CE706 – SU - Information Retrieval 2022

Assigment 2

Name: Renjith Nataraja Pillai
Student ID : 2111151

Test collection (Task 1)

Information need	Query
Open Jobs/Vacancies related to software engineering in USA: Job position, requirements, employer details, application process, salary/package etc	"Software Engineer Jobs USA"
Latest movie details of Marvel Avengers: Including the filim release date, actors details, etc	"Marvel avengers new movies"
Ongoing Projects/Research undertaken by Oxford University: Including project names, its description, time period, etc	"Oxford University projects"

Table:1

IR systems (Task 2)

After Building the test queries, a second IR system is created by re indexing the collection.

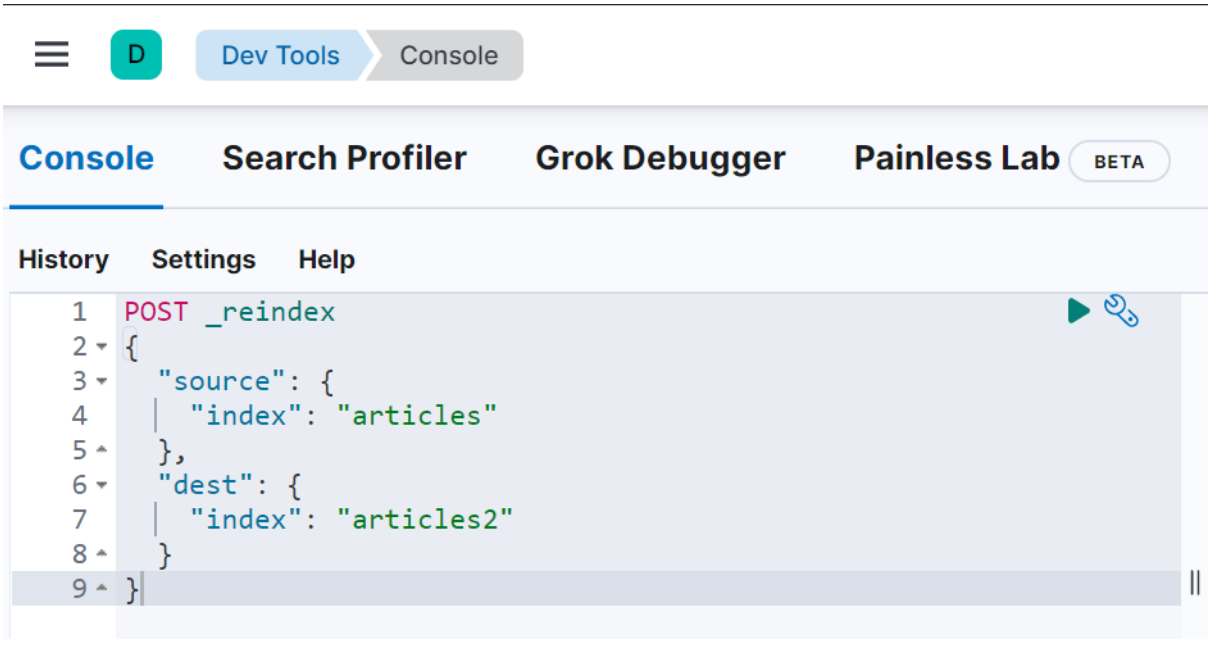

Fig:1 Reindexing

Fig:2: Indices after Re-indexing

Now we have 2 systems with index names, Article and Article2. The Parameters of the system 2 varies from the system 1. Lets display the settings of both system : Article and Article2

Fig:3: System 1: Settings

Fig:4: System 2: Settings

For the system 1 the similarity module used is BM25, which is a TF/IDF based similarity that has built-in tf normalization. It is a probabilistic retrieval framework in which the score is calculated based on TF and IDF components. Since it is based on BoW(Bag of Words) the main disadvantage is that, it does not capture position in text, semantics, co-occurrences in different documents, etc. The second system uses a boolen model for calculating the scores. A simple and basic model in which the scoring is done based on whether a term appears in the document or not. Therefore it has only two scores 0 and 1.

A custom tokenizer is used in the system 1 with token filters asciifolding, lowercase, and word_delimiter. The asciifolding filter will convert the Unicode characters in the query and the lowercase filter convert all letters into lower cases. Word_delimiter filter splits the tokens at special characters. Stop word removal is also used in system 1 which will removes all the words that are not useful while indexing and stop token filter will do the task. Word stemming is also done in system 2 in which the stemming filter kstem will reduce the words in the query into its root form. For system 2 the standard tokenizer is used with token filters lowercase, word_delimiter and apostrophe. Since we are using different pre-processing pipelines for both system for comparison, the stemmimg and Stop word removal is not applied on the second system.

Pool method (Task 3)

The ID of the top 10 documents that are retrieved for each of the queries by each of the systems:

	Query 1: "Software Engineer Jobs USA"	
	DOCUMENT ID	
Rank	**System 1**	**System 2**
1	9159abe7-854b-4527-af6d-9c5bd33b49a7	9159abe7-854b-4527-af6d-9c5bd33b49a7
2	9067506b-9aa8-4b0b-836a-c3d741cda166	09724065-a9a7-483e-9508-3c02cbdc0db8
3	3d0e3300-e003-4bf3-b0b5-ec529701e237	9067506b-9aa8-4b0b-836a-c3d741cda166
4	94293de6-e501-4214-97e3-2577c5b3ba2e	9d59a73a-c3d9-4c48-9631-bc3bf60bdf31
5	f10df5df-3d11-490a-8b18-a22166d49be5	a05c18a7-34cb-4116-9a5a-e3b0b9e0db34
6	b2a28b89-6ecc-4a5d-a2f8-d1e8041c41a7	d9fde138-c1dc-43ab-b840-4f17d62c7c89
7	033d40ab-1b84-4ce6-a55f-1c9f2beb3bfa	3d0e3300-e003-4bf3-b0b5-ec529701e237
8	518dca9e-4c9c-4328-a4aa-0e3bc202b1d0	c66947b7-9a2b-42bf-8667-8a6aaa95d845
9	39fd20d0-3e87-4bbf-87a8-3e160c9780f5	5e876db9-233a-4eb5-9b33-78d9f54d20cb
10	99e0ad5f-2998-4103-b645-98b213a32680	840e94f3-b81b-45ee-8a5e-380b5410c65e
# different documents	**17**	

Table:2

	Query 2: "Marvel avengers new movies"	
	DOCUMENT ID	
Rank	**System 1**	**System 2**
1	a4471538-6a5d-467a-af5a-66481f4ff7a6	a4471538-6a5d-467a-af5a-66481f4ff7a6
2	1f3fb696-1ea9-4c5b-8378-352678fa8755	1f3fb696-1ea9-4c5b-8378-352678fa8755
3	63b5ff06-c729-4edb-9912-272f35eddc05	a816d4df-7fea-4c3c-8dac-6c4d077675eb
4	c24f70f2-97b0-4cad-8928-0964c8e14604	26132e6b-748a-4654-ba72-35bcfeda9cbc
5	ad793d45-ac7c-4c19-b1bb-13a1d9604f37	a10c3827-0f7c-4d5f-965c-bd97ba3e512f
6	a816d4df-7fea-4c3c-8dac-6c4d077675eb	6e44253e-9856-41da-a534-2ee20eebe259
7	e7fd4833-0eb7-46a8-9b44-f5e16203c73f	84bc64e7-d69c-4970-a4ae-67bf5e79d432
8	399c614d-e0a0-4c8c-854b-f796eddd426b	b8fc24d1-66ac-4bb5-99c8-77f5d7254704
9	c11f6d41-c195-4464-ac80-d4d8676bb5c0	e978be72-d9d3-49db-b24b-01320af96e01
10	035d61a7-c487-48c4-a27a-c197fa04b354	3be00b73-4c66-465d-8be1-fdb2f7aeb59b
# different documents	**17**	

Table:3

	Query 3: " Oxford University projects "	
	DOCUMENT ID	
Rank	**System 1**	**System 2**
1	5af61a0c-8c95-4bdb-8abc-7d4b96c7ae98	a754e2ec-7254-4492-a706-608391c221b2
2	b815fe64-e679-454e-be7e-7571d86f7e58	982a86c9-7902-4f76-ab27-075552f4b674
3	63b47992-9906-4f28-88b3-64c74232f7bc	94dfbce1-e574-4d83-b496-eeb8c3d6b8ea
4	94dfbce1-e574-4d83-b496-eeb8c3d6b8ea	64b55d4b-5529-45b1-97d8-aea0a0b43e4c
5	01940125-de0c-4006-9379-b632a5fdc59d	b815fe64-e679-454e-be7e-7571d86f7e58
6	9e78ef85-c2ef-44ca-93c9-e74764382b42	8dcd280e-5b07-4525-890d-87b1b8c2e9f3
7	eace0b48-bfd0-4732-9274-ad5d56f04ae7	5617be39-2ce6-4d18-9115-c4b915f9c9ac
8	982a86c9-7902-4f76-ab27-075552f4b674	01940125-de0c-4006-9379-b632a5fdc59d

9	64b55d4b-5529-45b1-97d8-aea0a0b43e4c	63b47992-9906-4f28-88b3-64c74232f7bc
10	0a56456e-68dd-4892-a720-8cfe811fcf22	0a56456e-68dd-4892-a720-8cfe811fcf22
# different documents	**13**	

Table:4

Relevance assessments (Task 4)

Relevance criteria:

Assessing the relevance of the search result is very important in an IR system. There are may factors that affect the relevance of the reterived document. Some of them are topicality, novelty, freshness, authority,formatting, reading level etc.

Topicality : Whether the reterived documents is based on the same topic of the search query.

Freshnes : Whether the information reterived is upto date. For example if we are searching for news, information may change from time to time.

Authority : Whether the document reterived is from reliable resourses. In these days there are a lot of fake information spreading over the web. It is very important to check the authority of the document source.

Formatting/ Readability: Even if the reterived document have the required information, there is no use if it is not described well. The readability and Formatting is very essential in transferring the information.

Along with these general factors there are user specific criterion that determines the relevance of the document. This will change for each query and its purely depents upon individual preferences.
In this assignment we are using 3 queries, the personal criteria for each query is listed on the table below:

User Specific Criterion:

	Query	User Specific Criteria
Query 1	"Software Engineer Jobs USA"	Location specific: Must be from USA
		A brief description about job, including the position, job responsibilities, requirements, salary etc
		Should mention how to apply for the job
Query 2	"Marvel avengers new movies"	Including the filim release date, actors details, etc
		News related to the movies, new movie announcements etc
Query 3	"Oxford University projects"	Including project names, its description, time period, etc
		Supervisor details, Area of study
		Previous studies related to the projects etc

Table:5

Based on the above criteria we have sorted out the relevant results for each query and the list is given in the table below:

ID of the relevant documents

Query		ID of relevant documents
Query 1	1	9159abe7-854b-4527-af6d-9c5bd33b49a7
	2	9067506b-9aa8-4b0b-836a-c3d741cda166
	3	3d0e3300-e003-4bf3-b0b5-ec529701e237
	4	99e0ad5f-2998-4103-b645-98b213a32680
	5	09724065-a9a7-483e-9508-3c02cbdc0db8
	6	9d59a73a-c3d9-4c48-9631-bc3bf60bdf31
	7	5e876db9-233a-4eb5-9b33-78d9f54d20cb
Query 2	1	a4471538-6a5d-467a-af5a-66481f4ff7a6
	2	1f3fb696-1ea9-4c5b-8378-352678fa8755
	3	a816d4df-7fea-4c3c-8dac-6c4d077675eb
	4	e7fd4833-0eb7-46a8-9b44-f5e16203c73f
	5	26132e6b-748a-4654-ba72-35bcfeda9cbc
	6	a10c3827-0f7c-4d5f-965c-bd97ba3e512f
	7	84bc64e7-d69c-4970-a4ae-67bf5e79d432
	8	c11f6d41-c195-4464-ac80-d4d8676bb5c0
	9	035d61a7-c487-48c4-a27a-c197fa04b354
	10	e978be72-d9d3-49db-b24b-01320af96e01
Query 3	1	982a86c9-7902-4f76-ab27-075552f4b674
	2	64b55d4b-5529-45b1-97d8-aea0a0b43e4c
	3	0a56456e-68dd-4892-a720-8cfe811fcf22
	4	94dfbce1-e574-4d83-b496-eeb8c3d6b8ea

Table:6

The irrelevant documents against each query and reason for rejection is explained in the table below:

Query	ID	Reason
Query 1	033d40ab-1b84-4ce6-a55f-1c9f2beb3bfa	Not related to the search topic
	39fd20d0-3e87-4bbf-87a8-3e160c9780f5	Not related to the search topic
	518dca9e-4c9c-4328-a4aa-0e3bc202b1d0	Not related to the search topic
	840e94f3-b81b-45ee-8a5e-380b5410c65e	Dont have enough Information
	94293de6-e501-4214-97e3-2577c5b3ba2e	Not related to the search topic
	a05c18a7-34cb-4116-9a5a-e3b0b9e0db34	Not related to the search topic
	b2a28b89-6ecc-4a5d-a2f8-d1e8041c41a7	Not related to the search topic
	c66947b7-9a2b-42bf-8667-8a6aaa95d845	Not related to the search topic
	d9fde138-c1dc-43ab-b840-4f17d62c7c89	Dont have enough Information
	f10df5df-3d11-490a-8b18-a22166d49be5	Not related to the search topic
Query 2	63b5ff06-c729-4edb-9912-272f35eddc05	Not related to the search topic
	c24f70f2-97b0-4cad-8928-0964c8e14604	Dont have enough Information
	ad793d45-ac7c-4c19-b1bb-13a1d9604f37	Dont have enough Information
	399c614d-e0a0-4c8c-854b-f796eddd426b	Bad Formatting
	a816d4df-7fea-4c3c-8dac-6c4d077675eb	Not from an authorised source
	6e44253e-9856-41da-a534-2ee20eebe259	Not related to the search topic
	3be00b73-4c66-465d-8be1-fdb2f7aeb59b	Bad Formatting
Query 3	5af61a0c-8c95-4bdb-8abc-7d4b96c7ae98	Not related to the search topic
	b815fe64-e679-454e-be7e-7571d86f7e58	Not related to the search topic
	63b47992-9906-4f28-88b3-64c74232f7bc	Not related to the search topic
	94dfbce1-e574-4d83-b496-eeb8c3d6b8ea	Not related to the search topic
	01940125-de0c-4006-9379-b632a5fdc59d	Not related to the search topic
	9e78ef85-c2ef-44ca-93c9-e74764382b42	Dont have enough Information
	eace0b48-bfd0-4732-9274-ad5d56f04ae7	Dont have enough Information
	5617be39-2ce6-4d18-9115-c4b915f9c9ac	Not related to the search topic
	8dcd280e-5b07-4525-890d-87b1b8c2e9f3	Not related to the search topic

Table:7

Evaluation (Task 5)

Evaluation is the backbone of building an efficient search engine. There are different metrics used to evaluate the model. Here we are using precision and recall as an evaluation metrics:

Precision indicate what proportion of the documents returned are relevant, the equation for calculating the precision is :

Precision = (No. of relevant docs returned) / (No. of docs returned)

Recall indicates the proportion of relevant documents that are retrieved

Recall = (No. of relevant docs returned) / (Total No. of relevant docs)

By using these formula we have to calculate the value of precision and recall for each query, with K =5. Lets consider the query 1 and system 1:

Query 1 and System 1				
K	**DOCUMENT ID**	**P@K**	**R@K**	
1	9159abe7-854b-4527-af6d-9c5bd33b49a7	**(1/1) = 1.0**	**(1/7) = 0.14**	**Total No. of relevant docs = 7**
2	9067506b-9aa8-4b0b-836a-c3d741cda166	**(2/2)= 1.0**	**(2/7) = 0.29**	
3	3d0e3300-e003-4bf3-b0b5-ec529701e237	**(3/3)= 1.0**	**(3/7) = 0.429**	
4	94293de6-e501-4214-97e3-2577c5b3ba2e	**(3/4)= 0.75**	**(3/7) = 0.429**	
5	f10df5df-3d11-490a-8b18-a22166d49be5	**(3/5)= 0.6**	**(3/7) = 0.429**	
…..	…….	…….	……..	

	RELEVANT DOCUMENTS

Table:8

From the above example the no of relevant documents when K@5 is 3 and the precision is calculated as the ratio of No of relevant documents to the No of documents returned. In this case its (3/5 = 0.6). For calculating the recall we need to identify the total no of relevant documents reterived from both system for each query which we already identified. For query 1 we have total 7 relevant documents. So recall@5 is (3/7 = 0.429). Similarly the paremeter values of each query in both system is shown in the table below:

	Q1	**Q2**	**Q3**
No. of relevant docs returned: System 1	**3**	**2**	**2**
No. of relevant docs returned: System 2	**3**	**4**	**3**
No. of docs returned (**K = 5)**	**5**	**5**	**5**
Total No. of relevant docs	**7**	**10**	**4**

Table:9

Now we have all the values. Calculating the precision and recall for all the 3 queries in two system by using the above equations:

Final Table:

	System 1		System 2	
	P@5	**R@5**	**P@5**	**R@5**
Q1	0.6	0.429	0.6	0.429
Q2	0.4	0.20	0.8	0.4
Q3	0.4	0.28	0.6	0.75

Table:10

Web search (Task 6)

By comparing the evaluation metrics of both systems, the system 1 have an average precision of 0.46 and a recall of 0.303. Similarly for system 2, the average values for both precision and recall are 0.66 and 0.52 respectivily. It is clear that the system 2 performed well in the evaluation process. Even though we didn't used stemming and stop word removal in system 2 the boolen model worked well in simple text query. But if we used complex queries the system 1 may out perform the other. From analysing the evaluation metrics only, we cannot say that the system 1 is better than system 2. Now consider the ranking and relevance. During the retrieval phase most of the top ranked documents were relevant in system 2 compared to system 1. For example consider the query 3:

Query: 3		
Rank	**System 1**	**System 2**
1	5af61a0c-8c95-4bdb-8abc-7d4b96c7ae98	a754e2ec-7254-4492-a706-608391c221b2
2	b815fe64-e679-454e-be7e-7571d86f7e58	982a86c9-7902-4f76-ab27-075552f4b674
3	63b47992-9906-4f28-88b3-64c74232f7bc	94dfbce1-e574-4d83-b496-eeb8c3d6b8ea
4	94dfbce1-e574-4d83-b496-eeb8c3d6b8ea	64b55d4b-5529-45b1-97d8-aea0a0b43e4c
5	01940125-de0c-4006-9379-b632a5fdc59d	b815fe64-e679-454e-be7e-7571d86f7e58
6	9e78ef85-c2ef-44ca-93c9-e74764382b42	8dcd280e-5b07-4525-890d-87b1b8c2e9f3
7	eace0b48-bfd0-4732-9274-ad5d56f04ae7	5617be39-2ce6-4d18-9115-c4b915f9c9ac
8	982a86c9-7902-4f76-ab27-075552f4b674	01940125-de0c-4006-9379-b632a5fdc59d

9	64b55d4b-5529-45b1-97d8-aea0a0b43e4c	63b47992-9906-4f28-88b3-64c74232f7bc
10	0a56456e-68dd-4892-a720-8cfe811fcf22	0a56456e-68dd-4892-a720-8cfe811fcf22

	RELEVANT DOCUMENTS

Table:11

Here the first realevant document for the search query is in rank 8 for system 1, where as in the second system the documents in the rank 2, 3 and 4 were related to the search topic. While choosing a system for web search, higher rank of the first relevant document is always better. It will satisfy the user. The user may doesn't need all the relevant documents.

Also, the proportion of relevant documents that are retrieved is less in system 1, that's why it has low recall value. Both systems gave conssistant results in all runs. It is notable that even though most of the reterived documents for the first 2 queries were different in both systems the relevant documents were almost same. Since in the normal web search, most users uses simple text queries and words, the system 2 will be more efficient in those scenarios as per the evaluation results.