# Overview of the TREC 2018 CENTRE Track

Ian Soboroff
NIST

Nicola Ferro and Maria Maistro
University of Padua

Tetsuya Sakai
Waseda University

**Abstract**

The CLEF-NTCIR-TREC Reproducibility track (CENTRE) is a research replication and reproduction effort spanning three major information retrieval evaluation venues. In the TREC edition, CENTRE participants were asked to reproduce runs from either the TREC 2016 clinical decision support track, the 2013 web track, or the 2014 web track. Only one group participated in the track, and unfortunately the track will not continue in 2019.

## 1 Introduction

The goal of the cross-campaign CENTRE effort was to develop and tune a reproducibility evaluation protocol. In particular, we targeted two specific objectives:

- Replicability: different team, same experimental setup;
- Reproducibility: different team, different experimental setup.

Organizers selected, among the methods/systems submitted to the CLEF, NTCIR, and TREC ad-hoc tasks over the years, the top performing and most impactful ones. Each participating group was challenged to replicate and/or reproduce one or more of the selected systems using standard open source IR systems, like Lucene, Terrier, and others. Each participating group submitted one or more runs representing the output of their reproduced systems. They also developed and integrated into the open source IR system all the missing components and resources needed to replicate/reproduce the selected systems. Finally, they were asked to contribute back to open source all the developed components, resources, and configuration via a common repository.

Therefore, the goal of CENTRE was to run a joint CLEF/NTCIR/TREC task that challenges participants:

- to reproduce best results of best/most interesting systems in previous editions of CLEF, NTCIR, TREC by using standard open source IR systems;

- to contribute back to the community the additional components and resources developed to reproduce the results in order to improve existing open source systems.

The CENTRE tracks[1] took place at the three major evaluation venues, TREC, NTCIR, and CLEF.[12, 5, 4] For the TREC edition of the track, a particular goal was to develop and refine a report format that would serve as a template to other reproduction efforts in the IR community.

## 2 Tasks

The TREC CENTRE track had three tasks:

1. **Merck at TREC 2016 Clinical Decision Support:** Replicating runs from the Merck group participation in the TREC 2016 Clinical Decision Support (CDS) track.[11, 6] The CDS task was to retrieve biomedical journal articles relevant to a patient represented by a de-identified clinical note. The Merck system employed a number of advanced features including word embeddings, pseudo-relevance feedback, sophisticated query parsing, and learning-to-rank.

2. **Delaware at TREC 2013 Web:** Replicating runs from the University of Delaware (Fang) group in the TREC 2013 Web track.[2, 14] The 2013 Web Track task was to retrieve web pages from the ClueWeb12 collection in response to single- and multi-faceted topics. The Delaware system uses an axiomatic retrieval model.

3. **Glasgow at TREC 2014 Web:** Replicating runs from the University of Glasgow group in the TREC 2014 Web track.[3, 9] The TREC 2014 task was the same as in the 2013 Web Track, with a new topic set. The Glasgow system uses a risk-sensitive learning to rank approach.

In each task, participating groups were asked to attempt to reproduce the results of the target group/system. Whether the reproduction was successful would be measured using the original relevance assessments and comparison of the ranking of the reproduction system against the original systems. If resources allowed, we hoped to do additional relevance assessments to explore bias.

To facilitate the tasks, the respective topics, collections, relevance judgments, system outputs, overview paper, and track guidelines were collected on the track homepage.[2]

The track had one participating group, the Anserini team from the University of Waterloo.[15] That group submitted six runs to task 2 (2013 Web Track, UDel_Fang), three attempts for each original UDel_Fang group run. Additionally, the University of Padua submitted unofficial runs for task 1 after the track deadline, with three variations on the stoplist for each of the original three Merck runs.[10]

[1] http://www.centre-eval.org/

[2] http://www.centre-eval.org/trec2018/index.html

# 3 Task 1: Clinical 2016

The University of Padua submitted unofficial runs for task 1, which were accepted as no new relevance assessments were planned for task 1. The Merck 2016 runs identified for reproduction were *MRKUmlsSolr, MRKSumCln*, and *MRKPrfNote*, which explored different sources of terms for query expansion.[6] The Padua group submitted three runs covering various instantiations of the stoplist used in pseudo-relevance feedback, which was not specified in the original work. Table 1 shows the inferred average precision scores for the new Padua runs (in red) and the original runs (in black) from the track. Note that the rank ordering of the respective Merck runs is by and large reproduced in the Padua runs.

| run | infAP |
|---|---|
| ManualRun | 0.0535 |
| AutoSummary1 | 0.0454 |
| SumES | 0.0321 |
| CCNUSUMR1 | 0.0316 |
| cbnus1 | 0.0316 |
| MrkUmlsXgb | 0.0315 |
| ECNUrun5 | 0.0313 |
| UDelInfoCDS5 | 0.0311 |
| DUTHsaRPF | 0.0302 |
| udelSRef | 0.0302 |
| ECNUrun1 | 0.0296 |
| cbnus2 | 0.0295 |
| nkuRun1 | 0.0289 |
| SDPHBo1NE | 0.0287 |
| nkuRun3 | 0.0286 |
| **MRKUmlsSolr** | **0.0285** |
| udelSB | 0.0283 |
| DUTHmaRPF | 0.0281 |
| nkuRun5 | 0.0276 |
| ECNUrun3 | 0.0276 |
| UWM1 | 0.0274 |
| ETHSummRR | 0.0272 |
| **MRKSumCln** | **0.0272** |
| udelSDI | 0.0263 |
| ETHSumm | 0.0261 |
| AutoSummary | 0.0258 |
| sacmmf | 0.0257 |
| DAdescTM | 0.0255 |
| ETHNoteRR | 0.0254 |
| DAsummTM | 0.0253 |
| mayoas | 0.0252 |
| **Smart_ims_unipd-MRKUmlsSolr** | **0.0244** |

| run | infAP |
|---|---|
| ECNUmanual | 0.0243 |
| ECNUrun4 | 0.0242 |
| NLMrun1 | 0.0239 |
| ETHNote | 0.0239 |
| SumCmbRank | 0.0239 |
| **Extended_ims_unipd-MRKUmlsSolr** | **0.0239** |
| **Extended_ims_unipd-MRKSumCln** | **0.0238** |
| **Smart_ims_unipd-MRKSumCln** | **0.0232** |
| NLMrun2 | 0.0230 |
| NLMrun3 | 0.0228 |
| DAnoteTM | 0.0227 |
| NDPHBo1C | 0.0221 |
| **Lucene_ims_unipd-MRKUmlsSolr** | **0.0221** |
| cbnun1 | 0.0217 |
| UWM2 | 0.0214 |
| DAnoteRoc | 0.0214 |
| **Lucene_ims_unipd-MRKSumCln** | **0.0210** |
| UWM0 | 0.0209 |
| NDPHBo1CM | 0.0207 |
| run5 | 0.0205 |
| udelNRef | 0.0203 |
| NLMrun5 | 0.0202 |
| nkuRun2 | 0.0198 |
| run1 | 0.0196 |
| DUTHaaRPF | 0.0192 |
| DAnote | 0.0186 |
| run4 | 0.0185 |
| NoteES | 0.0185 |
| udelNB | 0.0181 |
| WHUIRGroup6 | 0.0179 |
| **MRKPrfNote** | **0.0179** |
| DDPHBo1CM | 0.0172 |
| CSIROmnul | 0.0168 |
| ETHDescRR | 0.0165 |
| run3 | 0.0162 |
| DDPHBo1MWRe | 0.0154 |
| UNTIIANA | 0.0153 |
| UNTIIASA | 0.0153 |
| NLMrun4 | 0.0146 |
| UNTIIANM | 0.0144 |
| SumClsRerank | 0.0143 |
| **Extended_ims_unipd-MRKPrfNote** | **0.0143** |
| nkuRun4 | 0.0142 |
| lssbs | 0.0141 |
| CCNUNOTER1 | 0.0140 |

| run | infAP |
|---|---|
| summUIOWAS3 | 0.0140 |
| run2 | 0.0140 |
| **Smart_ims_unipd-MRKPrfNote** | **0.0137** |
| AutoNote | 0.0134 |
| UNTIIANMERG | 0.0132 |
| AutoDes | 0.0131 |
| mayomn | 0.0130 |
| mayoad | 0.0121 |
| **Lucene_ims_unipd-MRKPrfNote** | **0.0119** |
| CSIROsumm | 0.0119 |
| UNTIIASMERG | 0.0113 |
| mayoan | 0.0113 |
| mayomd | 0.0109 |
| CCNUDESR2 | 0.0105 |
| WHUIRGroup1 | 0.0104 |
| prna1sum | 0.0102 |

Table 1: Runs from the 2016 Clinical Decision Support task, with the University of Padua reproductions (red) of the Merck runs, scored by inferred average precision. Runs with infAP scores less that 0.01 have been elided for space.

## 4 Task 2: Web 2013

The 2013 Web Track collection was constructed using very shallow pools of either depth 10 or 20 depending on the topic.[2] Because of this, the collection was mostly useful for measuring the 2013 participating runs on precision-oriented metrics, and it could be quite likely that a system developed later would retrieve many unjudged documents and therefore its effectiveness would be uncertain. Since CENTRE asks participants to reproduce an existing system (and so, in some sense a well-measured approach) but using open-source software (that would quite likely differ in many internal aspects such as segmentation and tokenization), we planned early in the TREC year to devote some assessment resources for the CENTRE track. These assessments could only be made for tasks 2 and 3, which did not require assessors with a medical background.

Since in the end we only had participation in task 2, we decided to revisit the pooling for that task entirely. In 2013, some topics were pooled to depth 10 and some to depth 20. For the CENTRE track, we re-pooled all the 2013 web track runs to depth 30, and included the Anserini runs. The assessors judged a random sample of 10% of the previously-judged documents as well as the unjudged documents in the pool. This gave us two sets of relevance assessments we could use to study the Anserini runs: the original relevance assessments from 2013 and the combined new set of judgments. Furthermore, we could study the agreement between the 2013 and 2018 judgments in the agreement sample.[3]

Figure 1 shows the pool sizes. These multiple sets of relevance assessments would allow us to compute inter-annotator agreeement between the 2013 and the 2018 assessments, to make a reasonable comparison of the Anserini and original UDel_Fang runs, to examine whether the Anserini runs would have been unfairly measured using only the original judgments, and to see if the new judgments imply a different run ranking than the original judgments had in 2013.

## 4.1 Agreement

We measured agreement between the original assessors and the CENTRE track assessors. Agreement is not a strong concern for ranking systems generally [13] but here there is a special concern because we are pooling and judging with different assessors to measure specific runs which can introduce bias. Since the 2013 web track judgments are on a graded relevance scale (junk, not relevant, relevant, highly relevant, key, navigational query target), simple overlap measures are not suitable. We used Krippendorff's alpha measure [7], which allows for multiple raters and categorical scales in a principled way.[4] Alpha lies on a $[-1, 1]$ scale where 1.0 is perfect agreement, 0.0 is no relationship between the assessors, and negative values indicate systematic disagreement.

Figure 2 shows that assessor agreement on the sample varies quite widely, perhaps due to the sample size compared with the number of categories on the relevance scale. We designated two sets of relevance judgments: the union of original and new judgments for all topics ("centre"), and the union for topics where $\alpha > 0.2$ ("pruned").

We then measured all the runs from the TREC 2013 track plus the new runs from Anserini using three sets of relevance judgments: the originals from 2013, the "centre" set, and the "pruned" set. The main track metrics for the adhoc task were expected reciprocal rank (ERR) [1] and Normalized Discounted Cumulative Gain (nDCG) [8]. The Kendall's tau correlations among the three system rankings are (left column pair):

| | with Anserini | | without Anserini | |
|---|---|---|---|---|
| | ERR@10 | nDCG@10 | ERR@10 | nDCG@10 |
| $\tau(2013, \text{pruned})$ | 0.31 | 0.39 | 0.39 | 0.50 |
| $\tau(2013, \text{centre})$ | 0.56 | 0.55 | 0.65 | 0.69 |
| $\tau(\text{pruned}, \text{centre})$ | 0.52 | 0.53 | 0.53 | 0.50 |

We conclude that the new assessments rank the systems quite differently, even when the new judgments are pruned to topics with (relatively) high agreement. Comparing the correlations where we alternately include and leave out the Anserini runs indicates that the original pools are biased against the new

[3] Relevance assessment in information retrieval typically exhibits low agreement, but in practice differences in relevance do not significantly affect the relative rankings of systems.[13]

[4] The paper by Hayes and Krippendorff [7], available freely from his website (`afhayes.com/public/cmm2007.pdf`), gives a succinct discussion of all the major agreement measures proposed since the 1950s, describes Krippendorff's $\alpha$, and gives an implementation in SPSS.

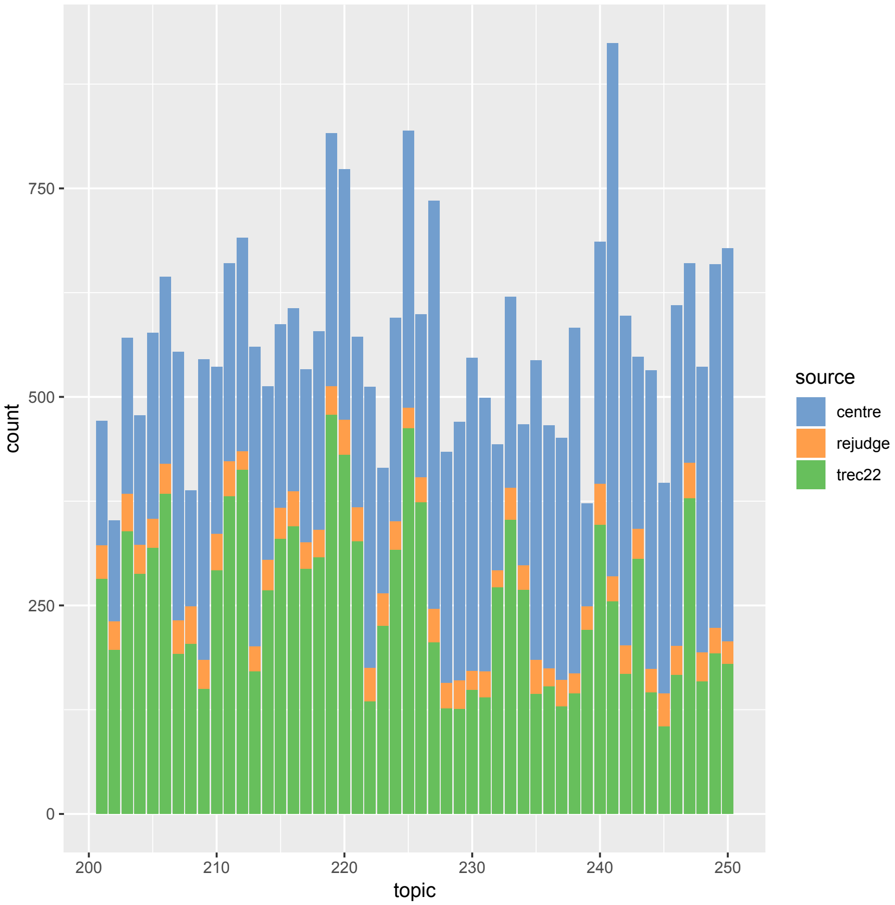

Figure 1: The number of documents assessed for the 2013 web track task. Documents marked *trec22* are from the original 2013 pools. We then deepened those pools and included the Anserini submissions (*centre*). 10% of the documents from the original runs were randomly selected for re-assessment (*rejudge*).

runs, but that making new judgments based on the new runs can bias the measurement in the other direction.

## 4.2 Evaluation

Figures 3 and 4 compare the ERR and nDCG scores using the original and augmented relevance judgments. The Anserini and UDel runs are called out in the plots. Table 2 gives the ERR scores for all runs against each set of relevance judgments.

In the augmented relevance judgements, ERR scores increase for all runs, and additionally the values are a bit more spread out, helping visually distinguish some points that were very close together originally. For nDCG, all scores decrease, which is sensible since the *centre* relevance judgments have essentially lengthened the ideal gain vector for everyone.

The critical point is to observe the relative placement of the Anserini and UDel WEB1 and WEB2 runs. The orders are not consistent between UDel and Anserini: for UDel, WEB2 is quite a bit higher-scoring than WEB1, whereas for Anserini the reverse is true. This indicates that the Anserini submission is not perfectly reproducing the original runs.

| run | pruned | centre | trec22 |
|---|---|---|---|
| **Anserini-UDInfolabWEB1-2** | **0.1671** | **0.2259** | **0.1303** |
| **UDInfolabWEB2** | **0.1438** | **0.1962** | **0.1666** |
| **UDInfolabWEB2R** | **0.1438** | **0.1962** | **0.1666** |
| **Anserini-UDInfolabWEB1-1** | **0.1247** | **0.1942** | **0.0942** |
| clustmrfaf | 0.1429 | 0.1938 | 0.1749 |
| udemFbWikiR | 0.1960 | 0.1885 | 0.1433 |
| udemQlm1lFbWiki | 0.1960 | 0.1885 | 0.1433 |
| uogTrAS2Lb | 0.1474 | 0.1825 | 0.1388 |
| uogTrAIwLmb | 0.0763 | 0.1802 | 0.1508 |
| ut22base | 0.0971 | 0.1799 | 0.1396 |
| **Anserini-UDInfolabWEB2-2** | **0.1215** | **0.1784** | **0.0929** |
| UJS13LCRAd1 | 0.1238 | 0.1761 | 0.0903 |
| clustmrfbf | 0.1364 | 0.1760 | 0.1152 |
| ICTNET13RSR3 | 0.1225 | 0.1753 | 0.1395 |
| mmrbf | 0.1328 | 0.1750 | 0.1169 |
| uogTrAS1Lb | 0.1387 | 0.1735 | 0.1392 |
| udelManExp | 0.1396 | 0.1725 | 0.1496 |
| udelPseudo2 | 0.1269 | 0.1714 | 0.1372 |
| ICTNET13ADR2 | 0.1229 | 0.1702 | 0.1205 |
| ICTNET13RSR1 | 0.1189 | 0.1664 | 0.1338 |
| cwiwt13cps | 0.1120 | 0.1663 | 0.1209 |
| ICTNET13ADR1 | 0.1240 | 0.1661 | 0.1170 |
| uogTrADnLrb | 0.1090 | 0.1642 | 0.1321 |
| ut22xact | 0.0852 | 0.1614 | 0.1444 |
| udemQlm1l | 0.1471 | 0.1604 | 0.1203 |

| run | pruned | centre | trec22 |
|---|---|---|---|
| udemQlml1R | 0.1471 | 0.1604 | 0.1203 |
| ICTNET13RSR2 | 0.0980 | 0.1600 | 0.1487 |
| udemQlm1lFb | 0.1504 | 0.1577 | 0.1052 |
| UJS13LCRAd2 | 0.0823 | 0.1547 | 0.1003 |
| UJS13Risk2 | 0.1090 | 0.1547 | 0.0999 |
| uogTrBDnLaxw | 0.0739 | 0.1540 | 0.1062 |
| **Anserini-UDInfolabWEB2-3** | **0.1415** | **0.1533** | **0.0554** |
| **Anserini-UDInfolabWEB1-3** | **0.1392** | **0.1515** | **0.0584** |
| cwiwt13cpe | 0.1111 | 0.1491 | 0.1012 |
| UJS13Risk1 | 0.1164 | 0.1484 | 0.0965 |
| **Anserini-UDInfolabWEB2-1** | **0.0761** | **0.1484** | **0.1165** |
| ut22spam | 0.1401 | 0.1444 | 0.1101 |
| udelPseudo1 | 0.0907 | 0.1411 | 0.0973 |
| UWCWEB13RISK01 | 0.0773 | 0.1410 | 0.0667 |
| uogTrBDnLmxw | 0.0750 | 0.1408 | 0.1132 |
| webishybrid | 0.1085 | 0.1382 | 0.0832 |
| wistud.runA | 0.0840 | 0.1358 | 0.1056 |
| udemQlml1FbR | 0.0783 | 0.1313 | 0.1062 |
| UWCWEB13RISK02 | 0.0924 | 0.1295 | 0.0798 |
| udelCombUD | 0.1041 | 0.1273 | 0.1195 |
| **UDInfolabWEB1** | **0.0786** | **0.1272** | **0.1073** |
| **UDInfolabWEB1R** | **0.0786** | **0.1272** | **0.1073** |
| msr_alpha0_95_4 | 0.1011 | 0.1261 | 0.0871 |
| wistud.runB | 0.0764 | 0.1254 | 0.0959 |
| msr_alpha1 | 0.1001 | 0.1247 | 0.0856 |
| msr_alpha10 | 0.0999 | 0.1237 | 0.0846 |
| msr_alpha0 | 0.0978 | 0.1235 | 0.0850 |
| wistud.runC | 0.0720 | 0.1233 | 0.0951 |
| wistud.runD | 0.0969 | 0.1226 | 0.1255 |
| msr_alpha5 | 0.0973 | 0.1220 | 0.0836 |
| RMITSC75 | 0.1002 | 0.1217 | 0.0931 |
| RMITSC | 0.1002 | 0.1216 | 0.0931 |
| webisrandom | 0.0723 | 0.1215 | 0.0933 |
| RMITSCTh | 0.0996 | 0.1212 | 0.0930 |
| webismixed | 0.0785 | 0.1160 | 0.0861 |
| ICTNET13ADR3 | 0.0723 | 0.1102 | 0.0883 |
| cwiwt13kld | 0.0925 | 0.1090 | 0.0774 |
| webiswtbaseline | 0.0721 | 0.1087 | 0.0819 |
| webiswikibased | 0.0945 | 0.1085 | 0.0862 |
| webisnaive | 0.0854 | 0.1042 | 0.0833 |
| udelPseudo1LM | 0.0624 | 0.0936 | 0.0813 |
| dlde | 0.0046 | 0.0247 | 0.0077 |

| run | pruned | centre | trec22 |
|---|---|---|---|

Table 2: ERR@10 scores for all runs against all three sets of relevance judgments.

# 5 Conclusion

With only two participants (one official) and in the midst of the first CENTRE cycle, we are hesitant to report conclusions beyond the limited experience presented here. Perhaps the strongest point we can make is that, indeed, replicating and reproducing experiments is hard. We supposed that the fairly rigorous and regular methodology of test collection experiments would help, and indeed it has, to the extent that the data used and the procedure followed are well defined. However, this perspective allowed us to more closely examine results, to consider the effect of assessor disagreement and test collection mismatch, and even to consider what it means to actually reproduce a result: are we looking for scores, or for effects, or for rank agreement, or something else?

There are formidable social challenges to reproducibility beyond the technical ones. Reproducing experiments is hard, but not very publishable, since without careful presentation no new research is done. The CENTRE tracks tried to substitute a social environment to encourage reproduction, with it must be said poor results in the TREC edition. We note that there are a wide array of social incentives being created for reproduction, from special tracks in conferences and journals to lauding open source reproductions to increase their social capital. Perhaps we are at an early stage of a long journey.

Reproduction of results with test collections that have not been built to handle recall is fraught. In the Web track task in TREC CENTRE, the participant paid close attention to detail, and effect sizes and rank orders seem to indicate success. But the runs found many unjudged documents, and reassessing documents five years after the original collection yielded low assessor agreement. This meant that we are not very confident of the evaluation either with the original relevance judgments (due to poor coverage) or with the augmented judgments (due to poor agreement). In order to support reproducing experiments on large data sets, we either need to solve the agreement problem, solve the pooling problem in large data sets, or come with another solution.

# References

[1] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 621–630, New York, NY, USA, 2009. ACM.

[2] Kevyn Collins-Thompson, Paul Bennett, Charlie Clarke, and Ellen M. Voorhees. TREC 2013 web track overview. In *Proceedings of the Twenty-*

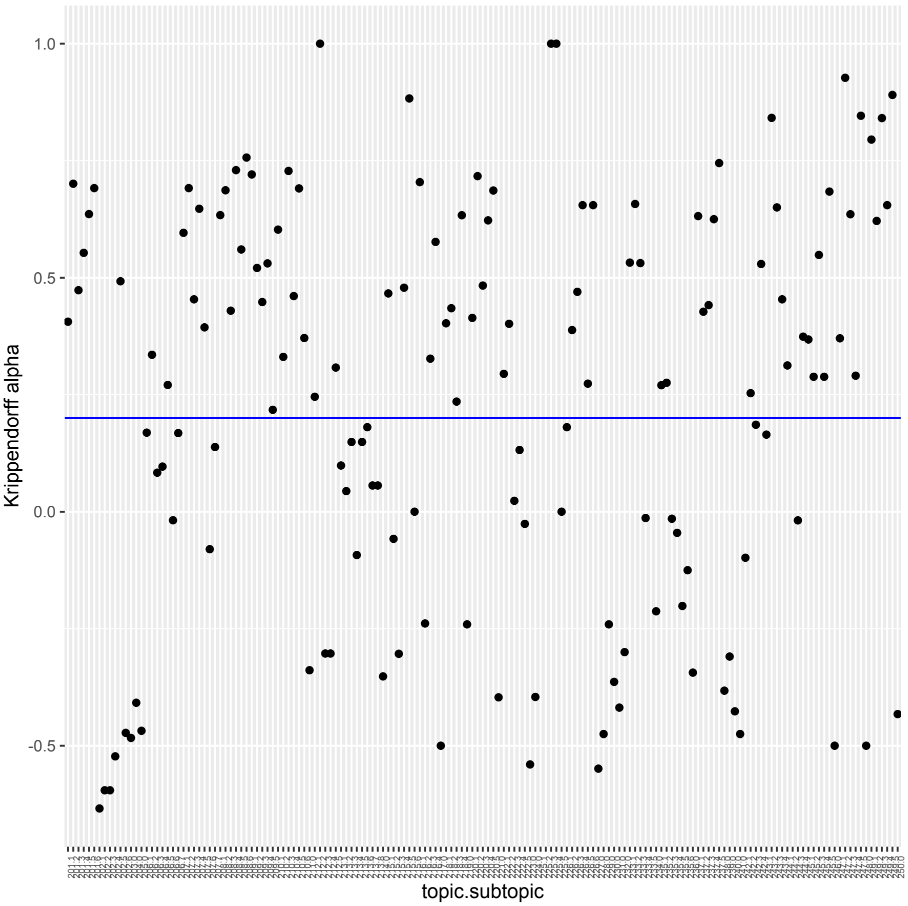

Figure 2: Values of Krippendorff's alpha for the 10% sample of documents judged by both the 2013 and the 2018 assessors. The blue line is the $\alpha > 0.2$ threshold used for selecting the topics to include in the *pruned* relevance judgment set.

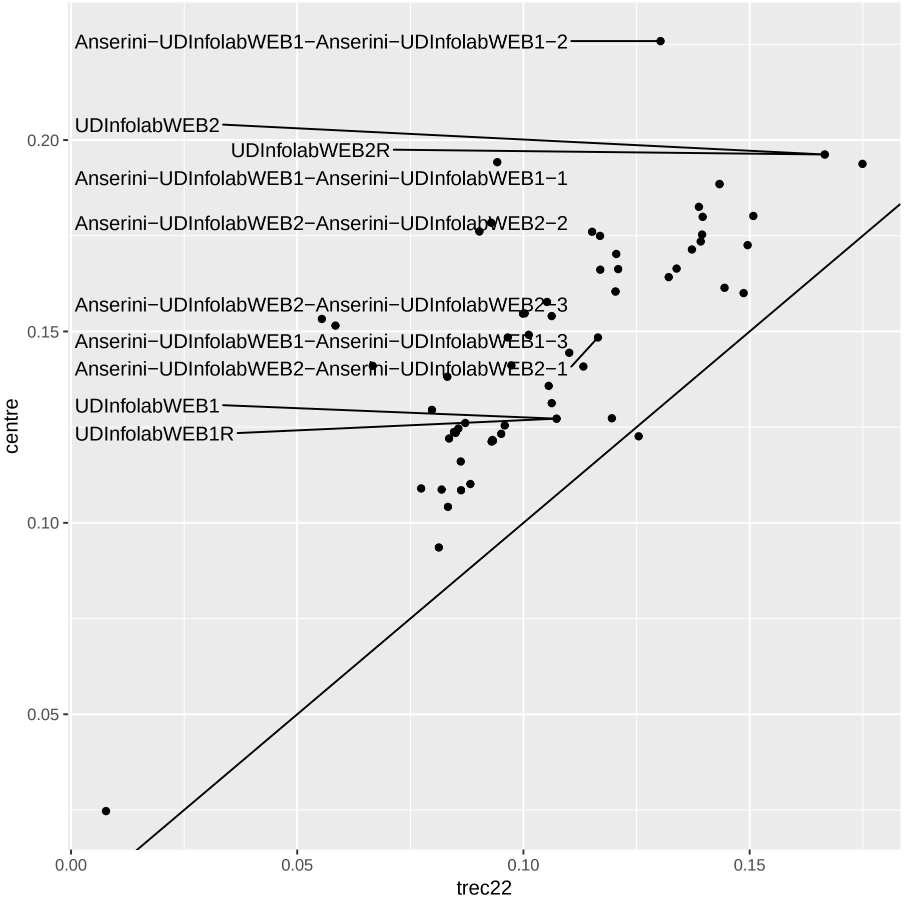

Figure 3: ERR@10 scores, based on the TREC 2013 qrels (x-axis) and the augmented CENTRE qrels (y-axis).

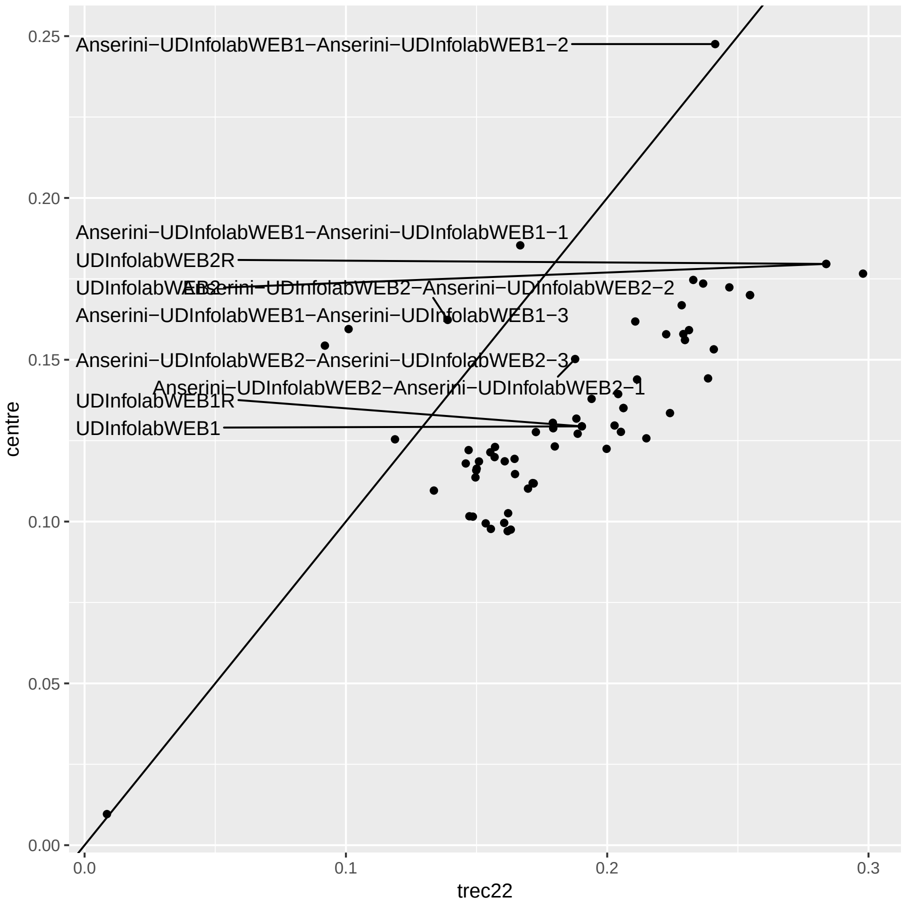

Figure 4: nDCG@10 scores, based on the TREC 2013 qrels (x-axis) and the augmented CENTRE qrels (y-axis).

*Second Text REtrieval Conference (TREC 2013)*, Gaithersburg, MD, November 2013.

[3] Kevyn Collins-Thompson, Craig Macdonald, Paul Bennett, and Ellen M. Voorhees. TREC 2014 web track overview. In *Proceedings of the Twenty-Third Text REtrieval Conference (TREC 2014)*, Gaithersburg, MD, November 2014.

[4] Nicola Ferro, Norbert Fuhr, Maria Maistro, Tetsuya Sakai, and Ian Soboroff. CENTRE@CLEF2019: Sequel in the systematic reproducibility realm. In *Proceedings of CLEF 2019*, LNCS 11696, pages pp. 287–300. Springer, 2019.

[5] Nicola Ferro, Maria Maistro, Tetsuya Sakai, and Ian Soboroff. Overview of CENTRE@CLEF2018: a first tale in the systematic reproducibility realm. In *Proceedings of CLEF 2018*, LNCS 11018, pages pp 239–246. Springer, 2018.

[6] Harsha Gurulingappa, Alexander Bauer, Luca Toldo, Claudia Schepers, and Gerard Megaro. Semi-supervised information retrieval system for clinical decision support. In *Proceedings of the Twenty-Fifth Text REtrieval Conference (TREC 2016)*, Gaithersburg, MD, November 2016.

[7] Andrew F. Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89, 2007.

[8] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002.

[9] Richard McCreadie, Romain Deveaud, M-Dyaa Albakour, Stuart Mackie, Nut Limsopatham, Craig Macdonald, Iadh Ounis, Thibaut Thonet, and Bekir Taner Dinçer. TREC 2014 web track overview. In *Proceedings of the Twenty-Third Text REtrieval Conference (TREC 2014)*, Gaithersburg, MD, November 2014.

[10] Giorgio Maria Di Nunzio and Stefano Marchesin. The University of Padua IMS research group at CENTRE@TREC 2018. In *Proceedings of the Twenty-Seventh Text REtrieval Conference (TREC 2018)*, Gaithersburg, MD, November 2018.

[11] Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, and William R. Hersh. Overview of the TREC 2016 clinical decision support track. In *Proceedings of the Twenty-Fifth Text REtrieval Conference (TREC 2016)*, Gaithersburg, MD, November 2016.

[12] Tetsuya Sakai, Nicola Ferro, Ian Soboroff, Zhaohao Zeng, Peng Xiao, and Maria Maistro. Overview of the NTCIR-14 CENTRE task. In *Proceedings of NTCIR-14*, pages pp. 494–509, 2019.

[13] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 315–323, New York, NY, USA, 1998. ACM.

[14] Peilin Yang and Hui Fang. Evaluating the effectiveness of axiomatic approaches in web track. In *Proceedings of the Twenty-Second Text REtrieval Conference (TREC 2013)*, Gaithersburg, MD, November 2013.

[15] Peiling Yang and Jimmy Lin. Anserini at TREC 2018: CENTRE, common core, and news tracks. In *Proceedings of the Twenty-Seventh Text REtrieval Conference (TREC 2018)*, Gaithersburg, MD, November 2018.