What’s new? Promoting diversity by re-ranking on hidden subtopics using LDA

Lilian de Jong
lilian.dejong@student.ru.nl
s1041699

Rosa van Ree
rosa.vanree@student.ru.nl
s4631889

Edwin Wenink
edwin.wenink@student.ru.nl
s4156072

Abstract

We propose to use a Latent Dirichlet Allocation (LDA) model to find subtopics in a collection of news articles in order to increase topical diversity. We evaluate our approach on the TREC Washington Post dataset using the nDCG metric. We do not improve the relevance scores of the baseline, but our approach does show a trade-off between diversity and relevance as well as the difficulties of assessing diversity.

Keywords: Bayesian learning, information retrieval, re-ranking, LDA, diversity

Code: github.com/EdwinWenink/bayesian-reranking

1 Introduction

With the rise of internet, search engines have grown to be the main method for gaining information online. Current discussions about filter bubbles and bias (for example the documentary The Social Dilemma) urge for innovative methods to improve diversity in search results of information retrieval systems. In this project, a diverse set of search results means that a ranking algorithm presents as many as possible aspects of the query. This is called topical diversity [7]. Because in TREC parlance *queries* are called *topics* we will speak of subtopic diversity. For example, searching for news articles about the 2020 United States presidential election should not only result in tweets of Donald Trump, but also in articles about the fly landing on Joe Biden, voting procedures etc.

The aim of this project was to promote diversity by re-ranking on hidden subtopics using Latent Dirichlet Allocation (LDA) in the domain of newspapers. Furthermore, we want to explore the trade-off between diversity and relevance.

Following Huang and Hu [4], we used Bayesian learning to promote diversity by re-ranking search results from an existing information retrieval system. We assume that for each topic we can distinguish an amount of hidden subtopics and then express the likelihood of each article given each hidden subtopic. If we then also define a prior distribution over these hidden topics, we can approximate the posterior distribution over topics for each document. Huang and Hu then use the maximum a posterior estimation (MAP) to pick the most likely subtopic for each document and group the documents that have their MAP in common. Some topics may not have any documents assigned to them, so this grouping process can decrease the amount of subtopics. By iteratively applying these steps until the document distribution over subtopics does not change anymore, Huang and Hu’s procedure is effectively also a way of dynamically finding the amount of subtopics. During re-ranking, articles are then picked from the different subtopics alternately to promote diversity in the ranking.

Huang and Hu apply their method to passage retrieval of biomedical information. We applied their proposed method on another domain and task: retrieval of newspaper articles. We did this because of two reasons. Firstly, we wanted to examine the effectiveness of their Bayesian learning approach onto another domain. Secondly, promoting different perspectives of news is important in the context of filter bubbles and selective exposure of information online [5]. By applying Huang and Hu’s approach to the domain of newspaper articles we aim to promote subtopical diversity and to contribute to a broader range of information presented in top ranked results by IR systems.

The remainder of this paper is organised as follows: section 2 contains a brief overview of related work, section 3 describes our methodology, section 4 contains our experimental setup, and we end this paper with results and discussion in section 5.

2 Related work

Search result diversification has been studied in a wide range of works. Maximal Marginal Relevance (MMR) is one of the first diversity aware ranking measures [1]. A document has a high marginal relevance if it is both relevant to the query and has minimal similarity to other retrieved documents.

MMR penalises redundancy to promote diversity. Our proposed method builds forth on this idea by re-ranking the search results such that documents that have different subtopics are prioritised over those that share the same subtopic, while maintaining the original relevance scores within each subtopic. In other words, if we find k subtopics, the top k re-ranked results consist of the most relevant document of each subtopic.

Another approach to promote diversity is with the use of Bayesian learning, like Huang and Hu [4]. Their method is explained in detail in section 1.

Diversifying search results is also possible with Latent Dirichlet Allocation (LDA), which is a form of probabilistic topic modelling that can discover the main themes in an unstructured collection of documents. Chen et al. used an LDA model to discover the subtopic distribution of passages from retrieved documents, and then re-ranked based on the distances between subtopic distributions [2]. They evaluated their approach to two baseline methods, which showed promising improvements.

Unfortunately, promoting diversity in search results by re-ranking does not necessarily imply a higher relevance score. GRASSHOPPPER is an example of a re-ranking algorithm, based on random walks in an absorbing Markov chain, that actually hurts relevance scores [9]. Zhu et al. show that GRASSHOPPER diversified search results, but consequently caused a drop in performance compared to the original result. With GRASSHOPPER, Zhu et al. illustrate that there is a trade-off between relevance and diversity for some ranking algorithms.

As stated in our introduction, we focused on subtopic diversity for this project, meaning that we promote a ranking that covers as many as possible aspects of the query. The diversity of a ranking can be evaluated with various metrics, that consider several aspects of what makes search results diverse [6]. Queries are often ambiguous and have different meanings or intents across different users. Ideally, a search engine returns results that reflect the ambiguity of a query by presenting a mix of results that address different subtopics of the query. Also, a search engine should accommodate for the diversity in users' information need.

One metric that is used to evaluate a ranking's diversity is α-nDCG, where the α parameter controls a trade-off between relevance, and diversity and novelty [3]. α-nDCG is based on the assumption that different users' information needs can be represented as a set of information nuggets. Because α-nDCG with higher α values rewards documents that express different information nuggets, this measure can be said to be intent-aware.

3 Methods

Our aim is to diversify the search results of a search engine through Bayesian learning of subtopics. We do this by re-ranking the results of a baseline method using an adjusted version of the algorithm proposed by Huang and Hu [4].

There are multiple differences between Huang and Hu's [4] and our approach that are relevant to mention. First, Huang and Hu [4] work with retrieved passages, whereas we retrieve entire news articles. Second, they use three indices based on three different passage extracting methods. Because we do not extract passages and because news articles are already relatively short, we instead use a single document-based index. This choice also required us to change our choice of the prior distribution over topics. Huang and Hu used a Poisson distribution which expresses the probability of some rare event given its *expected* counts of occurrence. In this case, the occurrence of a passage is interpreted as such a rare event, and a hidden subtopic is interpreted as the expected amount of a passages' occurrence, i.e. as a parameter in the Poisson distribution. But in the case of our single document-based index, it is not meaningful to count article occurrences, as an article appears at most one time in the ranking.

We therefore opted for another prior, namely the Dirichlet distribution, which expresses the prior probability of occurrence for a set of disjunct events (in our case the subtopics) and is a conjugate prior to the multinomial distribution. By using this Dirichlet prior instead, we essentially transform the Bayesian approach of Huang and Hu into a well-known topic model called Latent Dirichlet Allocation (LDA). Interestingly, in their conclusions Huang and Hu already mention they want to extend their approach to use PLSA, of which LDA is an extension [4].

3.1 Preprocessing and index

We performed several common pre-processing steps. We used Lucene's Analyzer through Anserini[1], which includes tokenization, stemming, and stop word removal. To create our re-ranking, we represented news articles as bag of words. Our index is a standard positional index that stores document vectors and raw HTML documents and was created using Anserini.

3.2 Baseline ranking

We used Pyserini[2], the Python interface of Anserini, to obtain our baseline ranking. To compute our baseline ranking we used Anserini's implementation of the BM25 retrieval function.

3.3 LDA

In this project we used Latent Dirichlet Allocation (LDA) to find subtopics in a collection of news articles. Like most probabilistic topic models, LDA defines a hidden subtopic as a word distribution over the vocabulary of a corpus. This allows the model to describe complex topics and capture subtle semantic differences between them. This approach also naturally facilitates disambiguation [8]. Zhai and Massung give the example of the word "star", which means something else in the context of sports than in science. In both the "sports" and "science" topic, the word "star" may be assigned high likelihood, but when the words "football" and "play" also occur in the same document, these words combined will have a higher likelihood $p(d = w_1, ..., w_n|\theta_i)$ under the "sports" than under the "science" topic. LDA computes a topic coverage distribution $p(\theta|d)$ for each document d (i.e. a posterior distribution), so in this example the document is better covered by the "sports" topic. LDA is thus a mixture

[1] https://github.com/castorini/anserini
[2] https://github.com/castorini/pyserini

model where each document is represented as a weighted mixture of topics. This feature makes LDA more appropriate for application to whole documents because each document may address various subtopics. Conversely, Huang and Hu's [4] method was designed for passages that can be expected to be more specific than a whole document. Nevertheless our approach is consistent with Huang and Hu's general procedure, because we can still group documents by the subtopic that *best* describes a document.

3.4 Re-ranking algorithm

We implemented an adaptation of the re-ranking algorithm by Huang and Hu [4] for our method and dataset. A description of our algorithm is presented below.

0. Input
The ranking of a BM25 model, which returns for each TREC topic relevance scores for N documents.
1. Initialisation
Initialise k subsets with $min(20, N)$. Hu and Huang initialise with $k = N$, but we noticed that the model always converged on $k < 20$, so we initialise with fewer subtopics to speed up the algorithm's convergence. Initialise the LDA model with a symmetric Dirichlet prior over topics.
2. Iterate
Find k subtopics, group each document d under the MAP topic $argmax_{\theta_j}(p(\theta_j|d))$, and iterate until the distribution of document assignments no longer changes. Each iteration, set k to the amount of subtopics after grouping and use the previous topic distribution as the Dirichlet prior over topics.
3. Re-ranking
When the iteration process converges, each document is assigned to a subtopic. We re-rank the results based on these assignments and their initial relevance judgements using one of two ranking strategies: GREEDY or TOP-K-AVG (3.5).
4. Output
The output is a re-ranked list of documents for each topic.

3.5 Re-ranking strategies

When the iteration process converges, each document is assigned to a subtopic. We create the new ranking by alternately taking the most relevant document from each subtopic until all documents are included in the new ranking, following the approach from Huang and Hu [4]. The order in which we iterate over the subtopics is decided using one of two re-ranking strategies, that both use the original relevance scores of the documents in a subtopic.

The first strategy is a "greedy" method that sorts the subtopics based on their highest scoring document. This way the order of the most relevant documents is largely preserved, but when for example the two best scoring documents belong to the same subtopic, the second most relevant document in the original ranking re-appears only after the most relevant documents of other subtopics have been listed.

In the second strategy we instead determine the topic picking order by sorting on the average of the top k most relevant results of each subtopic.

3.6 Evaluation

To evaluate our results, we make use of the normalized discounted cumulative gain metric (nDCG). This measure is appropriate because the TREC Common Core 18 Track provides graded relevance judgements on the Washington Post dataset. We evaluate at various cut-offs, because we care most about diversity in the top ranked results.

α-nDCG is an evaluation metric based on nDCG, that rewards novelty and diversity. Unfortunately, the use of the α-nDCG was not appropriate for our dataset. We will further elaborate on this in section 5.2.

4 Experimental setup

4.1 Dataset

We use the Washington Post dataset[3], for which graded relevance judgements on 50 topics are available in the TREC Common Core 2018 Track. The Washington Post dataset contains 608,180 news articles and blog posts from January 2012 through August 2017. Examples of topics include *sony cyberattack*, *eggs in a healthy diet*, and *car hacking*.

4.2 Experimental conditions

We apply our method using two re-ranking strategies (GREEDY and TOP-5-AVG) and evaluate using nDCG@5, @10 and @20. Experiments with multiple values of k for the TOP-K-AVG strategy returned very similar results, so we only report on $k = 5$ for brevity. Because we care most about the top-ranked results, we initially rank $N = 100$ documents. LDA uses random initial topic assignments, so for reproducibility and comparability all runs use the same seed.

5 Results and discussion

5.1 Results

We evaluated two re-ranking strategies (GREEDY and TOP-5-AVG) using nDCG@5, @10, and @20, meaning only the top 5, 10 or 20 search results were considered when evaluating (Figure 1). Figure 1 shows that both strategies score significantly lower on relevance than the baseline, consistently across different cut-off values (p < .05 for GREEDY@5, and p < .01 for the other conditions, t-values can be found in Figures 2-7).

We can examine the performance of GREEDY and TOP-5-AVG in more detail by comparing the scores per individual topic. For example, Figures 10 and 7 show the performance of GREEDY per topic for nDCG@5 and nDCG@20. We see that for GREEDY@5, the biggest gain is 0.2042 and the biggest loss -0.4468. For GREEDY@20, the biggest gain is 0.1025 and

[3] https://trec.nist.gov/data/wapost/

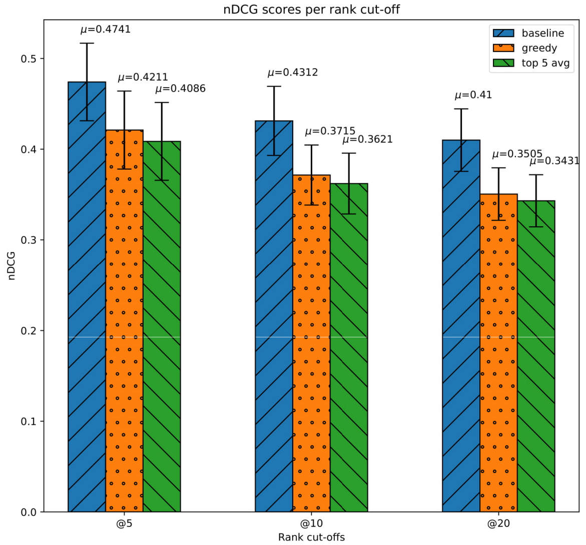

Figure 1. nDCG@5, @10, and @20 over all topics for GREEDY and TOP-5-AVG strategy compared to the BM25 baseline.

the biggest loss -0.3864. We see that even though GREEDY@5 has a higher average score, both the highest gain and loss are larger. We however see why the average of GREEDY@20 is lower, because only 9 topics have an increased score, in comparison to 16 improved topics for GREEDY@5. We see this pattern in the other runs as well. The plots with scores per topic for the other runs can be found in Appendix A.

5.2 Discussion

We observed large score differences between topics. But why does the relevance score improve for some queries and not for others? Recall that we tried to increase the subtopical diversity in a ranking, meaning that multiple different intentions of a query can be found in the top search results. The evaluation topics provided by the TREC Common Core 18 Track consisted of ad hoc queries, possibly multi-intentional, accompanied by specific, single-intent descriptions. This means that the relevance of documents in the dataset are judged based on only one possible intention of a query. When diversifying a ranking using hidden subtopics, there is a chance that more documents judged as relevant were pushed into the top k. Yet, there is also a high chance that we push documents to the top that, even though they do reflect different aspects of the query, are marked as irrelevant due to the very specific intent-descriptions of the TREC topics. This means that diversity is not necessarily rewarded in terms of relevance and that there may be a relevance-diversity trade-off.

5.2.1 Manual analysis. To support these intuitions, we manually analysed two subtopics that had a rise or drop in score compared to the baseline: topic 626 *human stampede* and topic 445 *women clergy*.

The top five results of the baseline@5 for *human stampede* included only two relevant articles, whereas GREEDY@5 scores higher because all top five results are considered highly relevant. The hidden subtopics found by the algorithm must have been different for them to appear in the top 5. For example, one article discusses how a stampede could have turned deadly so easily, whereas another document discusses a stampede in the political and quite different context of a U.N. meeting. Even though the subtopics are apparently diverse, four of the five articles do mention the same specific stampede in Mecca, which shows that it is quite hard to interpret the meaning of the found subtopics.

GREEDY@5 performed worse on the topic *women clergy* compared to the baseline. Just like the GREEDY@5 re-ranking for topic *human stampede*, the new top articles were about different subtopics and the ranking seemed more diverse. However, it seems that the articles pushed to the top were irrelevant for a user searching for information about women clergy. Because the intent of this query is very specific, it seems that diversity is actually not desirable here and hurts relevance.

Our model seems to effectively incorporate diversity into the top results of a ranking, but whether these results are relevant with respect to the intended interpretation of the query can go either way. This could explain why some queries obtain a gain in relevance and other queries cause a drop. To properly investigate the trade-off between relevance and diversity, we either need to do a user study, or we need a data set with relevant judgements on different aspects of the same query.

5.2.2 nDCG vs. α-nDCG. The nDCG metric does not by itself reward diversity, and in fact we have seen that increasing diversity can hurt the nDCG relevance score. This makes it harder to improve both the ranking's relevance and diversity.

α-nDCG instead assumes that a query can have multiple possible intents and rewards rankings with less redundancy in document's particular information nuggets. α-nDCG thus rewards diversity, but does require that relevance assessments for each query intent is available. Although the TREC Common Core 18 Track includes queries with several aspects, it provides relevance judgements based on only one specific intention per query. As a result, α-nDCG scores are not appropriate, even though we can still compute them by acting as if each topic is its own single subtopic (see appendix B).

5.2.3 Impact. Promoting diversity by re-ranking is important, because search engine results should reflect the ambiguity of the query and accommodate for the diversity in users' information need. We have tried to increase diversity, but at what cost? A challenge for future work is to find better methods for navigating the trade-off between relevance and diversity.

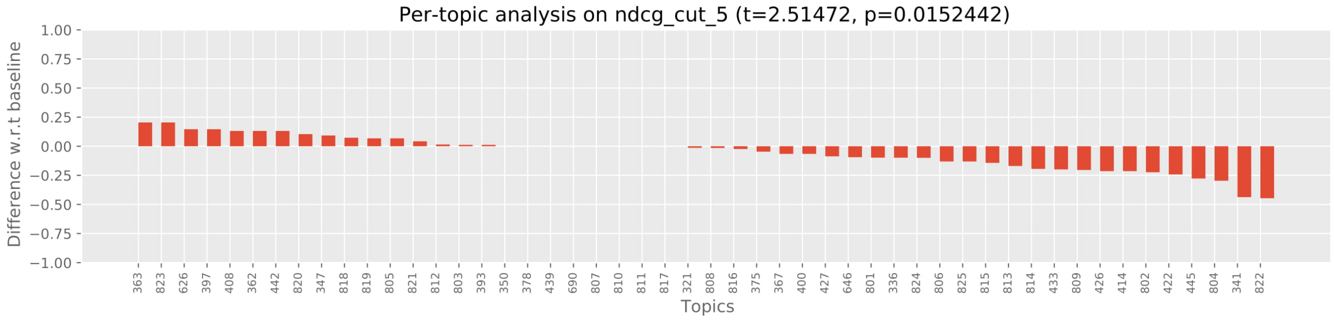

Figure 2. The performance of GREEDY per topic for nDCG@5.

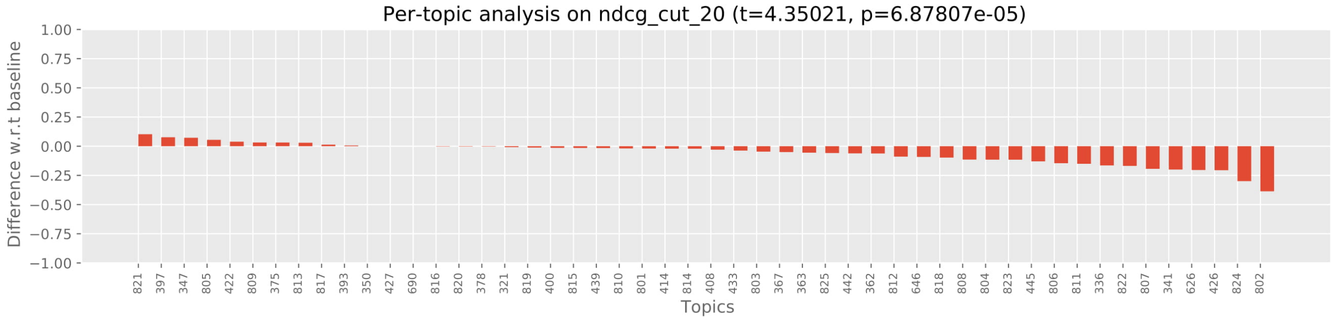

Figure 3. The performance of GREEDY per topic for nDCG@20.

References

[1] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR*. 335–336.

[2] Yan Chen, Xiaoshi Yin, Zhoujun Li, Xiaohua Hu, and Jimmy Xiangji Huang. 2012. A LDA-based approach to promoting ranking diversity for genomics information retrieval. *BMC Genomics* 13, 3 (2012), 1–10.

[3] Charles L. A. Clarke, Maheedhar Koll, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Buttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of SIGIR*. 659–666.

[4] Xiangji Huang and Qinmin Hu. 2009. A Bayesian Learning Approach to Promoting Diversity in Ranking for Biomedical Information Retrieval. In *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*.

[5] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin, New York, NY.

[6] Tetsuya Sakai, Nick Craswell, Ruihua Song, Stephen Robertson, Zhicheng Dou, and Chin-Yew Lin. 2010. Simple evaluation metrics for diversified search results. In *Proceedings of the 3rd International Workshop on Evaluating Information Access (EVIA)*.

[7] ChengXiang Zhai, William W. Cohen, and John D. Lafferty. 2003. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 10–17.

[8] ChengXiang Zhai and Sean Massung. 2016. *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Association for Computing Machinery and Morgan Claypool, Chapter 17.

[9] Xiaojin Zhu, Andrew B. Goldberg, Jurgen Van Gael, and David Andrzejewski. 2007. Improving Diversity in Ranking Using Absorbing Random Walks. In *Proceedings of NAACL HTL*. 97–104.

A nDCG scores per topic

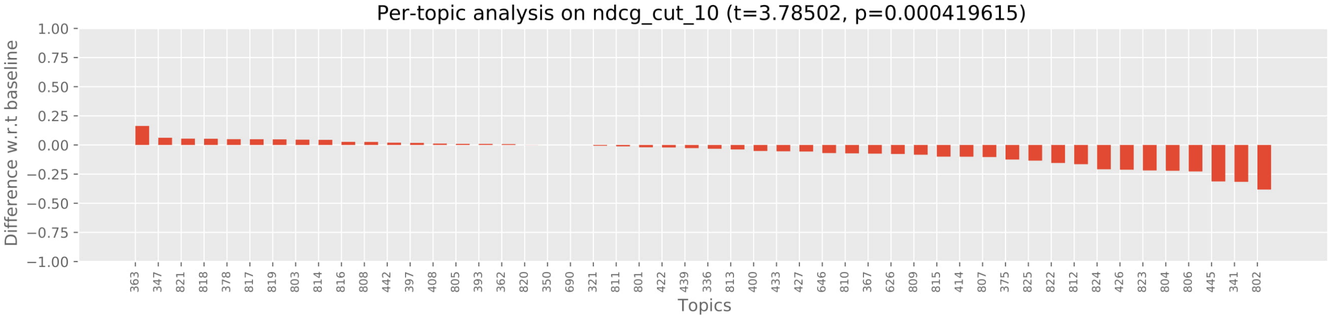

Figure 4. The performance of GREEDY per topic for nDCG@10.

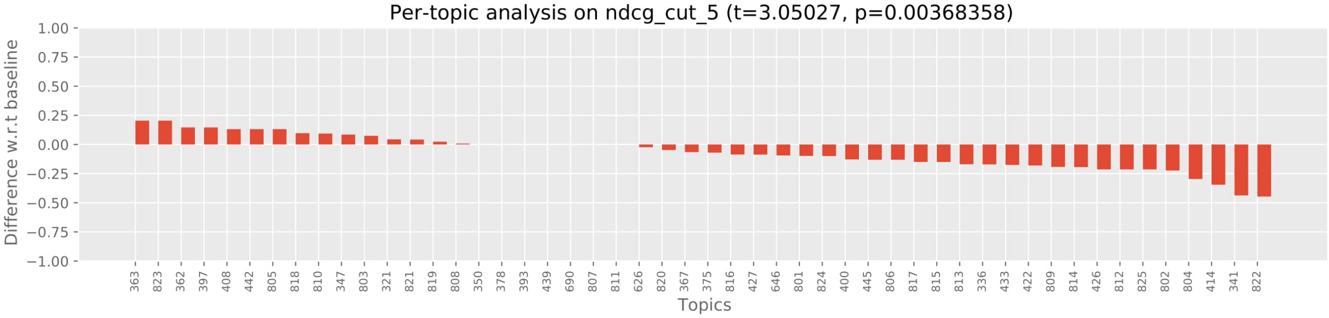

Figure 5. The performance of TOP-5-AVG per topic for nDCG@5.

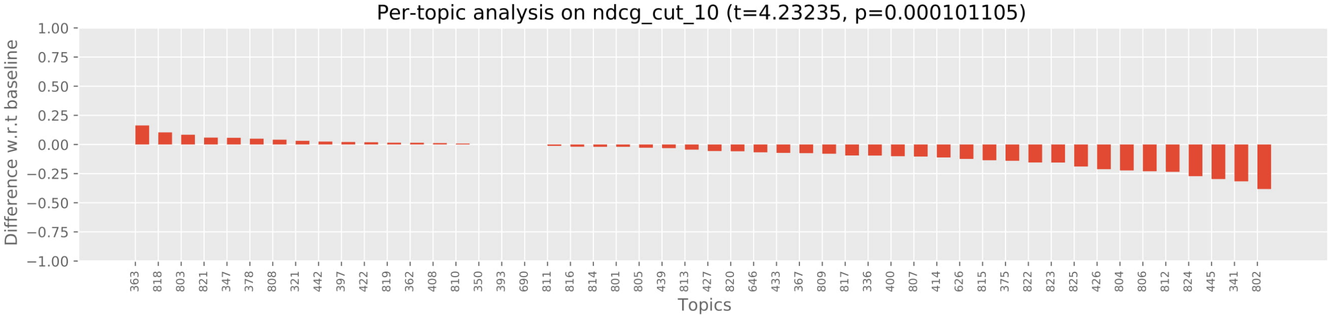

Figure 6. The performance of TOP-5-AVG per topic for nDCG@10.

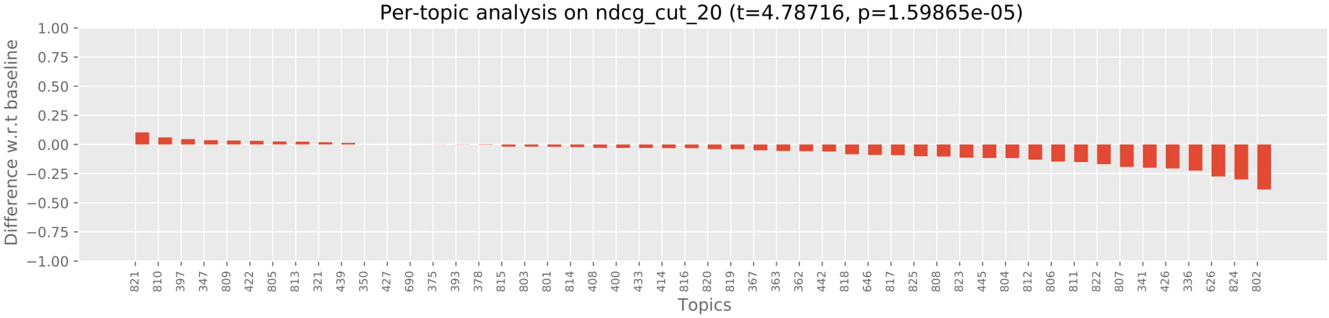

Figure 7. The performance of TOP-5-AVG per topic for nDCG@20.

B α-nDCG plots

This appendix contains plots that show the trajectory of α-nDCG for different values of α. We included this for completeness, but note these results are not appropriately balancing relevancy and diversity for the reasons mentioned earlier.

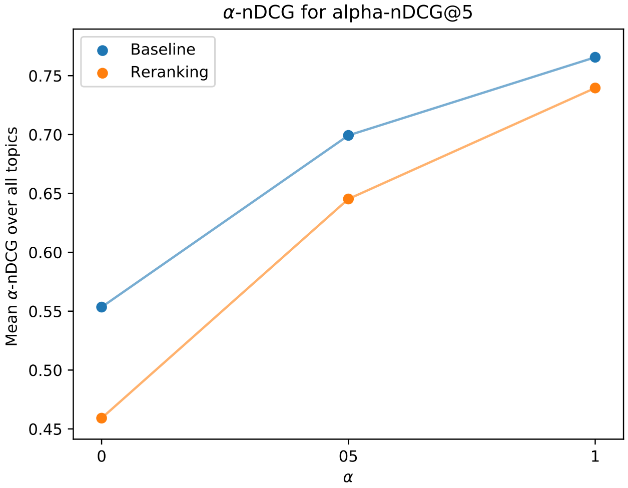

Figure 8. α-nDCG@5 over all topics for GREEDY and baseline, calculated for different values of α.

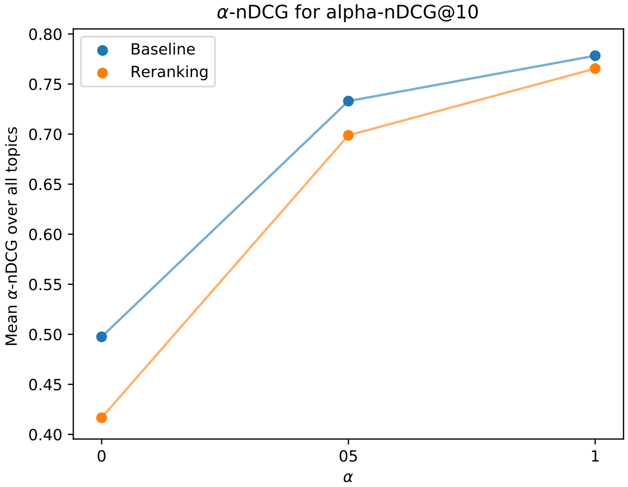

Figure 9. α-nDCG@10 over all topics for GREEDY and baseline, calculated for different values of α.

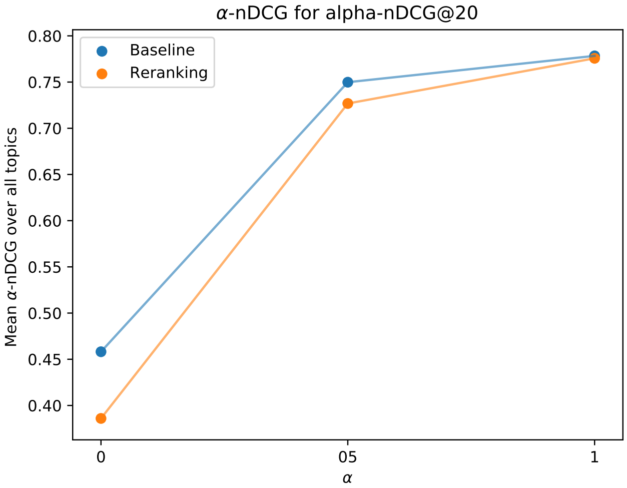

Figure 10. α-nDCG@20 over all topics for GREEDY and baseline, calculated for different values of α.