Stability of Dataset Comparisons for avg-idf (p=0.9) **Dataset Pair** beir-webis-touche2020 vs msmarco-passage 0.30 beir-webis-touche2020 vs msmarco-subsample msmarco-passage vs msmarco-subsample 0.25 0.20 0.10 0.05 0.00 2 3

Average Performance Delta

Swap Probability