Week 6: Evaluation

# Assignment 1: Test Collections

a) Find three other test collections used for information retrieval evaluation in the literature and describe them (number of documents, number of queries, how relevance was assessed).

|  |  |  |  |
| --- | --- | --- | --- |
| **Collection** | **Documents** | **Queries** | **Relevance assessment** |
| Cranfield 2 (1962) | 1,400 | 225 | Manually-rated references |
| ADI (1968) | 82 | 35 | Manual binary rating, based on abstracts |
| IRE-3 (1968) | 780 | 34 | Manual binary rating, based on abstracts |

“Doubt in relevance decisions was usually settled by accepting the document as relevant.”

Sources:

“Test Collection Based Evaluation of Information Retrieval Systems” by Mark Sanderson

<http://www-nlpir.nist.gov/projects/irlib/pubs/irs13/irs13_text/p1-09.txt>

# Assignment 2: Measures

Compute the following evaluation measures:

a) Precision and recall

Precision: 6 / 10 = 0.6

Recall: 6 / 40 = 0.15

b) Precision at 7 and recall at 7

Precision at 7: 4 / 7 = 0.57

Recall at 7: 4 / 40 = 0.1

*c) MAP*

MAP / average precision: (1 + 2/3 + 3/4 + 4/7 + 5/9 + 6/10) / 6 = 0.69

*d) NDCG (assume binary gain value)*

DCG: 1 + 1/log2(3) + 1/log2(4) + 1/log2(7) + 1/log2(9) + 1/log2(10) = 3.10

Perfect ranking: 1, 1, 1, 1, 1, 1, 0, 0, 0, 0

Ideal DCG: 1 + 1/log2(2) + 1/log2(3) + 1/log2(4) + 1/log2(5) + 1/log2(6) = 3.95

NDCG: 3.10 / 3.95 = 0.78

# Assignment 3: (Programming) Evaluation

Print the ranking of Wikipedia titles in your development set together with its NDCG values for the queries:

a) “Anschluss”

Anschluss (Soziologie)

Ang Lee

Anschlussfähigkeit

nDCG@10: 1.0

b) “Soziologie”

Anschluss (Soziologie)

nDCG@10: 0.0

The (questionable) NDCG rating of 1.0 in a) is caused by our filtering: We throw away all results in the “ideal” list that do not appear in our results.