CE706 – SU - Information Retrieval 2023

Assignment 2

Reg No: 2211433

Test collection (Task 1)

Below table contains information needed for the three identified queries:

Information need	Query
Men of Herefordshire engaged in the fight.	Men engage in fight.
They met at Eastcheap the next night	Meet at Eastcheap
His brother Lord Scroop's death took place at Bristol	Death at Bristol

Ground Truth Examples:

Example 1:

Query 1: Men engage in fight.

Expected Results: For Query 1 the following line_id and text_entry should be retrieved.

1. {"line_id":42, "text_entry": "Leading the men of Herefordshire to fight"}
2. {"line_id":3010, "text_entry": "They fight. KING HENRY being in danger, PRINCE HENRY enters"}
3. {"line_id":3200, "text_entry": "To fight with Glendower and the Earl of March."}
4. {"line_id":24,"text_entry":"We are impressed and engaged to fight,"}

Example 2:

Query 2: Meeting at Eastcheap

Expected Results: For Query 2 the following line_id and text_entry should be retrieved.

1. {"line_id":238: "text_entry": "supper tomorrow night in Eastcheap: we may do it"}
2. {"line_id":263, "text_entry": "count enance. Farewell: you shall find me in Eastcheap."}

Example 3:

Query 3: Death at Bristol

Expected Results: For Query 3 the following line_id and text_entry should be retrieved.

1. {"line_id":612,", "text_entry": "His brothers death at Bristol, the Lord Scroop."}

2. {"line_id":7784, "text_entry": "At Bristol I expect my soldiers;"}

IR systems (Task 2)

I made use of assignment 1 as IR system 1 which uses pre-processing filters such as, bigram, lowercase conversion, removing stop-words, stemming whereas for IR system 2 I created a new index which with pre-processing steps of converting the letters to lowercase and removing stop-words. Mapping each of the documents type along with the custom analyser of both systems and fetching top 3000 lines of both our created Information Retrieval System. Prediction that system 1 will perform better when compared to system 2 as in system 1 there are certain pre-processing steps that will enhance the scores of evaluations.

Pool method (Task 3)

For each query (3 queries) maximum of 5 documents were retrieved from both our IR system. The retrieved documents ids have been mentioned in the below table.

	Query 1		Query 2		Query 3	
Rank	**System 1**	**System 2**	**System 1**	**System 2**	**System 1**	**System 2**
1	359	42	536	297	207	612
2	24	24	297	709	612	761
3	42	920	639	890	1073	661
4	810	359	33	15	776	486
5	920	1253	323	238	1369	476
# different documents	2		6		8	

Judge every document in the pooling.

Query 1 for System 1 and 2:

Rank 2 and 4 in System 1 with Doc IDs 24 and 42 and Rank 1 and 4 in System 2 Doc IDs 42 and 24 are relevant as it answers the query 1. Men of Herefordshire engaged in fight along with other soldiers was the information I was looking for. Even though the other common Doc IDs 359, 920 from both systems resulted with the keywords it still was irrelevant to the query. The Doc ID 810 from system 1, resulted as “Some eight or ten.” Which has zero relevance to the query and Doc ID 253 from system 2 resulted as “as thou hast done, and then say it was in fight!”. Used keyword “fight” and resulted the document’s text entry.

Query 2 for System 1 and 2:

Rank 2 in system 1 and Rank 2 in system 2 with Doc ID 297and Doc ID 238 from System 2 is relevant to the query provided. The information of the query was about the meeting that held at Eastcheap the next night or following night for supper.

From system 1 the Doc IDs 639 and 33 used keyword “meet” and resulted the texts with the following IDs whereas the Doc ID 536 resulted as “By heaven, methinks it were an easy leap” and Doc ID 323 resulted as “Redeeming time when men think least I will” had no relevant keywords to match the query.

In system 2 the Doc ID 238 is relevant but the same was not displayed in system 1. The other Doc IDs 709, 890, 15 again using the keyword “meet” retrieved the irrelevant sentences.

Query 3 for System 1 and 2:

The Doc ID 612 at Rank 2 in system 1 and at Rank 1 in System 2 is relevant as it addresses the query with the actual information needed i.e, about the death of his brother Lord Scroop which took place at Bristol. The rest of the Doc IDs 207, 1073, 776, and 1369 in system 1 retrieved sentences but had no match texts or keywords with the query whereas in System 2 Doc IDs 761, 661, 486, 476 had either “die” or “death” which was used as a keyword to retrieve the sentences for the given query 3.

Relevance assessments (Task 4)

Relevance criteria: Given some details that can be used to answer the set of questions that will be provided by the information needs. Does not necessarily need to have answers but some relevant information based on the question.

Query 1: Men engage in fight.

Description: Details of the fight’s location, information on how many men got injured in the fight, reason why the fight happened, who won the fight, did someone die in the fight, weapons used in the fight.

Query 2: Meet at Eastcheap

Description: Location of the meeting, at what time did they meet, what was the discussion when they met, meeting between how many people, number of invites to the meeting, did they meet for dinner or supper or for lunch, did they still discuss about the fight when they met again, was the meeting about upcoming war, was it just a family reunion meeting.

Query 2: Death at Bristol.

Description: Who died at Bristol, what caused him/her to death, place of death, whether the death was of a man or woman, if man then whose brother/son/father was he, was he/she murdered, was she/he ruler of Scotland or was she/he the ruler of York, if so then who was crowned after his/her death, how many siblings did he/she have in total , was the person married and had children.

Below table consists of complete description of the query's relevance criteria along with the IDs of relevant documents in both IR System 1 and IR System 2:

	Description	ID of relevant documents
Query 1	Details of the fight's location, information on how many men got injured in the fight, reason why the fight happened, who won the fight, did someone die in the fight, weapons used in the fight.	42 in both system 24 in both system
Query 2	Location of the meeting, at what time did they meet, what was the discussion when they met, meeting between how many people, number of invites to the meeting, did they meet for dinner or supper or for lunch, did they still discuss about the fight when they met again, was the meeting about upcoming war, was it just a family reunion meeting,	297 in both system 238 in **System 2** alone
Query 3	Who died at Bristol, what caused him/her to death, place of death, whether the death was of a man or woman, if man then whose brother/son/father was he, was he/she murdered, was she/he ruler of Scotland or was she/he the ruler of York, if so then who was crowned after his/her death, how many siblings did he/she have in total , was the person married and had children	612 in both systems

Evaluation (Task 5)

To calculate the precision, a function was created for a retrieval system by comparing the relevant and retrieved documents, considering only the top 3 documents. It computes the proportion of relevant documents among the retrieved ones, providing a measure of the system's accuracy in returning relevant results. The precision metric measures the proportion of relevant documents among the retrieved ones, giving an indication of how accurate and focused the retrieval system is.

Recall measures the proportion of relevant documents that were successfully retrieved out of all the relevant documents. A function to calculate the recall of a retrieval system was created by comparing relevant and retrieved documents, considering only the top 3 documents. It computes the proportion of relevant documents successfully retrieved out of all the relevant documents, giving a measure of the system's ability to retrieve all relevant results.

```python
def precision(relevant_docs, retrieved_docs):
    relevant = set(relevant_docs[:3])  # Consider only the first 3 relevant documents
    retrieved = set(retrieved_docs[:3])  # Consider only the first 3 retrieved documents
    intersection = relevant.intersection(retrieved)  # Get the common documents between relevant and retrieved sets
    precision1 = len(intersection) / 3  # Calculate precision as the ratio of common documents to 3 (considered documents)
    return precision1


def recall(relevant_docs, retrieved_docs):
    relevant= set(relevant_docs[:3])  # Consider only the first 3 relevant documents
    retrieved = set(retrieved_docs[:3])  # Consider only the first 3 retrieved documents
    intersection = relevant.intersection(retrieved)  # Get the common documents between relevant and retrieved sets
    if len(relevant) == 0:
        recall1 = 0
    else:
        recall1 = len(intersection) / len(relevant)  # Calculate recall as the ratio of common documents to relevant documents
    return recall1
```

✓ 0.0s

```python
relevance_assessment = {
    "q1": {
        "q1System1" : ["42", '24'],
        "q1System2" : ["42", '24']
    },
    "q2": {
        "q2System1" : ["297"],
        "q2System2" : ["297", "238"]
    },
    "q3":{
        "q3System1" : ["612"],
        "q3System2" : ["612"]
    }
}
```

[109] ✓ 0.0s

```python
# Calculate P@3 and R@3 for each query and systen
p3 = {}
r3 = {}

for query in relevance_assessment:
    p3[query] = {}
    r3[query] = {}

    for system in relevance_assessment[query]:
        relevant_docs = relevance_assessment[query] [system]
        retrieved_documentss = relevance_assessment[query][system]
        precision2 = precision(relevant_docs, retrieved_documentss)
        recall2 = recall(relevant_docs, retrieved_documentss)
        p3[query][system] = precision2
        r3[query][system] = recall2
```

[110] ✓ 0.0s

	System 1		System 2	
	P@3	**R@3**	**P@3**	**R@3**
Q1	0.67	1.0	0.67	1.0
Q2	0.3	1.0	0.67	1.0
Q3	0.3	1.0	0.3	1.0

Plagiarism detection (Task 6)

Discussion: In the initial discussion of the two systems, my prediction was that system 1 would perform much better than system 2. However, the above results shows that the Precision and Recall for Queries 1 and 2 are same and that is because we had exact one match for queries 1 and 2 in both IR system but when it comes to Query 1 Precision score for both system is different on comparison with the other two queries. The reason being that for query 1 we had exact match of two documents leading to lower scores of precisions. As we have only 5 documents to be sampled with and to perform the evaluation with k=3 we got almost similar scores. Assumptions if k=20 and retrieved relevant document is 6 then the scores will change accordingly.

To retrieve documents that may be the sources of plagiarism for a suspicious document I would choose System 1 with Recall and Precision metrics. This result indicates that the model neither predicts false negatives nor false positive. However, this might change when more than 5 documents are retrieved. As in system 1 we used different pre-processing methods, by removing stop-words, stemming, conversion of letter to lowercase. Bigrams and this will also help in plagiarism detection for suspicious documents with better results on Recall and Precision as well.