CE306 or CE706 - Information Retrieval 2022

Assignment 2

1904341

Assessing relevance (Task 1)

Doc	Judgment	Reason
1.	Relevant	Document states that parking is free during Graduation week.
2.	Not Relevant	Document lists all University of Essex pages.
3.	Not Relevant	Document contains information regarding honorary graduates from the University of Essex.
4.	Not Relevant	Document describes the ways in which the University of Essex webpage is made accessible to as many people as possible.
5.	Relevant	Document gives information regarding disabled parking during graduation.
6.	Not Relevant	Document describes the 2008 Alumnus of the Year award winner Dotun Adebayo.
7.	Not Relevant	Document gives information regarding parking, but the information is not related to graduation.
8.	Not Relevant	Documents lists ways to get to the University of Essex.
9.	Not Relevant	Document details how the University of Essex takes pride in its way of teaching and how it is done.
10.	Not Relevant	Document explains how to get to the University of Essex and where to park, but it is not related to graduation.

Pooling (Task 2)

To deem which of the documents are relevant enough, a pool is constructed by combining the top 10 retrieval results from all three IR systems. Any documents not inside this pool is automatically considered not relevant.

The four documents not in the pool are 5, 10, 11 and 12.

This leaves the following documents in the pool: 1, 2, 3, 4, 6, 7, 8, 9, 13, 14, 15, 16, 17, 18, 19, 20.

Using the provided program "*f(x) = x%2*" to asses relevance, the following table was constructed.

Doc	**Judgment**	**Doc**	**Judgment**
1.	Relevant	**11.**	Not Relevant
2.	Not Relevant	**12.**	Not Relevant
3.	Relevant	**13.**	Relevant
4.	Not Relevant	**14.**	Not Relevant
5.	Not Relevant	**15.**	Relevant
6.	Not Relevant	**16.**	Not Relevant
7.	Relevant	**17.**	Relevant
8.	Not Relevant	**18.**	Not Relevant
9.	Relevant	**19.**	Relevant
10.	Not Relevant	**20.**	Not Relevant

P/R@5 (Task 3)

To find P@5 and R@5 for each of the three IR systems, the previous 4 precisions and recalls need to calculated.
The values of precision and recall at most positions is also needed for later parts of the report.
To not disrupt the layout of the report, the complete tables for all three IR systems have been included at the end of the report in the section "Precision-Recall Tables".

The following table was constructed with values from the tables in the mentioned section.

Systems	**P@5**	**R@5**
IR1:	(4/5) = 0.8	(4/8) = 0.5
IR2:	(2/5) = 0.4	(2/8) = 0.25
IR3:	(2/5) = 0.4	(2/8) = 0.25

Average Precision (Task 4)

To calculate the average precision, the following must be done.
The rankings of each IR system must be searched through one rank at a time. Then if the document is relevant, add the P@K to a total amount starting at 0. When R@K is equal to 1.0, average the sum. The resulting being the average precision.

This is done for each table in “Precision-Recall Tables”, where the relevant documents for each IR system has been labelled with a the letter R.

Systems	**Average Precision**
IR1:	*1.0 + 1.0 + 1.0 + 1.0 + 0.833 + 0.666 + 0.7 + 0.471 = 6.67* *6.67/8 = 0.834* ***Average Precision of IR1 is 0.834***
IR2:	*0.333 + 0.5 + 0.5 + 0.571 + 0.417 + 0.462 + 0.412 + 0.421 = 3.616* *3.616/8 = 0.452* ***Average Precision of IR2 is 0.452***
IR3:	*0.5 + 0.5 + 0.5 + 0.5 + 0.385 +0.4 + 0.438 + 0.471 = 3.694* *3.694/8 = 0.462* ***Average Precision of IR3 is 0.462***

Discounted Cumulative Gain (Task 5)

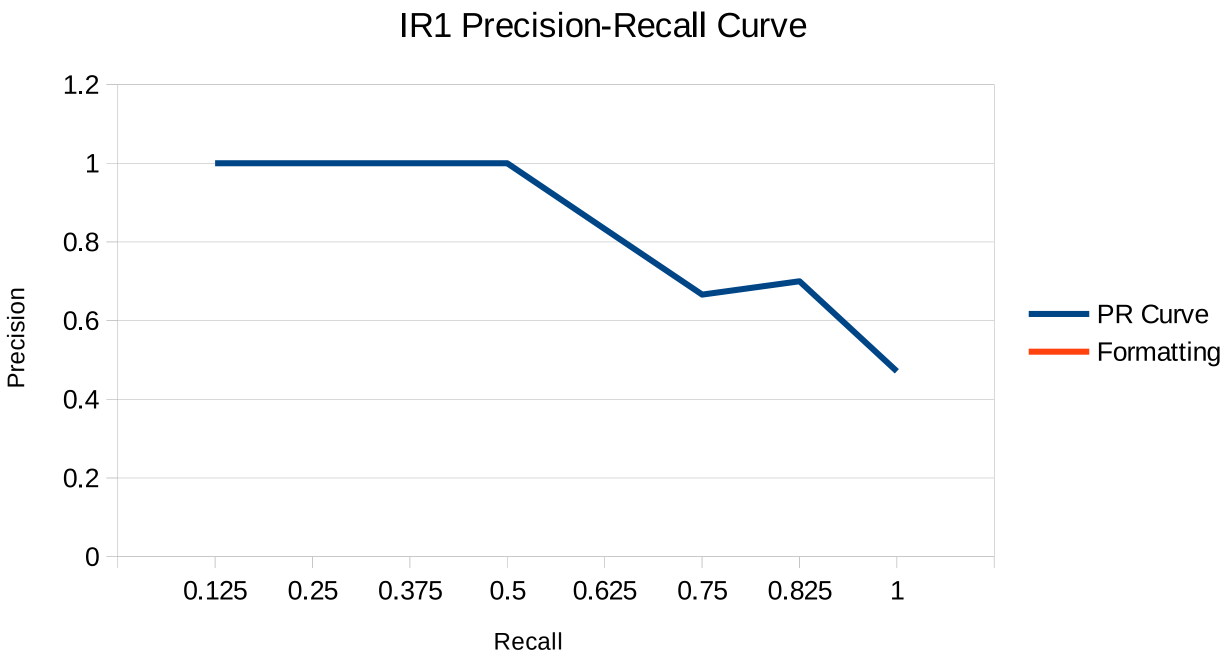
To calculate the discounted cumulative gain, the following formula must be applied to the three IR systems.

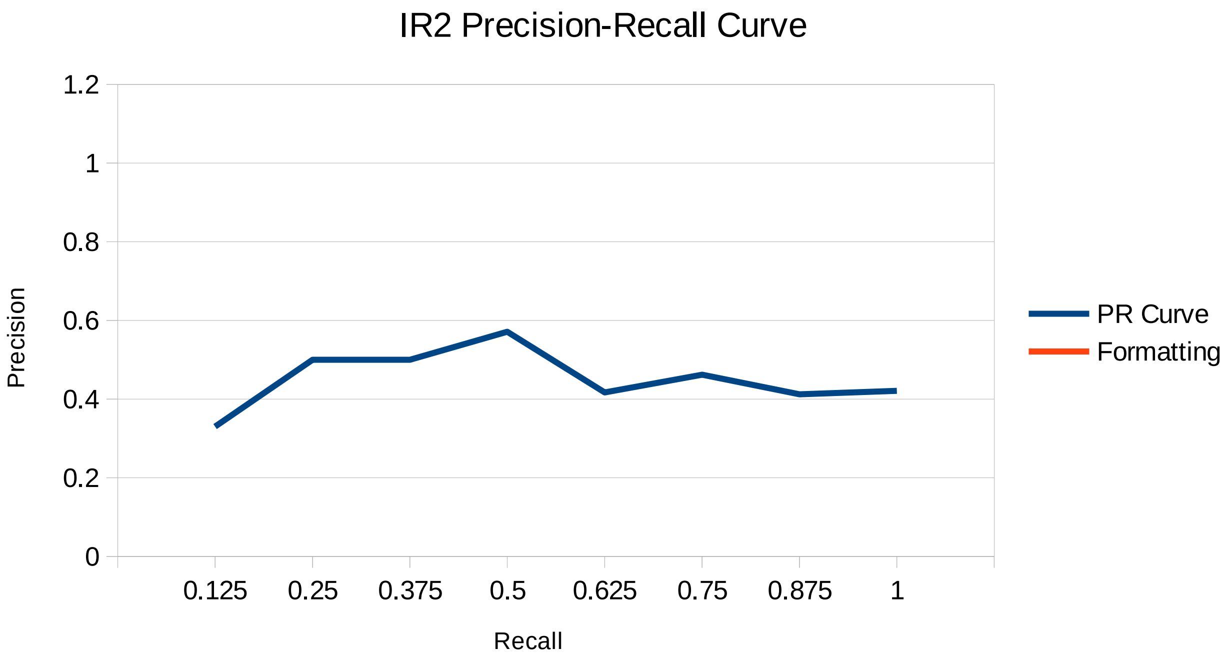
$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

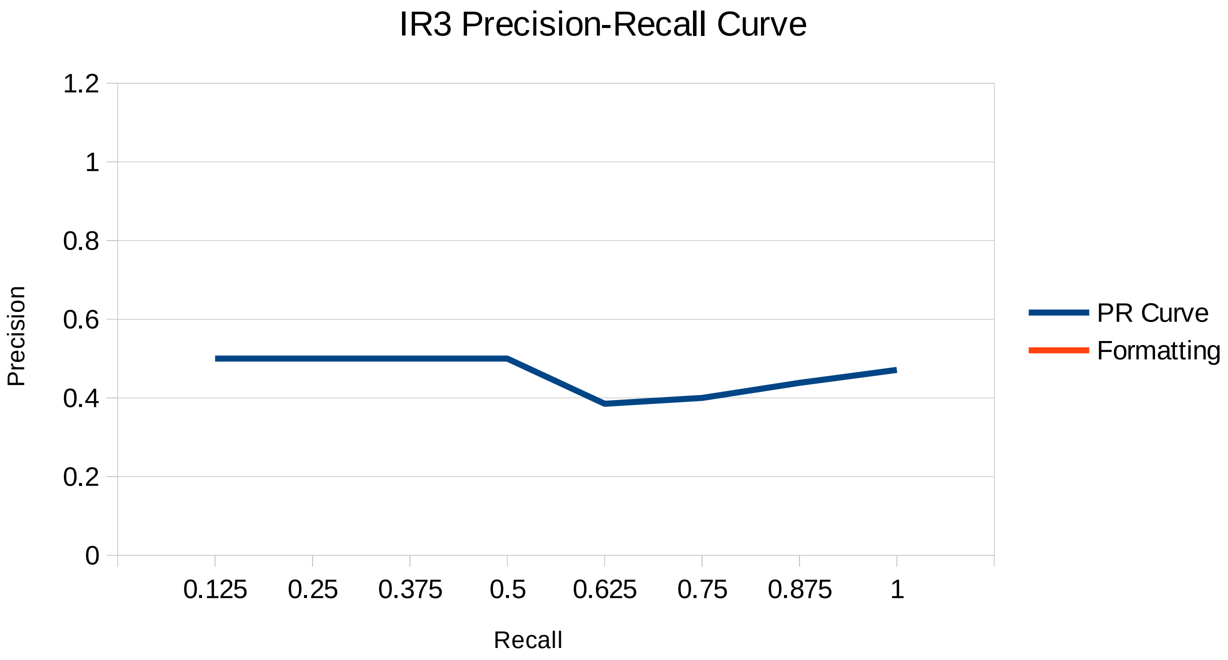
This formula is applied to every relevant document for each of the three IR systems, and the relevancy is being treated as being 1 for each relevant document. The *i* is the position of the relevant document, and is taken from the tables in “Precision-Recall Tables”

Systems	**Discounted Cumulative Gain**
IR1:	$1 + 1/log_2 2 + 1/log_2 3 + 1/log_2 4 + 1/log_2 6 + 1/log_2 9 + 1/log_2 10 + 1/log_2 17 =$ *1 + 1 + 0.631 + 0.5 + 0.387 + 0.315 + 0.301 + 0.245 = 4.379* ***Discounted Cumulative Gain of IR1 is 4.379***
IR2:	$0 + 1/log_2 3 + 1/log_2 4 + 1/log_2 6 + 1/log_2 7 + 1/log_2 12 + 1/log_2 13 + 1/log_2 17 + 1/log_2 19 =$ *0 + 0.631 + 0.5 + 0.387 + 0.356 + 0.279 + 0.270 + 0.245 + 0.235 = 2.903* ***Discounted Cumulative Gain of IR2 is 2.903***
IR3:	$0 + 1/log_2 2 + 1/log_2 4 + 1/log_2 6 + 1/log_2 8 + 1/log_2 13 + 1/log_2 15 + 1/log_2 16 + 1/log_2 17 =$ *0 + 1 + 0.5 + 0.387 + 0.333 + 0.270 + 0.256 + 0.25 + 0.245 = 3.241* ***Discounted Cumulative Gain of IR3 is 3.241***

Precision-Recall Curves (Task 6)







For the scholar search that requires 80% of recall, IR3 will be chosen as it has the lowest variance in retrieval result of the three systems.

Web search (Task 7)

The metric that will be used to determine which IR system is to be used for web search will be discounted cumulative gain, because with web search the top results are the more important ones as people often do not navigate to the next page or even scroll, they only look at the first couple of results.
The system with the highest discounted cumulative gain is IR1, so it is the one that will be chosen as the information retrieval system for web search.

Precision-Recall Tables

All numbers have been shortened to at most three decimals. Using the table with relevant and not relevant documents after pooling, these are the P/R@K tables for the three IR systems.

IR1

K	**P@K**	**R@K**
1 = R	(1/1) = 1.0	(1/8) = 0.125
2 = R	(2/2) = 1.0	(2/8) = 0.25
3 = R	(3/3) = 1.0	(3/8) = 0.375
4 = R	(4/4) = 1.0	(4/8) = 0.5
5	(4/5) = 0.8	(4/8) = 0.5
6 = R	(5/6) = 0.833	(5/8) = 0.625
7	(5/7) = 0.714	(5/8) = 0.625
8	(5/8) = 0.625	(5/8) = 0.625
9 = R	(6/9) = 0.666	(6/8) = 0.75
10 = R	(7/10) = 0.7	(7/8) = 0.875
11	(7/11) = 0.636	(7/8) = 0.875
12	(7/12) = 0.583	(7/8) = 0.875
13	(7/13) = 0.538	(7/8) = 0.875
14	(7/14) = 0.5	(7/8) = 0.875
15	(7/15) = 0.466	(7/8) = 0.875
16	(7/16) = 0.438	(7/8) = 0.875
17 = R	(8/17) = 0.471	(8/8) = 1.0
18	(8/18) = 0.444	(8/8) = 1.0
19	(8/19) = 0.421	(8/8) = 1.0
20	(8/20) = 0.4	(8/8) = 1.0

IR2

K	P@K	R@K
1	(0/1) = 0.0	(0/8) = 0.0
2	(0/2) = 0.0	(0/8) = 0.0
3 = R	(1/3) = 0.333	(1/8) = 0.125
4 = R	(2/4) = 0.5	(2/8) = 0.25
5	(2/5) = 0.4	(2/8) = 0.25
6 = R	(3/6) = 0.5	(3/8) = 0.375
7 = R	(4/7) = 0.571	(4/8) = 0.5
8	(4/8) = 0.5	(4/8) = 0.5
9	(4/9) = 0.444	(4/8) = 0.5
10	(4/10) = 0.4	(4/8) = 0.5
11	(4/11) = 0.636	(4/8) = 0.5
12 = R	(5/12) = 0.417	(5/8) = 0.625
13 = R	(6/13) = 0.462	(6/8) = 0.75
14	(6/14) = 0.429	(6/8) = 0.75
15	(6/15) = 0.4	(6/8) = 0.75
16	(6/16) = 0.375	(6/8) = 0.75
17 = R	(7/17) = 0.412	(7/8) = 0.875
18	(7/18) = 0.389	(7/8) = 0.875
19 = R	(8/19) = 0.421	(8/8) = 1.0
20	(8/20) = 0.4	(8/8) = 1.0

IR3

K	**P@K**	**R@K**
1	(0/1) = 0.0	(0/8) = 0.0
2 = R	(1/2) = 0.5	(1/8) = 0.125
3	(1/3) = 0.333	(1/8) = 0.125
4 = R	(2/4) = 0.5	(2/8) = 0.25
5	(2/5) = 0.4	(2/8) = 0.25
6 = R	(3/6) = 0.5	(3/8) = 0.375
7	(3/7) = 0.429	(3/8) = 0.375
8 = R	(4/8) = 0.5	(4/8) = 0.5
9	(4/9) = 0.444	(4/8) = 0.5
10	(4/10) = 0.4	(4/8) = 0.5
11	(4/11) = 0.364	(4/8) = 0.5
12	(4/12) = 0.333	(4/8) = 0.5
13 = R	(5/13) = 0.385	(5/8) = 0.625
14	(5/14) = 0.357	(5/8) = 0.625
15 = R	(6/15) = 0.4	(6/8) = 0.75
16 = R	(7/16) = 0.438	(7/8) = 0.875
17 = R	(8/17) = 0.471	(8/8) = 1.0
18	(8/18) = 0.444	(8/8) = 1.0
19	(8/19) = 0.421	(8/8) = 1.0
20	(8/20) = 0.4	(8/8) = 1.0