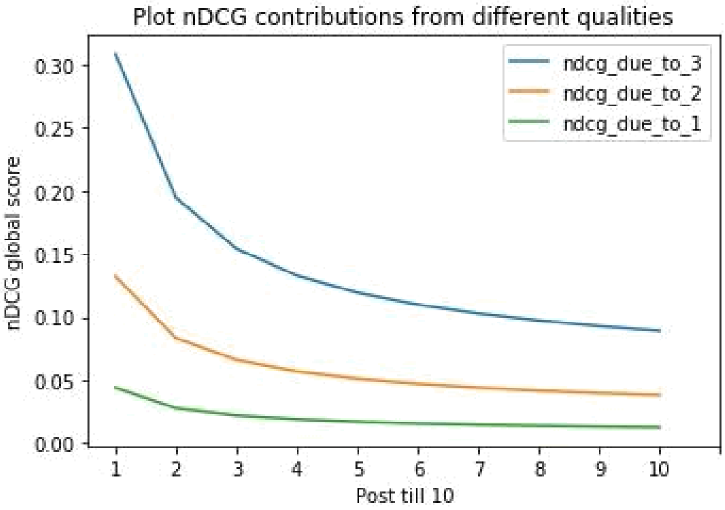
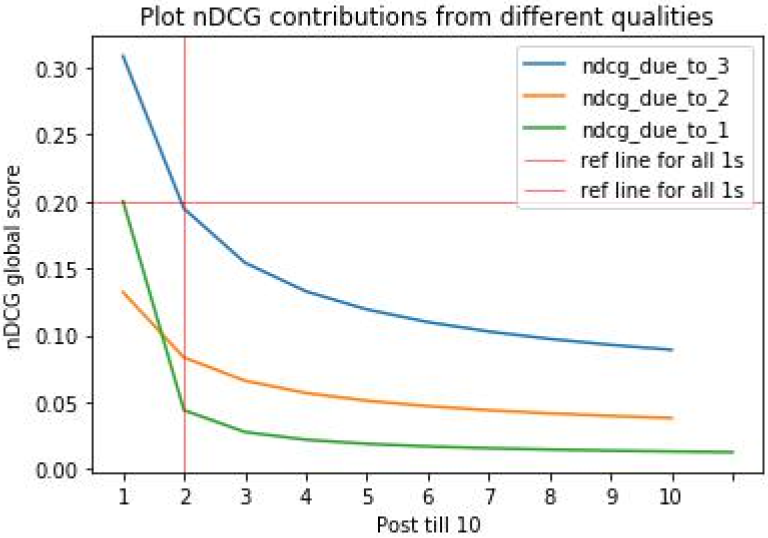
Ranking Metrics Sensitivity

Metrics - nDCG

1. **How does this impact metrics? - Case Study nDCG**
    a. 0s don't contribute at all
    b. 1s, 2, and 3s contribute proportional to their position in the results
    c. The below plot shows a contribution by a solo label in top 10, at different positions. As we can see a contribution from a 3 at any position is higher than the contribution by a 2 at any position than 1 at any position

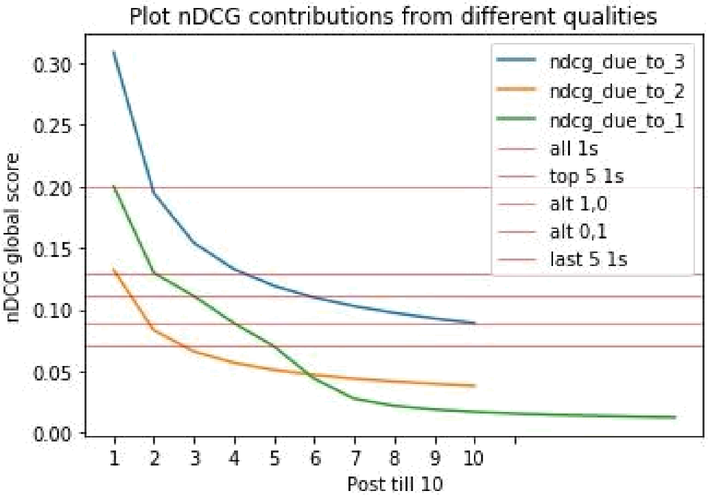


    d. When all the top 10 are 1s the nDCG value is higher than a two in the top 2 positions. **Meaning the contribution by a lot of 1s is than a contribution from a 2 at any position in the top 10, and a 3 at any position after the top 2.** Note: The green plot with 1 at different rank keeps sliding as we add more points.



    e. More granular cases - For the top 3 cases there is at least one position where a solo 3's contribution is lower than 1s and for all of them there is at least one position where a solo 2's contribution is lower than 1s

        i. For the green plot points are as follows
            1. All top 10 are 1s
            2. Top 5 are 1s
            3. 1,0s are alternating to simulate an average case of 50% 1s
            4. 0,1s are alternating to simulate an average case of 50% 1s
            5. 5 0s followed by 5 1s



    f.
    g. **Binary nDCG**
        i. nDCG can be computed with all the 3 and 2s changed to 1 (good) and all the 1s and 0s changed to 0 (bad)
        ii. What does this change?
            1. **Only labelled 2s and 3s can contribute towards the score How do we measure nDCG for unknown labels?**

1. We convert all the unknowns to label 0, i.e irrelevant.
2. This choice was made to prevent negative impact on the nDCG calculation as 0s don't contribute.
3. Even though the 0s don't add to the score, they could push known labels to lower positions.
4. Enter nDCG - Condensed metrics - These calculations ignore the unlabelled data points.
5. E.g. A random quality label order such as 2 UNK 3 1 1 0 UNK 3 2 UNK 1 0 0 becomes
    a. 2 0 3 1 1 0 0 3 2 0 1 0 0 for regular nDCG metrics
    b. 2 3 1 1 0 3 2 1 0 0 for the condensed metrics.
6. So a heavy presence of unknown labels negatively impacts nDCG with the current way of handling them
7. Consider another sequence 2 U U U U U U U 3 1 -> this becomes 2 3 1 for condensed, but we witness information loss as we use only 3 out of 10 labels, and this might not be sufficient to compare two systems.

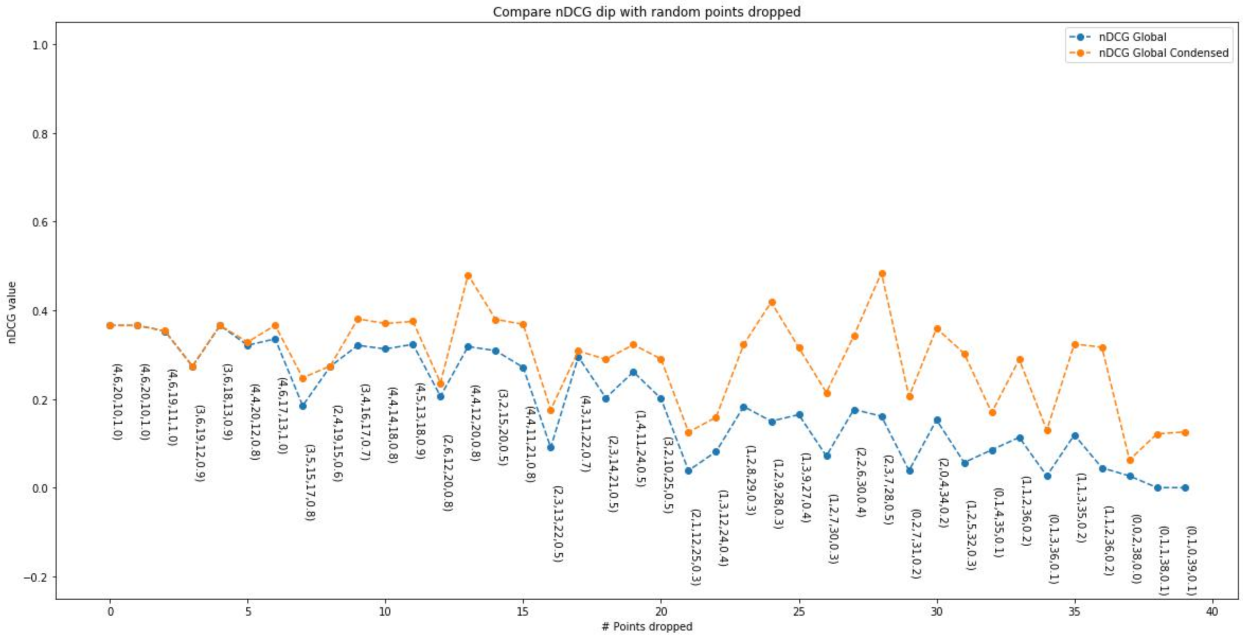
Dropping Labels

Ideal ordering of the labels: 3 * 4 + 2 * 6 + 1 * 20 + 0 * 10 A
random ranking system has sorted these labels like this

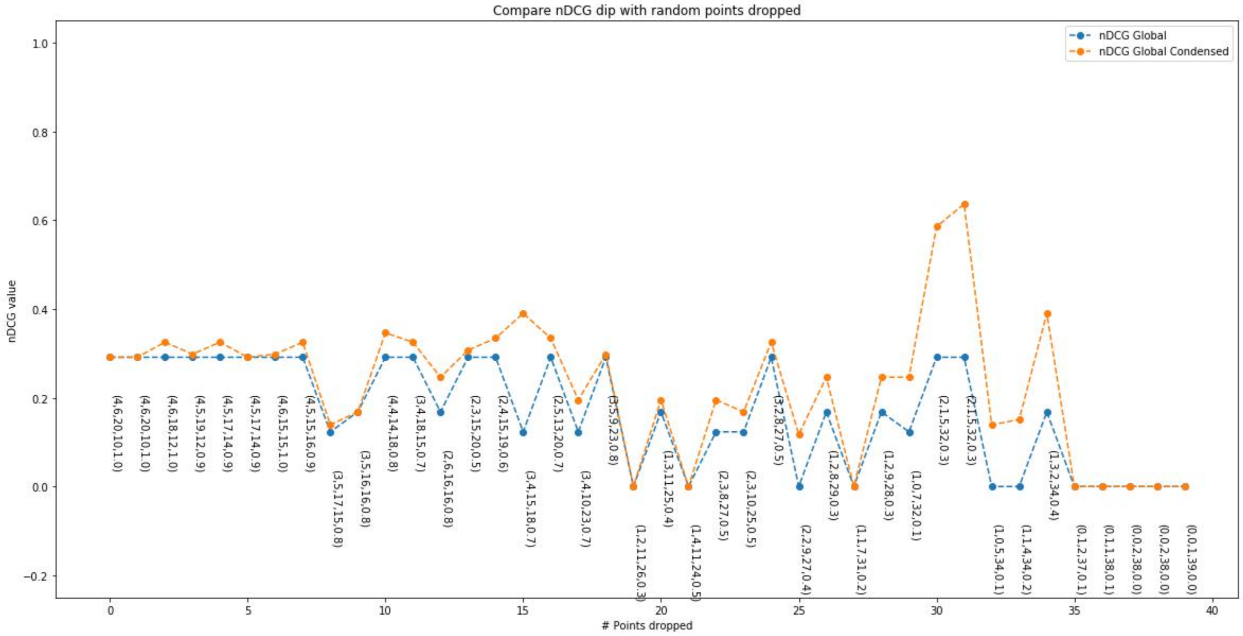
[0, 1, 1, 3, 1, 2, 1, 3, 1, 1, 0, 0, 0, 1, 1, 0, 2, 1, 0, 0, 2, 0, 1, 1, 0, 1, 1, 1, 3, 0, 1, 3, 1, 1, 1, 2, 2, 1, 1, 2]

The (a,b,c,d, e) tuple annotation near each point corresponds to the counts of each quality (#3, #2, #1, #0, % of 2s & 3s retained)

Let us drop a different percentage of random indices.



All qualities are considered for nDCG



Treat topical (1s) like Irrelevant (0s)

What do we notice?

1. Obviously dropping indices with high quality labels like 3, 2 dents the performance

2. As the #drop increases, the condensed and regular metrics start to diverge, i.e the impact of missing labels is low on condensed
3. Condensed and Regular metrics vary similarly when topical answers are removed from the dataset

Conclusions

1. A differing amount of missing labels between systems makes it difficult for a fair comparison between systems.
2. It is decided to only use the RELEVANT and HIGHLY RELEVANT labels for nDCG calculations.
3. Research should be done into exploring other metrics that can help overcome this problem.
4. A well labelled dataset will enable direct comparison for ranking tasksd