UNICAMP-IR
Applying LLMs to build a native pt-BR Information Retrieval dataset

Eduardo Seiti de Oliveira, Leodécio Braz da Silva Segundo

July 2023

Abstract

Building an Information Retrieval (IR) dataset is always expensive and time-consuming, and the documents annotation limited depth affects the dataset capacity to evaluate future IR systems. To attenuate those problems, we explore Large Language Models (LLMs) for key steps in a semi-automated documents annotation pipeline, leveraging their ability to evaluate query annotation, as well as query proposal. We use the Clueweb22 dataset corpus, as it provides thousands of curated Brazilian Portuguese (pt-BR) web pages, and execute two different LLM-based pipelines for creating a validation and training splits of a new IR dataset built entirely from pt-BR original content, making it more suitable for evaluating IR systems for that language.

1 Introduction

Dataset annotation part is an expensive and time-consuming task, and for textual Information Retrieval (IR) the final result might suffer from problems which limits its capacity to evaluate new retrieval systems [2]. The depth of documents annotation is directly limited by the budget available, and for larger corpus that limitation can impact the dataset evaluation capacity for new IR systems, as they become more capable of extract meaning and find actual relevance in non-annotated documents.

The advances in language translation techniques allows building IR datasets for new languages, which can successfully used to evaluate IR systems focused on corpus on those languages [1, 3]. However, despite their ability to evaluate language comprehension, translated datasets fall short in providing regional context for the IR systems, which might end up hindering their capacity of handling live, real-world content, specially for embedding-based systems, which capture both the language structure and meaning.

To engaged those aspects of IR dataset creation, we propose o apply Large Language Models (LLMs) to perform key steps of the dataset creation pipeline, starting from a large corpus of documents originally written in Brazilian Portuguese (pt-BR) language. Our goal is to build a new IR dataset — UNICAMP-IR — focused to evaluate IR systems operating primarily over pt-BR documents and queries, proposing a less-expensive pipeline.

All the source code used for this work is publicly available[1].

2 Methodology

The proposed pipeline for applying LLMs for IR dataset creation is slightly different from validation and training splits. That because powerful LLMs have a non-negligible usage cost — even though smaller than applying human annotators — and training dataset is expected to much larger, and might contain noisier annotations, when compared to validation dataset.

The figure 1 depicts the overall working pipeline.

2.1 Validation split creation

The validation split is composed by the following steps:

1. Manual queries creation;
2. Passages retrieval for the created queries, applying different retrieval systems;
3. LLM usage to evaluate the relevance of the passages returned to each query, in order to build the final annotation.

Human created queries can have a high quality, if the proper instructions are given. For this task, we asked the people to create questions considering an ideal retrieval system, guiding them using a taxonomy of themes, and questions characteristics:

- Themes: **Geography, Politics, Economy, Culture, Culinary, Tourism, Leisure, Sports**.
- Scope: **General** (exploring a broad theme or subject) or **Specific** (exploring a narrow theme or subject).
- Type: **Opinion** (asking for an opinion about something) or **Factual** (asking for a fact or data which has little dependency on one's opinion)

For this research stage, our goal is to create a whole total of 100 validation queries, combining the results of every person.

The first step of the queries annotation is applying the retrieval systems to return the relevant passages. They resemble the human annotators as they provide a relevance score, which the expectation that the more documents returned, less relevant documents will also be included. We apply multiple IR

[1] `https://github.com/eduseiti/ia368v_dd_final`

systems to mimic human evaluators variability, as well as to try capturing most diverse documents from the corpus. For this step we will apply the following IR systems:

- **BM25**: as strong baseline for retrieval: we use the BM25 Pyserini implementation [5]. We just change the language to Portuguese on the default settings. At inference time, was necessary to declare an analyzer (Lucene Analyser) for Portuguese ('pt') and we retrieve the top 1000 documents per query.
- **BM25 + mT5**: a two-stage pipeline with BM25, as first stage, followed by mT5 for rerank.
- **ColBERT-X** [6], a multilingual ColBERT-v1, as a state-of-the-art dense retrieval.

The final step of the queries annotation is to apply a LLM to evaluate the retrieved passages relevance for each query. We leverage the LLM text comprehension capacity and broad knowledge to not only check if the passage is relevant or not to answer to the proposed question, but more specifically to request the LLM to provide a relevance score of the passage given the query. In this sense, returning non-relevant documents during the first step is still valuable, and applying a variety of IR systems is valuable.

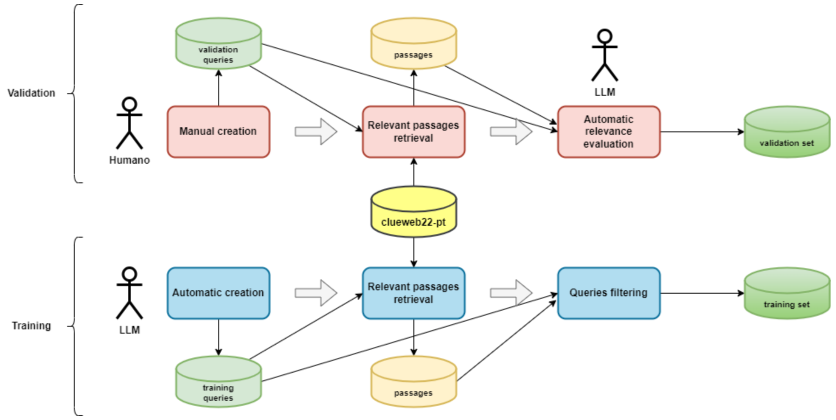

Figure 1: Proposed IR dataset creation pipeline.

2.2 Training split creation

For the training split creation, the pipeline follows the same sequence, but the steps changes to allow creating a larger set of annotated queries under a reduced cost:

1. Automatic queries creation using LLM;

2. Relevant documents retrieval, using a state-of-the-art retrieval;
3. Low-quality queries filtering.

A LLM is applied to create a large number of queries, using a prompt asking to create a query for a given passage, and sampling several passages from the corpus.

Then, the same ColBERT-X system is used to retrieve the relevant documents for each synthetic query, and the returned passages can be used as a basic quality check: the sampled passage, passed to the LLM for the query creation, must have returned within the first 5 best-scored ones; if the passage fails to be well evaluated, it is a strong indicator of low-quality question, as it fails to provide clear enough information to retrieve the original passage.

For this research stage we aim to create a train split containing at least 1000 questions, after the quality assurance filtering.

2.3 Final dataset evaluation

Once the training and validation splits are created, the new dataset evaluation can be done, under the assumption that retrieval models trained using this new dataset should perform better in pt-BR retrieval tasks.

Hence, we will fine-tune a ColBERT-X retrieval using the new dataset and compare the final nDCG@10 on the validation split, against the performance ColBERT-X baseline. We will also do the same comparison against the mRobst04-pt dataset [3]. As a way of verifying the overall dataset quality, we will check the judge@20 metric for the evaluation done over our new dataset.

3 Dataset

The Clueweb22 dataset general category [8] contains a total of 4 million web pages in pt-BR which can be used as the base corpus for the new IR dataset. However, this initial set is very noisy, containing some advertising and marketplaces webpages which produced text-only documents with very few information. When applying document segmentation, to produce passages suitable for IR pipeline steps using Neural Networks-based Language Models — which have in its vast majority, limited input string length capacity — the final number of passages increases 10x.

Considering we select Colbert-v1 as our state-of-the-art dense retriever, the required memory to build the corpus index is an important bottleneck for the experiments, as Colbert-v1 is extremely memory intensive.

To make the experiments feasible under reasonable memory budget (Google Colab's [2] A100 VM of 83.5GB RAM, 40GB GPU, and 166.8GB of disk) we applied the following dataset cleanup and sampling steps:

[2] https://colab.research.google.com

1. Removal of documents with more the 20% of new-line characters, considering that an indicator of text fragmentation and low-content webpages;

2. Segmentation applying a Window aiming a passage of 480 tokens, considering the ColBERT-X base Language Model is a cross-encoder, with the limitation of 512 tokens input.

3. Replace of <`tab`>, <`new-line`>, and <`carriage-return`> characters by space character, as those characters affects the ColBERT-X scripts execution.

4. Sampling of 10 million of the resulting passages, in order to have a manageable, but still representative, corpus size.

4 Experiments

4.1 Retrieval systems to return the relevant passages

For this step we follow a pipeline quite similar to that adopted by [1]. First we retrieve a ranked list of passages using BM25 [9, 4] with queries as input. Further, we rerank the returned list of passages using a multilingual pretrained[3] mT5 [7]. We do not fine-tuned the mT5 model.

4.2 Fine-tuning Colbert-X for pt-BR

We fine-tuned the ColBERT-X retriever applying the Translate-Train approach [6] using the provided scripts and the pt-BR translated version of the MS-MARCO passage dataset [1]. We trained the original `xlm-roberta-large`[4] model during 20K steps, passing through 204k mMARCO triplets. That training took 1h30 using a Google Colab A100 VM.

In order to limit the final memory footprint of the ColBERT indexes, we considered only 48-dimension embeddings — reducing from the original 128-dimension setting.

4.3 Creating the passage evaluation prompt

We applied few-shot learning prompt to instruct the LLM to evaluate the passage relevance given a query. Inspired by Faggioli et al. [2], we instructed the LLM to produce a relevance score along with an explanation, but we asked for a 0 to 10 score, aiming a finer granularity.

We evaluated the performance of 3 different LLMs provided by OpenAI: `text-davinci-003`, `gpt-3.5-turbo`, and `gpt-4`.

The final prompt is presented in InfoBox 1.

As expected, the `gpt-4` model performed better to execute the passages relevance task being able to generate better scores, in a simple human qualitative

[3] The pre-trained model used is on: `https://huggingface.co/unicamp-dl/ptt5-base-pt-msmarco-100k-v2`

[4] `https://huggingface.co/xlm-roberta-large`

Você avalia se uma passagem de texto responde a uma pergunta, indicando uma pontuação de 0 à 10, onde 0 indica que a passagem não responde e 10 que a passagem responde de forma correta e clara. Você desconsidera informações que o texto diz que vai apresentar mas não apresenta. Siga os exemplos abaixo.

Exemplo 1:
Passagem: "O cirurgião faz uma incisão no quadril, remove a articulação do quadril danificada e a substitui por uma articulação artificial que é uma liga metálica ou, em alguns casos, cerâmica. A cirurgia geralmente leva cerca de 60 a 90 minutos para ser concluída." Pergunta: de que metal são feitas as próteses de quadril?"

Pontuação: 2; Razão: não responde a pergunta de forma clara, pois apenas indica indiretamente que a prótese pode ser de uma liga metálica, mas não explicita quais metais. O assunto da passagem é sobre cirurgia de colocação de prótese que, embora relacionado, não é diretamente o assunto da pergunta.

Exemplo 2:
Passagem: "O Brasil possui muitas belezas naturais. Neste artigo vamos indicar os melhores lugares para passear no Brasil." Pergunta: "Onde passear no Brasil?"

Pontuação: 1; Razão: a passagem apenas indica que o Brasil tem muitas belezas naturais, mas não indica nenhum exemplo. Embora a passagem indique que artigo vai falar sobre lugares para passear no Brasil, o trecho apresentado não lista nenhum lugar específico para passear no Brasil.

InfoBox 1: LLM prompt for passage relevance evaluation.

evaluation of 5-question subset. The most common evaluation error was the LLM failure to identify, and score, when the passage answered poorly to the question, due to incomplete information, or when the passage indicated the question would be answered by the whole webpage, but the answer itself was not in that particular webpage segment — the passage did not contained the actual answer.

4.4 Creating the query generation prompt

For the query generation prompt, to instruct the LLM to generate queries for sampled queries, we also applied one-shot learning, to illustrate the task and the desired output format. We tried to guide the query generation using the **scope** characteristic we applied for the validation queries manual creation, asking the LLM to generate one question for the passage theme, and another question related a specific information or conclusion within the passage. The final prompt applied is presented in InfoBox 2.

On `gpt-3.5-turbo` model, that prompt produced reasonable results, with the major problems related to sometimes the LLM generated questions which only made sense after reading the passage first, despite the fact the prompt explicitly asked the model to avoid that

4.5 Generating synthetic queries

Given the query generation prompt, we decided to sample 1000 passages from the corpus and feed them to the LLM. As an attempt to enhance the queries quality, we first filtered out all the passages from documents with more than 3% of new-line characters; that reduced to 8.5 million the number of available passages.

LLM processing of the 1000 passages took 37 minutes, and cost U$ 1.25.

The resulting 2000 queries were filtered using the retrieval approach previously described — keep only the question for which ColBERT-X retriever return the original passage (the passage sent to LLM for the query creation) within the best 5. Around 50% (1080) of the generated queries remained. The proposed filtering process was able to remove low quality queries, as the examples in InfoBox 3 can show.

4.6 Verifying the LLM evaluations

As a sanity check for the validation split pipeline, we selected only the first returned passage by each retrieval system for the manually created validation queries, to execute the LLM relevance check, as `|gpt-4` usage is still expensive, and each relevance check cost around U$ 0.027. This way we performed the LLM verification for a total of 300 query/passage scored combinations, considering 1 passage returned for each question (100) for each retrieval system (3); the cost for that verification round was U$ 7.92 and it took 19 minutes.

Você sugere 2 perguntas a partir da leitura de uma passagem de texto. A primeira pergunta explora o tema da passagem, e a segunda pergunta explora uma informação ou conclusão específica possível a partir da leitura da passagem. Suas perguntas devem fazer sentido para alguém que não leu a passagem. Siga o formato do exemplo:

Exemplo : Passagem: "Como acontece com todos os tratamentos naturais, a qualidade do produto utilizado para o tratamento de decide o resultado. Portanto, se você deseja obter os melhores resultados com o tratamento óleo de rosa mosqueta para a acne, você deve tentar encontrar o melhor e mais puro óleo de rosa mosqueta orgânica. Antes de comprar um produto, certifique-se que você leia os rótulos das embalagens adequadamente para verificar se ele contém óleo de rosa mosqueta puro ou de uma mistura de outros óleos essenciais. Leia as instruções de uso recomendadas pelo fabricante, porque alguns produtos requerem lavagem após alguns minutos da aplicação, enquanto que alguns precisam ser mantidos durante a noite.óleo de rosa mosqueta tem um cheiro desagradável e desagradável e muitas pessoas podem não gostar. Se você tem crianças em casa, eles podem ser desligados de você devido ao cheiro. Por isso, certifique-se de que você adicionar uma certa quantidade de óleo essencial aromático, tal como lavanda ou jasmim para travar para baixo o cheiro.[Ler: Como usar o óleo de abacate para acne?]Considerações ao usar o Óleo de Rosa Mosqueta"

Pergunta 1: Quais tratamentos para acne? Pergunta 2: Como é possível evitar o forte cheiro da rosa mosqueta no tratamento de pele?

InfoBox 2: LLM prompt for query creation.

```
O que é essa lista?  (None)
Qual empresa está na posição 32 da lista?  (994)

Qual é o objeto do contrato mencionado na passagem?  (None)
Qual é o acréscimo em reais no valor total do contrato após a
aplicação do reajuste contratual?  (3)

Qual foi a reação da pessoa ao abrir o aplicativo Pou?  (111)
O que aconteceu com a pessoa depois de jogar o Pou?  (88):

Qual é o título da tabela apresentada?  (None)
Qual é o salário base para o nível PM - 1,00?  (2)
```

InfoBox 3: Examples of synthetic queries filtering results. The values in parenthesis indicate the original passage retrieval position.

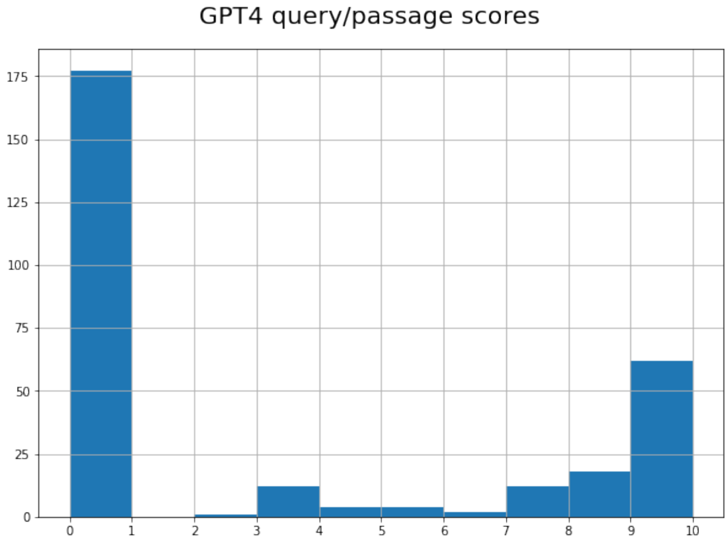

Figure 2: `gpt-4` relevance scores distribution for the pipeline sanity check.

The histogram in figure 2 indicates the `gpt-4` relevance scores are concentrated extreme notes, but covering practically the whole score range. Considering 5 as the starting threshold for a minimum passage relevance, around 1/3 (98) query/passages where deemed relevant, and 2/3 not. However, checking the scores per retriever, we realized the BM25 + mT5 had lower scores — BM25 + mT5: 1.541667 mean, 3.261441 std; BM25: 3.067708 mean, 4.225403 std; ColBERT-X: 4.786458 mean, 4.457034 std. We figure out that difference was due to an experimental error, and the selected BM25 + mT5 passages were not the best returned for each query.

Following [2], we intend to apply human verification over the LLM passage relevance check in two different steps:

1. Basic verification, sampling some of the LLM relevance check scores and asking a human to evaluate the result;
2. Extensive verification, applying human evaluation over the entire — or at least as much as possible — validation split query/passage associations.

We started the basic verification performing a single person analysis over a sample of 20 query/passage scored associations (from the 300 total), providing the query, passage and LLM score, asking the person to provide her own evaluation. With this very reduced set, we reached a Kendall Tau Correlation coefficient of 0.79 between the `gpt-4` and human relevance evaluation, which is not an impressive value. However, considering coefficients above 0.8 are considered to be within the Inter Annotators Agreement (IAA), this initial result is not far. Initial analysis indicated the `gpt-4` is still incurring in some comprehension errors — far more common for `gpt-3.5` and `text-davinci-003` — which leads overestimating the passage score; some passages which illustrate that behavior are presented in InfoBox 4 and analyzed here:

- Example 1: The LLM failed to recognize the passage contained a test statement, where the test question were exactly the proposed query: while the answer was one of the multiple-choice options presented, the passage itself does not provide the requested information; anyway, the LLM returned a score of 10, which way more than a human annotator provided, 1.
- Example 2: The LLM failed to recognize the passage was a biography of a volleyball player, which indeed won two times the Grand Prix competition, but that was not the overall Women Volleyball National Team results. Hence, the score of 10 is too high, since the required information is only partially provided — one can know the National Team won at least 2 times the Grand Prix competition. Human annotator score were 1.
- Example 3: In this final example the LLM failed to properly score a passage which indicates the answer for the query is available in the complete document — from which the passage was extracted — but itself does not

Example 1:
Query: Qual o relevo predominante do Brasil?
Passage: 32. (UDESC) Segundo Aziz Nacib Ab Saber, geógrafo, o relevo predominante no Brasil é:a) Depressão Central.b) Planícies e Terras Baixas.c) Planalto Brasileiro.d) Planície Costeira.e) Planalto das Guianas.33. ...
LLM score: 10
LLM reasoning: a passagem responde claramente à pergunta, indicando que, segundo o geógrafo Aziz Nacib Ab Saber, o relevo predominante no Brasil é o Planalto Brasileiro.

Example 2:
Query: Quantas vezes o voleibol feminino venceu o grand prix?
Passage: ... TítulosPela Seleção BrasileiraJuvenil ... Campeã do Grand Prix - 2005 ... Campeã do Grand Prix - 2008 ... Vice-campeã do Grand Prix de Vôlei - Ningbo 2010 ... Vice-campeã do Grand Prix de Vôlei - Macau 2011 ... Vice-campeã do Grand Prix de Vôlei - Ningbo 2012 ...
LLM score: 10
LLM reasoning: a passagem responde claramente à pergunta, indicando que o voleibol feminino venceu o Grand Prix três vezes (2005, 2008 e 2011).

Example 3:
Query: Existem vantagens ao definir uma moeda única?
Passage: 2.6 - Como se manifestam as vantagens do euro ... 8 2.7- Vantagens à escala mundial ... 82.8 - Tirar pleno partido das vantagens ...
LLM score: 4
LLM reasoning: a passagem menciona que existem vantagens da moeda única, mas não especifica quais são essas vantagens. Portanto, a passagem não responde completamente à pergunta.

InfoBox 4: Examples of LLM relevance evaluation errors, when compared to human annotators.

contain the answer, as it is a document summary. The LLM identified that scenario — which is explicit by its reasoning — but the provided score is higher than what a human annotator would provided — 1 in this case.

It is worth to state that due to a corpus pre-processing error, the special characters removal step replaced <`tab`>, <`new-line`>, and <`carriage-return`> by empty character, instead of blank one. That caused the final text to contain several linked words, which might have hindered the LLM comprehension.

As we have not completed the validation split relevance evaluation, we were not able to perform the extensive verification step.

4.7 New dataset performance

We have not executed yet the new dataset performance evaluation.

5 Conclusion

The developed work indicated modern LLMs are able to automatize key aspects of text IR datasets creation process, reducing the final cost to produce passage annotation, and that can be leveraged to created language specific datasets, which would allow better evaluation tools for IR systems operating in a given Language. However, some post-processing is still required in order to guarantee the dataset overall quality, specially if we apply LLMs to create new queries for a given text corpus.

6 Future Work

A lot of work remain to complete the goal of creating an IR dataset for Brazilian Portuguese native text corpus. More specifically, we identify the following tasks:

1. Fix the corpus pre-processing in order to remove the introduced noise after removing the special characters;
2. Expand the training split, generating more questions;
3. Perform additional prompt engineering work, trying to divert the LLM from the errors it still incur while performing the passage relevance evaluation;
4. Finish creating the validation split, performing the passage relevance evaluation for a bigger set of retrieved passages;
5. Perform the extensive verification of the LLM relevance evaluation over the final validation split.

6. Perform the final dataset evaluation, fine-tuning the ColBERT-X model and checking its performance against the mMARCO-pt fine-tuned one.

Once those steps are finished, we have already identified some future research topics to explore the use of Large Language Models for dataset annotation:

- Try identifying possible LLM biases when attributing scores for passage relevance, and how to better instruct it to provided judgements human annotators could disagree less.
- Explore the impacts of sending to the LLM multiple passages for relevance evaluation, possibly through an interactive execution, or using the recent LLMs with larger inputs.
- Distill a Language Model to reproduce the LLM passage relevance evaluation, continuing to reduce the final cost of creating a new IR dataset.
- Explore inter- and intra-annotation agreement across different LLMs, with the hypothesis that LLMs are more stable and knowledgeable on different subjects.
- Explore the annotation quality variation as the LLM is exposed — e.g. fine-tuned — to specific subject or topics.

References

[1] Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. mmarco: A multilingual version of the ms marco passage ranking dataset. *arXiv preprint arXiv:2108.13897*, 2021.

[2] Guglielmo Faggioli, Laura Dietz, Charles Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, et al. Perspectives on large language models for relevance judgment. *arXiv preprint arXiv:2304.09161*, 2023.

[3] Vitor Jeronymo, Mauricio Nascimento, Roberto Lotufo, and Rodrigo Nogueira. mrobust04: A multilingual version of the trec robust 2004 benchmark. *arXiv preprint arXiv:2209.13738*, 2022.

[4] Canjia Li, Andrew Yates, Sean MacAvaney, Ben He, and Yingfei Sun. Parade: Passage representation aggregation for document reranking. *ACM Transactions on Information Systems*, 2020.

[5] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362, 2021.

[6] Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W Oard. Transfer learning approaches for building cross-language dense retrieval models. In *European Conference on Information Retrieval*, pages 382–396. Springer, 2022.

[7] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713*, 2020.

[8] Arnold Overwijk, Chenyan Xiong, and Jamie Callan. Clueweb22: 10 billion web documents with rich information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3360–3362, 2022.

[9] Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *arXiv preprint arXiv:2101.05667*, 2021.