Ranking

CS276 Information Retrieval and Web Search
Programming Assignment 3

Maha El Choubassi (MELCHOUB) Christoph Wertz (CWERTZ)

Weights Choices

From our experiments, for each method, there are many choices for parameters to achieve near 0.85 NDCG, we choose the set of weights that:

- is verified to generalize to new development set with reasonable performance.
- the relative values of parameters reflect the relative importance of the fields. Statistically, we will show that our empirical tests agree with this intuition.

	url	title	header	body	anchor	smooth'g body	B	B_{url}	B_{title}	B_{header}	B_{body}	B_{anchor}	K1	λ	λ'	$NDCG_{train}$	$NDCG_{test}$
task1	1	0.6	0.4	0.2	0.2	500	-	-	-	-		-	-	-	-	0.8506	0.857
task2	1	0.8	0.6	0.4	0.2	-	-	0.8	1	0.2	0.4	0.6	1.7	1	0	0.8442	0.8493
task3	0.8	0.8	0.6	0.4	0.2	500	2	-	-	-		-	-	-	-	0.8554	0.8558

0.8492578132901406

Cosine Similarity Ranking Weights

We ran multigrid search with 6250 combinations for the url, title, header, body length, and anchor weights and the smoothing body length constant to be added to all body lengths:

5 possible values for each weight in [0.2, 1.0], and 2 values for smoothing body length in {500, 1000}

For these values, the NDCG score ranges from 0.841425 to 0.851205.Please see in figure 1 the histograms of the top 200 scores and the bottom 200 scores.

In figure 2, the histograms of every parameter for the top 200 scores are in blue, and the those for the bottom 200 scores are in red. Here are the general trends. The top scores are most likely to have:

- high url weight: expected as the URL can be thought of as an attempt by the site creator to organize pages by relevant topics as indicated by the terms in the URL.
- high title weight: similarly as the URL, the title of a relevant document should be highly dense in relevant terms.
- a little lower headers weight: it is expected not to be small.
- low body weight: it is not surprising, since the body can be very long and have the terms from the query scattered and not necessarily reflecting relevance of the document.
- smoothing body length of 500 is much more preferable to 1000.

As for the anchor weight, for low NDCG scores, most often the anchor weight is low. But even for high scores, the anchor can have low weight, we believe that is compensated for by the magnification of the anchor importance with its multiplicity number.

For sublinear scaling, the only fields that may need this are the body and the anchors. With long bodies and highly repeated anchors, there might be a big range in the possible values of the term frequencies. Sublinear scaling can down tone this discrepancy. We did try with and without sublinear scaling on the body and on the anchors. We did not see much difference and we sticked without using sublinear scaling.

We chose weights: task1_W_url = 1; task1_W_title = 0.6; task1_W_header = 0.4; task1_W_body = 0.2; task1_W_anchor = 0.2 with task1_NDCG_train = 0.8506 and task1_NDCG_test = 0.857

task2_W_url = 1.0; task2_W_title = 0.8; task2_W_header = 0.6; task2_W_body = 0.4; task2_W_anchor = 0.2

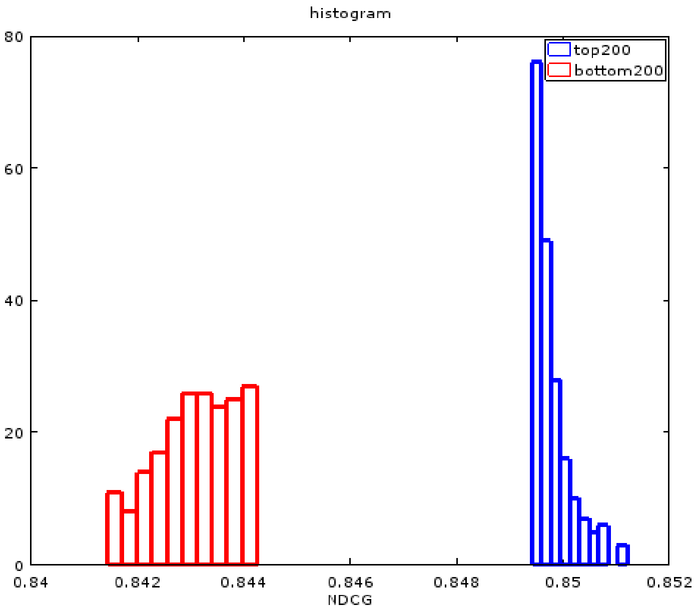
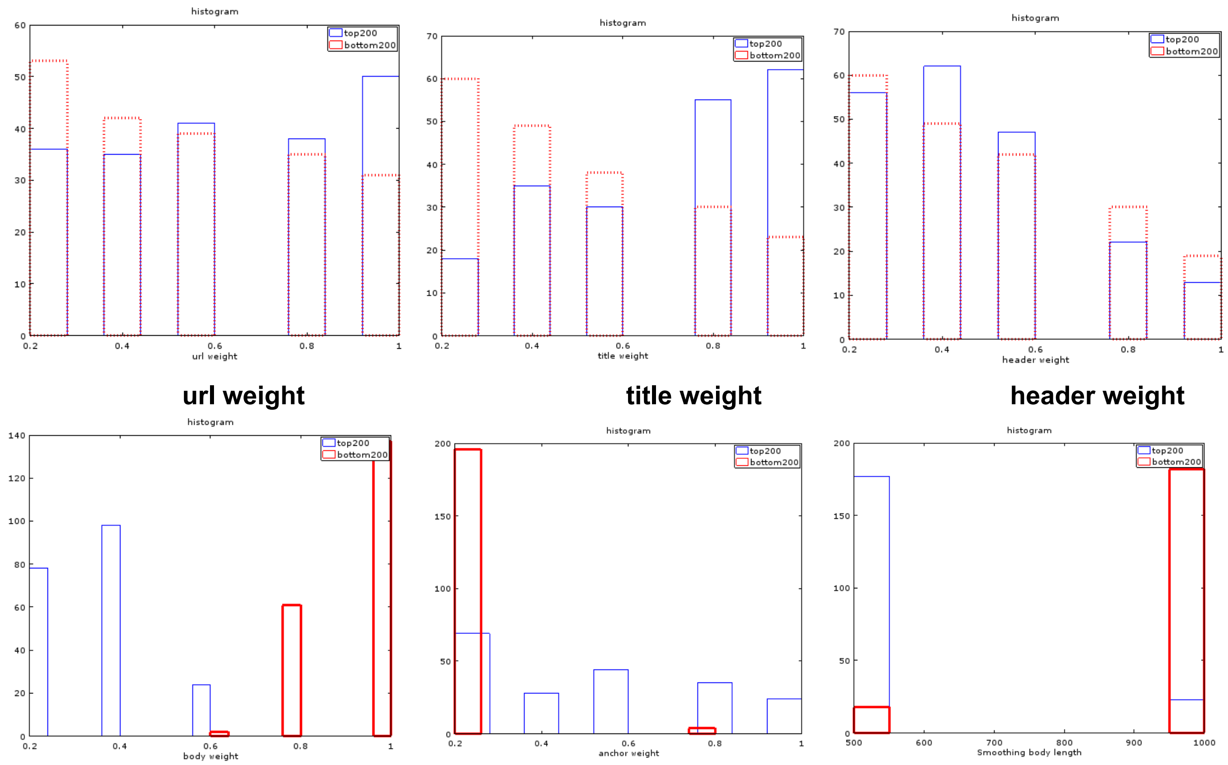


Figure 1. Histograms of top (blue) and bottom (red) 200 NDCG scores for cosine similarity scorer.



url weight

title weight

header weight

body weight

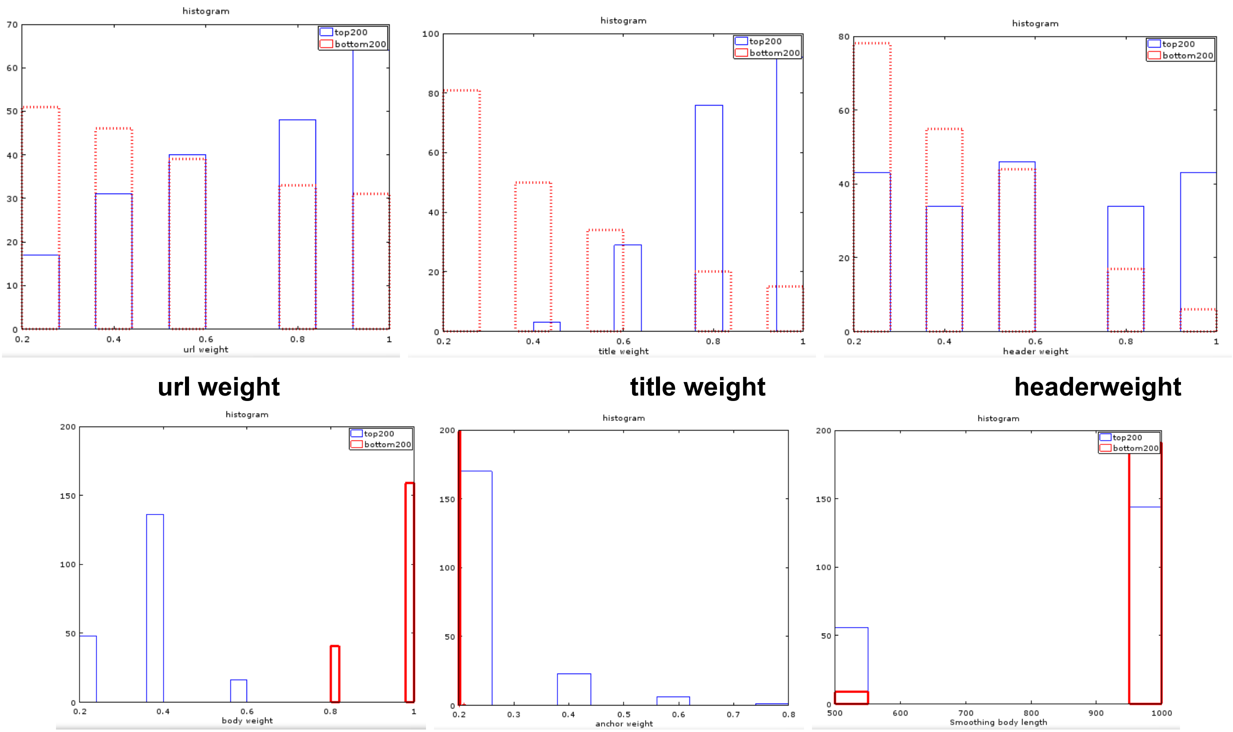
anchor weight

smoothing body length

Figure 2. Histograms of tuning parameters for cosine similarity scores corresponding to top (blue) and bottom (red) 200 NDCG scores.

Smallest Window on Cosine Similarity Scorer

For both exponential decay in boosting, we ran multigrid search with 31250 combinations, including the 6250 combinations mentioned earlier and with 5 possible values for boosting parameter task3_B. In figure 3, the histograms of the weight parameters follow very similar trend in the weights for high and low scores. We repeated the same experiment using 1/x decay of the boosting multiplier and figure 4 has similar results.



url weight

title weight

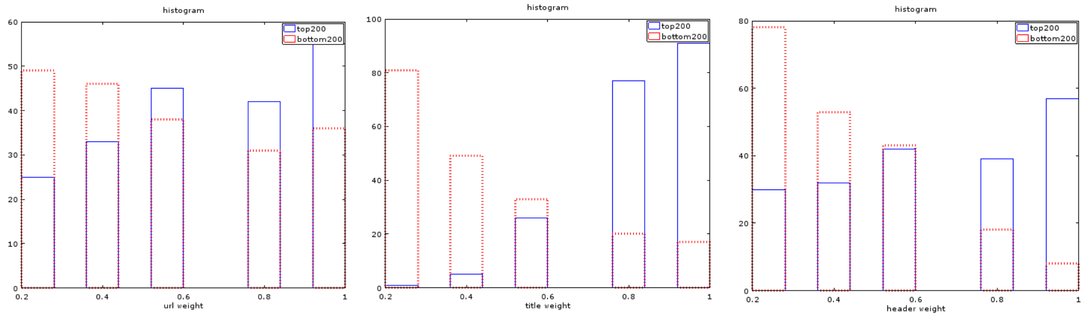
headerweight

body weight

anchor weight

smoothing body length

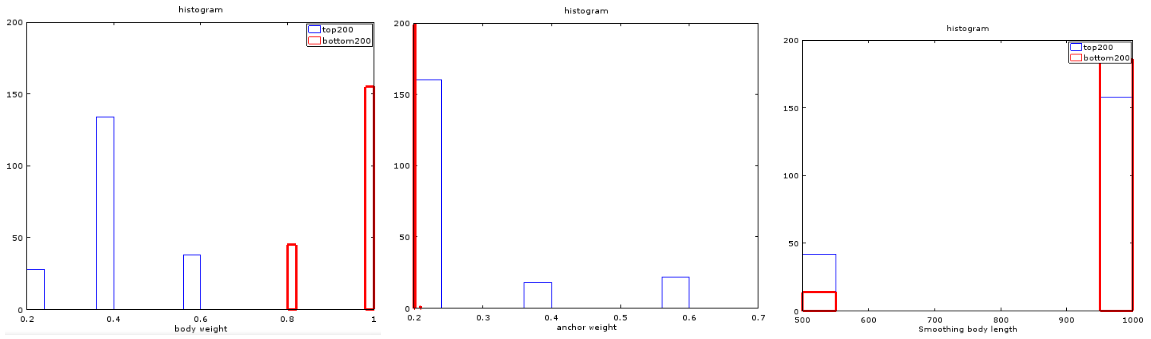
Figure 3. Histograms of tuning parameters for smallest window cosine similarity scorer (exponential decay) corresponding to top(blue) and bottom(red) 200 NDCG scores.



url weight

title weight

header weight



body weight **anchor weight** **smoothing body length**

Figure 4. Histograms of tuning parameters for smallest window cosine similarity scorer (1/xl decay) corresponding to top(blue) and bottom(red) 200 NDCG scores.

BM25 Weights Choice

Starting with task1_W_url; task1_W_title; task1_W_header; task1_W_body; task1_W_anchor weights from cosine similarity scorer analysis, we ran grid search on all 120 permutations of task2_B_url; task2_B_title; task2_B_header; task2_B_body; task2_B_anchor rankings. We then ran initial grid search with 0.1 increments for task2_k_1, task2_lambda, and task2_lambda' for a total of 11 x 11 x 11 = 1331 combinations. Further analysis, explained below, was done to optimize task2_k_1, task2_Vj(pagerank), task2_lambda, and task2_lambda'.

Particular properties of the documents:

a. **High quality:** Unlike the big web, Stanford web pages have high quality and satisfy a much higher standard than the average web page quality. These high standards automatically lead to higher pagerank, better structured pages, more reliable content with much less errors, better content, and more meaningful fields such as representative titles and headers.
b. **URL:** For the same reason, in particular the URL of these documents is of better quality than typical URLs. Stanford most likely has some guidelines/a policy for URL names to make them meaningful, categorize, and organize the content.
c. **Smaller scope:** The stanford collection has much less documents than the documents on the web. This smaller scale results in lower computational cost and most likely different system design options vs the web search problem. Additionally, the smaller defined scope impacts the accuracy the results. The spectrum covered by Stanford documents must be denser and more coherent than all the diverse documents on the web. Even users' queries on local search engine such as stanford would be automatically bounded to fit this scope.

Other Metrics

Query-independent

In this assignment, we use pagerank. Additionally the body length is involved in bm25 computation, also in the normalization step for cosine and window. Hubs and authorities scores are also related to pagerank.

Another metric is the quality of the page. For example, the number of misspelling errors is a good indicator of the quality, additionally, the page design and layout may fall under quality too.

Another source of information is context. Context can constrain the scope and help to retrieve more relevant pages. Context examples include: language and GPS.

Another metric is not only the relevance of the retrieved results but also their diversity, emphasizing the balance between exploitation and exploration. A good system should, at least, prune duplicate results.

Query-dependent

We used smallest window in this assignment and got a boost from 0.8506 to 0.8554.

We could also use bi-gram probabilities to calculate the probability of word pairs in queries, and for bi-grams with high probability, rank pages higher that have corresponding bi-grams, e.g. “computer science”.
Personalized pagerank (for specific terms) could be used to get a better scoring for terms that have special semantics or significance based on user behavior, e.g. “java”.
Trending queries could be used to rank results higher if there are specific events that are of interest to a significant portion of users. This could be an indication of “freshness” or something newsworthy (CNN, TMZ).

Tuning of BM25

In addition to the tuning of the weights for cosine: task1_W_url; task1_W_title; task1_W_header; task1_W_body; task1_W_anchor, we also tuned weights for task2_B_url; task2_B_title; task2_B_header; task2_B_body; task2_B_anchor. We performed grid search with relative field weights of 1.0, 0.8, 0.6, 0.4, and 0.2 for each of the 120 permutations of the 5 fields. Highest NDCG scores were found with the following weights: task2_B_title = 1.0; task2_B_url = 0.8; task2_B_anchor = 0.6; task2_B_body = 0.4; task2_B_header = 0.2;

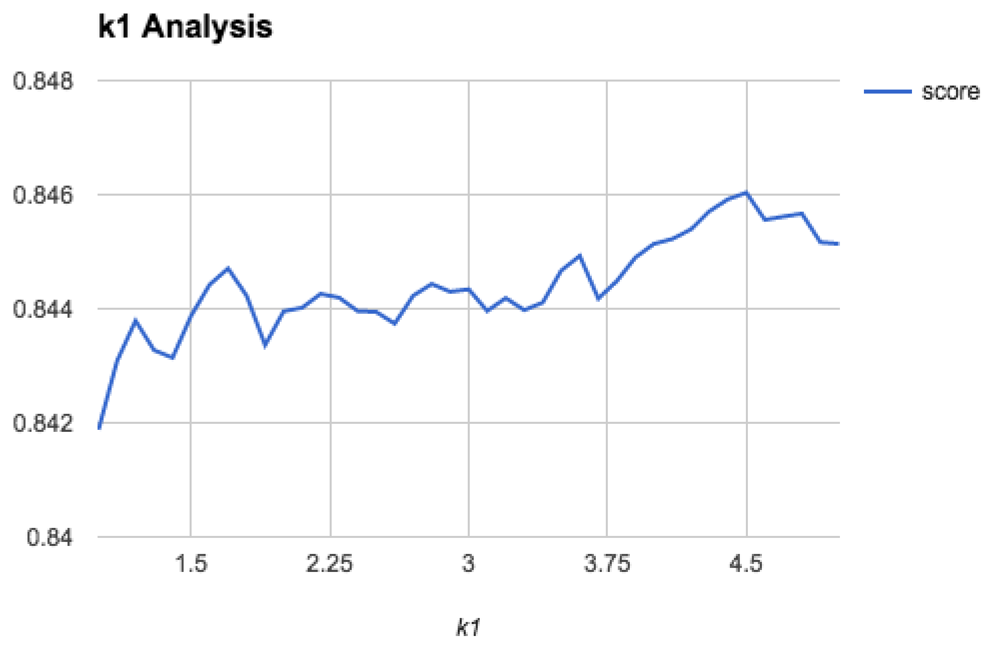


Figure 5. Tuning parameter task2_k_1 with task2_Vj = log(pagerank), task2_lambda = 1.0, and task2_lambda’ = 0.1

With weights constrained, initial testing with task2_Vj(pagerank) using the log function confirmed that an optimal task2_lambda = 1, and task2_lambda’ = 0.1 or ~0. We then tuned task2_k_1 for values between 0.0 and 5.0. We found a local maximum at task2_k_1 = 1.7 that produced the highest score until task2_k_1 = 3.6 and task2_k_1 = 3.9. This value of task2_k_1 = 1.7 was used for the subsequent testing of log, saturation, and sigmoid for task2_Vj(pagerank).

BM25F Tuning of task2_Vj Function

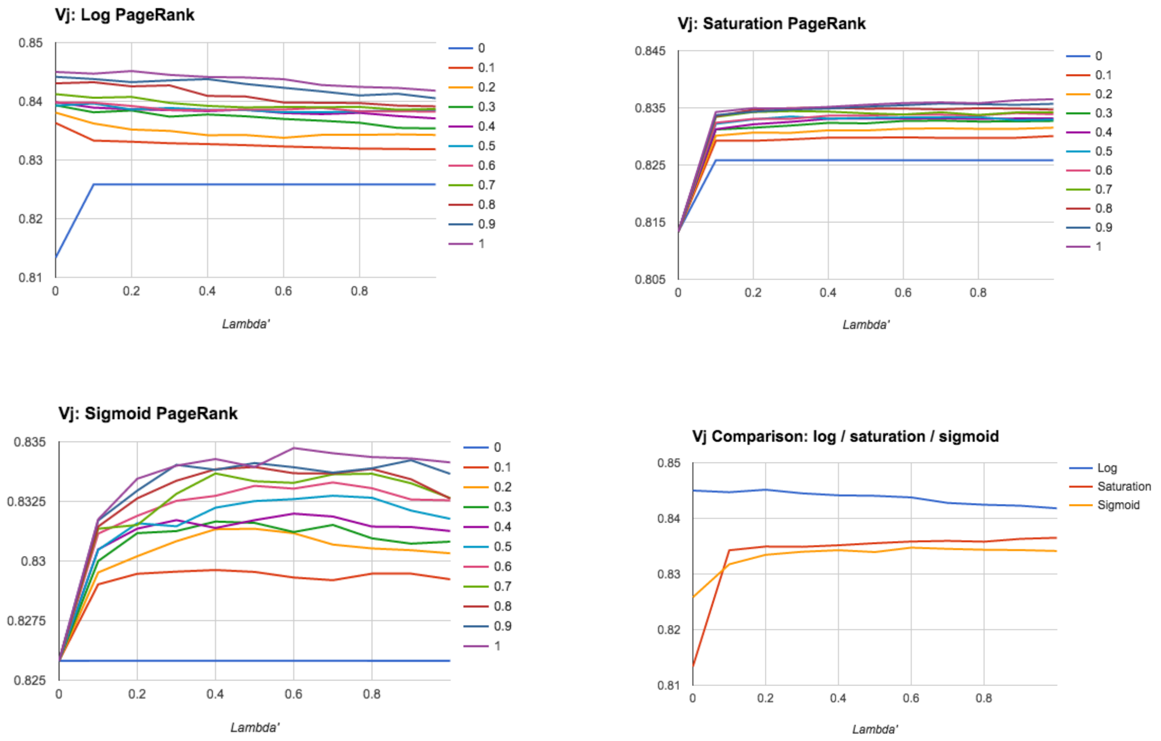


Figure 6. Tuning task2_Vj for log, saturation, or sigmoid of pagerank

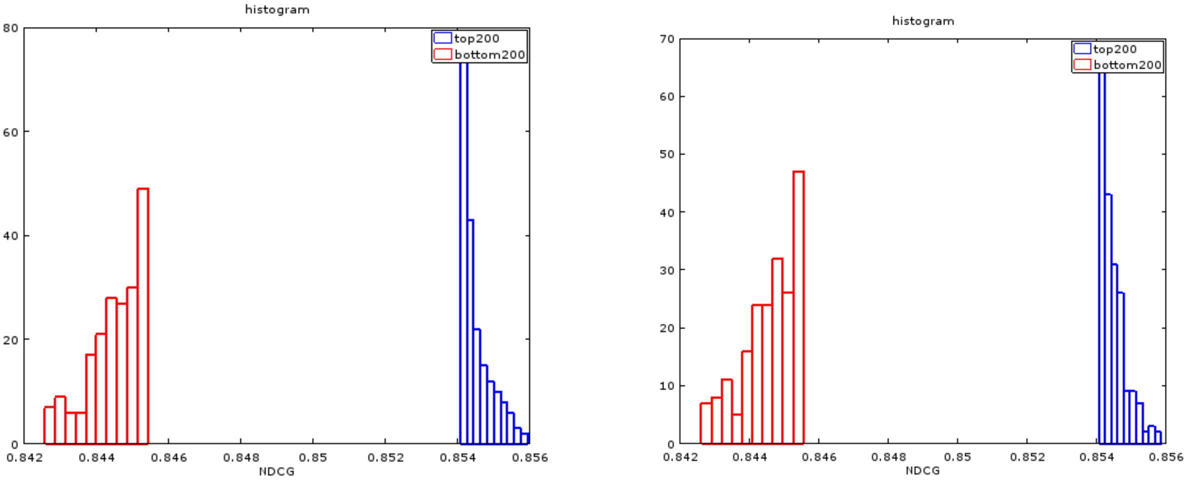
For values of task2_lambda between 0.0 and 1.0, and task2_lambda’ between 0.0 and 1.0 we tested task2_Vj(pagerank) for log, saturation, and sigmoid. For log, NDCG score is higher for task2_lambda = 1.0 for all corresponding values of task2_lambda’. Highest score for log pagerank was with task2_lambda = 1.0 and task2_lambda’ = 0.0. For saturation, NDCG score is mostly higher for task2_lambda = 1.0 for all values of task2_lambda’. Highest score for saturation pagerank was with task2_lambda = 1.0 and task2_lambda’ = 1.0. For sigmoid, Once again, NDCG score is maximized with a task2_lambda = 1.0. Highest score for sigmoid pagerank was with task2_lambda = 1.0 and task2_lambda’ = 0.6.

When comparing the top scores for log, saturation, and sigmoid functions, task2_lambda = 1.0 had the highest scores for all three functions. Within those functions, log outperforms saturation and sigmoid. The highest NCDJ score for task2_Vj(pagerank) is using log function with task2_lambda = 1.0 and task2_lambda’ = 0.0.

Smallest Window: exponential decay vs 1/x decay and boosting parameter B

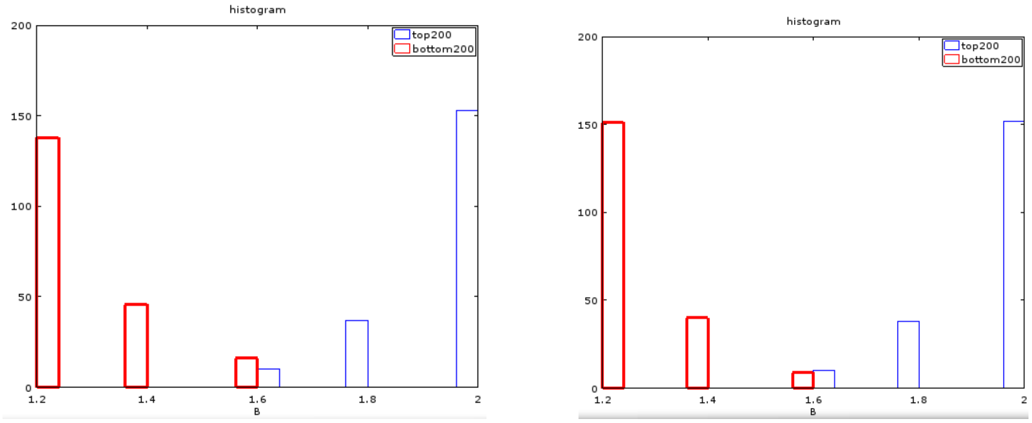
As mentioned in question 1, we ran experiments with 31250 combinations of parameters for smallest window cosine similarity scorer with both exponential and 1/x decay. The boosting parameter B could have 5 values in [1.2, 2.0]. As shown below, the NDCG scores for both exponential and 1/x decay have similar distributions.Although exponential decay has more parameters in the high score region. Similarly, for both decays, values of boosting parameter B close to 2.0 have higher scores. Values of B closer to 1 have lower scores. We eventually chose

task3_W_url = 0.8; task3_W_title = 0.8; task3_W_header = 0.6; task3_W_body = 0.2;
task3_W_anchor = 0.2 with task3_NDCG_train = 0.8554 and task3_NDCG_test = 0.8558



exponential decay **1/x decay**

Figure 7. Histograms of top (blue) and bottom (red) 200 NDCG scores for small window cosine



B boosting (exponential decay) **B boosting (1/x decay)**

Figure 8. Histograms of boosting parameter B for top (blue) and bottom (red) 200 NDCG scores