

**Faculty of Engineering & Technology – Electrical & Computer Engineering Department**

**Second Semester 2022 – 2023**

**Natural language processing (NLP)-ENCS5342**

**Assignment #4**

**Name: Amal Ziad**

**ID: 1192141**

**Section: 1**

**Instructor: Dr. Adnan Yahya**

**Date: 19th June 2023**

1:**IR Evaluation:** Assume the following ranking for a given query(only results1-10are shown).

* The column ‘rank’ gives the rank of the document.
* The column ‘docid’ gives the document ID associated with the document at that rank.
* The column ‘gradedrelevance’ gives the relevance grade associated with the document (4=perfect,3=excellent,2=good,1=fair, and 0=bad).
* The column ‘binary relevance’ provides two values of relevance (1=relevantand0=non-relevant).

The assumption is that anything with a relevance grade of ‘fair’ or better is **relevant** and that any thing with a relevance grade of ‘poor’ is **non-relevant**. Also, assume that this query has only 7 documents with a relevance grade of fair or better. All happen to be ranked within the top10 in this given ranking.

Answer thequestions below. For P@K, R@K, and average precision (AP) assume binary relevance. For those metrics, use the ‘binary relevance’ column.

DCG and NDCG assume graded relevance. For those metrics, use the ‘gradedrelevance’ column. Show all your work.

|  |  |  |  |
| --- | --- | --- | --- |
| Rank | docid | graderelevance | binaryrelevance |
| 1 | 43 | 3 | 1 |
| 2 | 531 | 0 | 0 |
| 3 | 183 | 4 | 1 |
| 4 | 102 | 2 | 1 |
| 5 | 10 | 2 | 1 |
| 6 | 1051 | 0 | 0 |
| 7 | 1031 | 1 | 1 |
| 8 | 332 | 1 | 1 |
| 9 | 573 | 0 | 0 |
| 10 | 128 | 2 | 1 |

(a) Compute P@6 and P@10.

[P@6=4/6=0.66](mailto:P@6=4/6=0.66)

P@10=7/10=0.7

(b) Compute R@6 and R@10.

R@6=4/7=0.57

R@10=7/7=1

(c)Provide an example ranking for this query that maximizes P@6[5points]

Rank docid gradedrelevance binaryrelevance

1. 43 3 1
2. 1031 1 1
3. 183 4 1
4. 102 2 1
5. 10 2 1
6. 332 1 1

P@6=1

(d)Provide an example ranking for this query that maximizes P@10[5points]

Rank docid gradedrelevance binaryrelevance

1. 43 3 1
2. 128 2 1
3. 183 4 1
4. 102 2 1
5. 10 2 1
6. 332 1 1
7. 1031 1 1
8. 1051 0 0
9. 573 0 0
10. 531 0 0

P@10=1

(e)Provide an example ranking for this query that maximizes R@6[5points]

1. DocID 43
2. DocID 183
3. DocID 102
4. DocID 10
5. DocID 1031
6. DocID 332

R@6=1

(f)Provide an example ranking for this query that maximizes R@10[5points]

1. DocID 43
2. DocID 183
3. DocID 102
4. DocID 10
5. DocID 1031
6. DocID 332
7. DocID 128
8. DocID 531
9. DocID 1051
10. DocID 573

(g)Compute average precision (AP).[10points]

For the main data

|  |  |  |
| --- | --- | --- |
| Pos | R | P |
| 1 | 1/7 | 1 |
| 2 | 1/7 | 1/2 |
| 3 | 2/7 | 2/3 |
| 4 | 3/7 | ¾ |
| 5 | 4/7 | 4/5 |
| 6 | 4/7 | 4/6 |
| 7 | 5/7 | 5/7 |
| 8 | 6/7 | 6/8 |
| 9 | 6/7 | 6/9 |
| 10 | 1 | 7/10 |

**AP=0.7194**

(h)Provide an example ranking for this query that maximizes average precision(AP).[5points]

1. DocID 183 (gradedrelevance: 4, binaryrelevance: 1)
2. DocID 43 (gradedrelevance: 3, binaryrelevance: 1)
3. DocID 102 (gradedrelevance: 2, binaryrelevance: 1)
4. DocID 10 (gradedrelevance: 2, binaryrelevance: 1)
5. DocID 128 (gradedrelevance: 2, binaryrelevance: 1)
6. DocID 332 (gradedrelevance: 1, binaryrelevance: 1)
7. DocID 1031 (gradedrelevance: 1, binaryrelevance: 1)
8. DocID 531 (gradedrelevance: 0, binaryrelevance: 0)
9. DocID 573 (gradedrelevance: 0, binaryrelevance: 0)
10. DocID 1051 (gradedrelevance: 0, binaryrelevance: 0)

|  |  |  |
| --- | --- | --- |
| Pos | R | P |
| 1 | 1/7 | 1/1 |
| 2 | 2/7 | 2/2 |
| 3 | 3/7 | 3/3 |
| 4 | 4/7 | 4/4 |
| 5 | 5/7 | 5/5 |
| 6 | 6/7 | 6/6 |
| 7 | 7/7 | 7/7 |
| 8 | 7/7 | 7/8 |
| 9 | 7/7 | 7/9 |
| 10 | 7/7 | 7/10 |

**AP=0.935**

(i)Compute DCG5 (i.e., the discounted cumulative gain at rank 5).[10points]

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Rank | docid | Graderelevance (rel) | CG | logn | Rel/log | DCGn |
| 1 | 43 | 3 | 3 | - |  | 3 |
| 2 | 531 | 0 | 3 | 1 | 0 | 3 |
| 3 | 183 | 4 | 7 | 1.58 | 2.53 | 5.53 |
| 4 | 102 | 2 | 9 | 2 | 1 | 6.53 |
| 5 | 10 | 2 | 11 | 2.32 | 0.86 | 7.39 |
| 6 | 1051 | 0 | 11 | 2.58 | 0 | 7.39 |
| 7 | 1031 | 1 | 12 | 2.81 | 0.36 | 7.75 |
| 8 | 332 | 1 | 13 | 3 | 0.33 | 8.08 |
| 9 | 573 | 0 | 13 | 3.17 | 0 | 8.08 |
| 10 | 128 | 2 | 15 | 3.32 | 0.6 | 8.68 |

**DCG5=7.39**

(j)NDCG5 is given by NDCG5=DCG5/IDCG5 , where IDCG5 is the DCG5 associated with the ideal top-5 ranking associated with this query. Computing NDCG5 requires three steps.[10points]

(i)What is the ideal top-5 ranking associated with this query (notice that the query has 1 perfect document,1 excellent document, 3 good documents, 2 fair documents, and the rest are bad).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Rank | docid | Graderelevance (rel) | CG | logn | IDCGn |
| 1 | 183 | 4 | 4 | - | 4 |
| 2 | 43 | 3 | 7 | 1 | 5 |
| 3 | 102 | 2 | 9 | 1.58 | 6.58 |
| 4 | 10 | 2 | 11 | 2 | 8.58 |
| 5 | 128 | 2 | 13 | 2.32 | 10.9 |

(ii)IDCG5 is the DCG5 associated with the ideal ranking. ComputeIDCG5. (Hint: compute DCG5 for your ranking proposed in part(i).)

IDCG5 is for rank 4 in the ideal ranking 🡪 IDCG5=8.58

(iii)Compute NDCG5 using the formula above.

Rank 5 before now is rank 4 🡪 NDCG5=7.39/8.58=0.8613

(j):Interpolated Precision-Recall Curves

A Precision-Recall (PR) curve expresses precision as a function of recall. Usually, a PR-curve is computed for each query in the evaluation set and then averaged. For simplicity, the goal in this question is to draw a PR-curve for a single query. Draw the PR-curve associated with the ranking above (same query, same results, computed for 11 recall points).

1. **Auction:** Five companies are bidding for a given word. Using Second Price Auction discussed determine the amount paid by each bidder for a user Click. Work with the following table.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Company | Bid | CTR | Bid\*CTR | Rank | Paid |
| A | 1 | 0.08 | **0.08** | **5** | **(min)** |
| B | 2 | 0.05 | **0.1** | **4** | **1.61** |
| C | 3 | 0.07 | **0.21** | **1** | **2.86** |
| D | 4 | 0.04 | **0.16** | **3** | **2.51** |
| E | 5 | 0.04 | **0.2** | **2** | **4.01** |
|  |  |  |  |  |  |

1. **Duplicate Detection:** Use three permutations to detect near duplicates. Determine how related (near duplicate) are each pair of documents given that we have 5 shingles overall and D1 has S1, S2, S3; D2 has S1, S2, S4; D3 has S2, S3, S4;S5.

|  |  |  |  |
| --- | --- | --- | --- |
| Sn | D1 | D2 | D3 |
| S1 | 1 | 1 | 0 |
| S2 | 1 | 1 | 1 |
| S3 | 1 | 0 | 1 |
| S4 | 0 | 1 | 1 |
| S5 | 0 | 0 | 1 |

Assume functions🡪 h(x)=x mod 5 , g(x)=(2x+1)mod 5

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | D1 | slot | D2 | slot | D3 | slot |
| H |  | ∞ |  | ∞ |  | ∞ |
| g |  | ∞ |  | ∞ |  | ∞ |
| H(1) | 1 | 1 | 1 | 1 | - | ∞ |
| G(1) | 3 | 3 | 3 | 3 | - | ∞ |
| H(2) | 2 | 1 | 2 | 1 | 2 | 2 |
| G(2) | 0 | 0 | 0 | 0 | 0 | 0 |
| H(3) | 3 | 1 | - | 1 | 3 | 2 |
| G(3) | 2 | 0 | - | 0 | 2 | 0 |
| H(4) | - | 1 | 4 | 1 | 4 | 2 |
| G(4) | - | 0 | 4 | 0 | 4 | 0 |
| H(5) | - | 1 | - | 1 | 0 | 0 |
| G(5) | - | 0 | - | 0 | 1 | 0 |

J(D1,D2)=2/2=1 (100% duplicate)

J(D1,D3)=1/2=0.5 (50% similarity)

J(D2,D3)=1/2=0.5 (50% similarity)

# Good Luck