## Round 4 results — Run ILPS_UvA_big_diverse submitted from ILPS_UvA

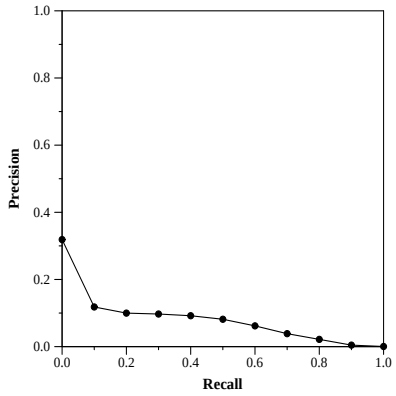
### Run Description

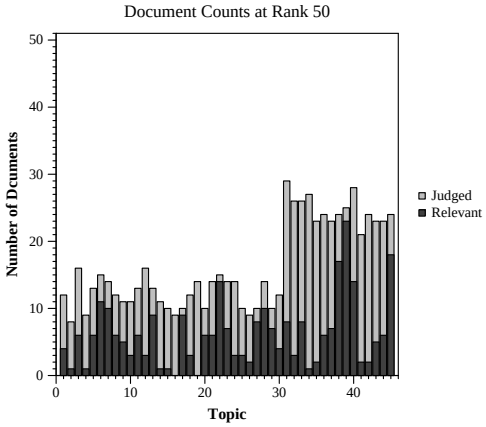
We used a weak supervision technique to rerank with BERT (only abstracts). This run was selected due to a big amount of training examples used and a big gap between our soft and hard evaluation protocol, which could suggest that diverse predictions are made by the model.

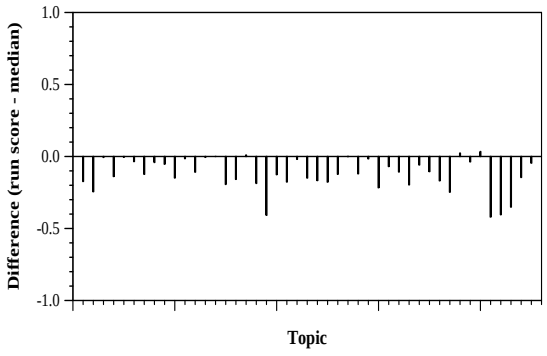
| Summary Statistics | |
|---|---|
| Run ID | ILPS_UvA_big_diverse |
| Topic type | feedback |
| Contributed to judgment sets? | yes |

| Overall measures | |
|---|---|
| Number of topics | 45 |
| Total number retrieved | 35945 |
| Total relevant | 5824 |
| Total relevant retrieved | 3512 |
| MAP | 0.0625 |
| Mean Bpref | 0.3921 |
| Mean NDCG@20 | 0.0970 |
| Mean RBP(p=0.5) | 0.0716 +0.8351 |

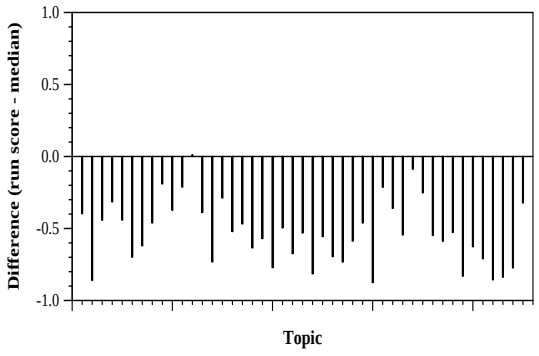
| Document Level Averages | |
|---|---|
| | Precision |
| At 5 docs | 0.0800 |
| At 10 docs | 0.1000 |
| At 15 docs | 0.1156 |
| At 20 docs | 0.1400 |
| At 30 docs | 0.1422 |
| R-Precision | |
| Exact | 0.0793 |



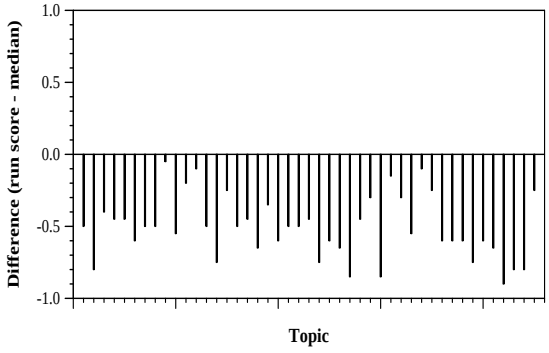



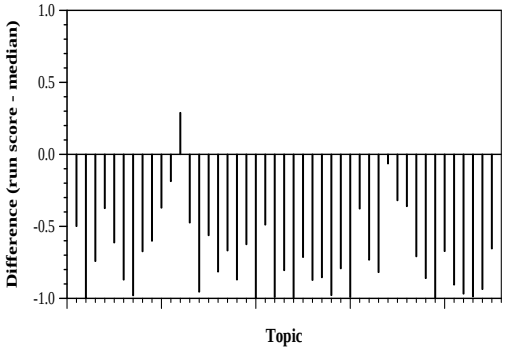

Per-topic difference from median bpref for all Round 4 runs




Per-topic difference from median NDCG@20 for all Round 4 runs




Per-topic difference from median P@20 for all Round 4 runs




Per-topic difference from median RBP(p=0.5) for all Round 4 runs