
Sreyas Ravichandran 2017A7PS0275P
Abhay Kanodia 2017A5PS1108P

REPORT

Query 1	Top 10 documents	Score	Is the document relevant?
Syrian brigader Note: spelling mistake In **brigadier.**	History of Syria	1.0	yes
	Kurdish Future Movement in Syria	1.0	yes
	Second Battle of the Shaer gas field	1.0	yes
	Rebecca Sieff Hospital	1.0	yes
	North Lebanon clashes (2014)	1.0	yes
	Mamelukes of the Imperial Guard	1.0	no
	Parichamuttukali	1.0	no
	Zouheir Shourbagi	1.0	yes
	2014 Idlib city raid	1.0	yes
	Palestinian stone-throwin	1.0	yes

Query 2	Top 10 documents	Score	Is the document relevant?
Tycoons from Fiat	Face Off (season 8)	0.9928	yes
	NBMR-3	0.8129	yes
	History of Syria	0.1194	no
	Langeria	0.1194	no
	Katie Ardill	0.1194	no
	Jane Sissmore	0.1194	no
	Jane Doe No. 14 v. Internet Brands, Inc.	0.1194	no
	Effects of Hurricane Floyd in Pennsylvania	0.1194	no
	Port Dunford	0.1194	no
	Charles W. Scharf	0.1194	no

Query 3	Top 10 documents	Score	Is the document relevant?
New Orleans Note: Exists in a lot of documents present in the corpus.	Man Down (film)	1.0318	yes
	Trewartha climate classification	1.0318	yes
	Nefi Ogando	1.0318	yes
	Joe T. Cawthorn	1.0318	yes
	Estevan Hall	1.0318	yes
	Elizabeth Pickett (judge)	1.0318	yes
	Louis Cella	1.0318	yes
	Carlos T. Mock	1.0318	yes

	Bobby Thompson (defensive back)	1.0318	yes
	2015–16 NBA season	1.0318	yes

Query 4	Top 10 documents	Score	Is the document relevant?
HINDI QUERY: **मन की बात**	Mann Ki Baat	1.0401	yes
	Tarwara	0.5453	no

Query 5	Top 10 documents	Score	Is the document relevant?
Conservaton in Pakistan Note: Correct query = **Conservation** in Pakistan	Hala (Pakistan) railway station	1.0531	yes
	Arian Road railway station	0.9987	no
	Attock Khurd railway station	0.9987	no
	Badin railway station	0.9987	no
	Bannu railway station	0.9987	no
	Bhakkar railway station	0.9987	no
	Bhalwal railway station	0.9987	no
	Bhera railway station	0.9987	no
	Bin Qasim railway station	0.9987	no
	Boundry Pillar railway station	0.9987	no

Query 6	Top 10 documents	Score	Is the document relevant?
Sports Representative	French Hill (politician)	0.7872	yes
	Georgy Toloraya	0.7872	yes
	Carlos Curbelo (politician)	0.7872	yes
	Elisha T. Gardner	0.7872	yes
	Steven Nielson	0.7872	yes
	Tebogo Ditshego	0.7872	yes
	Coleridge's theory of life	0.7872	yes
	Martin Barber	0.7872	yes
	International Yoga Day	0.7872	yes
	Dallas Love Field	0.7872	yes

Query 7	Top 10 documents	Score	Is the document relevant?
Microfadeometry	Microfadeometry	1.0	yes

Query 8	Top 10 documents	Score	Is the document relevant?
Australian Individual Speedway Championship	1998 Australian Individual Speedway Championship	0.9726	yes
	1983 Australian Individual Speedway Championship	0.9652	yes
	1995 Australian Individual Speedway Championship	0.9368	yes

	1986 Australian Individual Speedway Championship	0.9368	yes
	1996 Australian Individual Speedway Championship	0.9363	yes
	1997 Australian Individual Speedway Championship	0.9157	yes
	1981 Juniors Track World Championships	0.7687	no
	2014 Ford EcoBoost 400	0.7571	no
	John Hindhaugh	0.6610	no
	2014 Quicken Loans Race for Heroes 500	0.6244	no

Query 9	Top 10 documents	Score	Is the document relevant?
expedition of the United States Army	2014 United States World Cup team	0.5236	no
	United States historical military districts	0.4864	yes
	Marble Valley, Alabama	0.4859	no
	John de Vars Hazard	0.4760	no
	Ziegler, Wisconsin	0.4758	no
	Shhh (film)	0.4587	no
	Yellowstone Expedition of 1873	0.4514	yes
	Westons Mills	0.4512	no
	Sunset Bay	0.4512	no
	Frog Jump, Crockett County, Tennessee	0.4512	no

Query 10	Top 10 documents	Score	Is the document relevant?
President of Ukraine	Anton Kotlyar	0.8078	no
	Pyrausta pavidalis	0.8078	no
	Timeline of the war in Donbass (July–September 2014)	0.6877	yes
	Timeline of the war in Donbass (April–June 2014)	0.6874	yes
	1991 KFK competitions (Ukraine)	0.6488	no
	Carl Peterson (disambiguation)	0.5860	no
	People's Democratic Union &quot;New Ukraine&quot;	0.5733	no
	Party of Slavic Unity of Ukraine	0.5549	yes
	Timeline of the war in Donbass (October–December 2014)	0.5357	yes
	Viktoriya Sasonkina	0.5356	no

ADVANCED

Query 10	Top 10 documents	Score	Is the document relevant?
President of Ukraine	Timeline of the war in Donbass (July–September 2014)	1.5176	yes
	Timeline of the war in Donbass (April–June 2014)	1.4973	yes
	People's Democratic Union &quot;New Ukraine&quot;	1.4032	yes
	Party of Slavic Unity of Ukraine	1.3848	yes
	Timeline of the war in Donbass (October–December 2014)	1.3656	yes
	National Expert Commission of Ukraine on the Protection of Public MoralitY	1.3476	yes
	Zastup (political party)	1.3326	yes
	Kamianka, Skole Raion	1.2619	no
	Party of Veterans of Afghanistan	1.2496	yes
	Taras Vintsiuk	1.2167	yes

Query 5	Top 10 documents	Score	Is the document relevant?
Conservaton in Pakistan Note: Correct query = **Conservation** in Pakistan	Conservation in Pakistan	2.3448	yes
	Beachport Conservation Park	0.8912	no
	Little Dip Conservation Park	0.8912	no
	Douglas Point Conservation Park	0.8912	no
	Hala (Pakistan) railway station	0.7750	no
	Papua grassland mosaic-tailed rat	0.7234	no
	Allahdadani railway station	0.7155	no
	Alluwali railway station	0.7155	no
	Alozai railway station	0.7155	no
	Amirpur Halt railway station	0.7155	no

Query 1	Top 10 documents	Score	Is the document relevant?
Syrian brigader Note: spelling mistake In **brigadier.**	Battle of Buna–Gona	0.7783	yes
	Italian Somali Divisions (101 and 102)	0.7783	yes
	Timeline of the war in Donbass (July–September 2014)	0.7783	yes
	History of Syria	0.6278	yes
	Second Battle of the Shaer gas field	0.6278	yes
	North Lebanon clashes (2014)	0.6278	yes

	Al-Nusra Front–SRF/Hazzm Movement conflict	0.6278	yes
	Ein Qiniyye	0.6278	no
	Suheil Al Hassan	0.6278	no
	Kurdish Future Movement in Syria	0.6278	yes

Assumptions:

1. Non existence of any previous bias for any term.
2. Index creation can be done in memory.
3. Document Ids considered as 1,2,...,N.
4. User submits free text queries.
5. The case folding is not performed while indexing.

Limitations:

1. Documents retrieved do not consider the context of the query.
2. The vector ranking model considers a document as a bag of words.
3. Index generation requires a huge amount of main memory.
4. Previous user queries are not considered while answering future queries. Thus our model is not learning and adapting.

Algorithms:

1. No specific algorithms used for index generation.
2. Just the documents are parsed and tokens are used to create indexes in memory.
3. For ranking tf-idf with cosine normalization is used.

Advance Query:

We have constructed an advanced query module which simultaneously performs 2 upgrades onto our vanilla vector space ranking model.

The upgrades done are:

1. **Spelling Correction:**
    a. The vanilla model made did not work properly when any word was not present in the vocabulary, maybe because someone might have typed a wrong spelling or used the word in plural.

b. Our upgrade uses levenshtein distance to calculate the word in the vocabulary closest to the query word entered. Thus taking care of not only spelling mistakes but also singular/plural and other minute word differences.
c. This module will parse through the string to find any out of vocabulary word, and once found calculating the word nearest to it, and replacing it in the query.
d. Though the module will not consider context while correcting the query and also will not replace the inappropriate word if it exists in the query. For example: If someone types : “male in India” , then the query is considered to be correct, even if the user wanted “Make in India”
e. Improvement example: Query 5 Advance: Conservaton in Pakistan vs Conservation in Pakistan.

2. **Bi-word Indexes:**
    a. The existing model sees a query as a bag of words rather than some phrase which has some contextual meaning. For example John Legend is seen as two separate words only-”John” and “Legend”, thus the results by the system can be irrelevant.
    b. Our upgrade uses Bi-word indexes to take care of these phrasal and collocations implications. Thus another inverted index is created using Bi-words and the ranking scores are updated with the scores because of bi-word matching too.
    c. This module will rank documents using not only the terms present but also by the presence of the bi-words in it. Thus documents whose words appear together as in query are ranked higher.
    d. A corner case might be when the bi-words are common but the multi-words are not. Example “Let it be”
    e. Improvement example: Query 10 Advance: “President of Ukraine”. “President” “of” “Ukraine” vs “President of” , “of Ukraine”