COS 470/570 Natural Language Processing
Lab 09

Shea Durgin, Sean Fletcher

April 17, 2023

**code related to this lab can be found at:
(https://github.com/sheaDurgin/NLP23/tree/main/Lab_9)

1. Parameters

Re-ranking:

We replaced tfidf with the DPH weighting model:

dph = pt.BatchRetrieve(index, wmodel="DPH")

Query expansion:

We added a query expansion technique:

bo1 = pt.rewrite.Bo1QueryExpansion(index)

Pipe-line:

Our pipeline needed to be updated:

pipeline = (bm25 % 20) » bo1 » dph

2. Results

Baseline:

precision@1: 0.0213

mrr@20: 0.1362

Our submission:

precision@1: 0.0390

mrr@20: 0.1569

3. Question analyses:

*please see the end of this document to read the question titles and bodies.

Pair (4231, 1985) Predicted correctly:

Both questions clearly give examples of AI generating work that would be identical or similar enough to already published work that it would likely trigger lawsuits. Both have a caveat explaining that it can be proven that the AI in question did not receive the work as an input source. These questions seem very similar to one another. We can see why our model also thought so.

Pair (15279, 15177) Predicted correctly:

These two questions are quite short and while the vocabularies don't seem to overlap too much, they do both have the name (name entity?) Clinton, and the use of the term "pardon." We're guessing that this name (or name entity) might be why the model was good at predicting them as similar.

Pair (3526, 1837) Predicted incorrectly:

These two questions actually seem like different questions to us. They're similar, but only in vocabulary. One is asking what something means; it's look for a definition. The other is asking about website maintenance and about how the law sees the terms in question, i.e. is one format preferable to others in the eyes of the law. While they might share some content and vocab, they're asking for different things. This is probably why our model thought they were dissimilar, because they are.

Pair (4043, 1985) Predicted incorrectly:

While these two questions seem very similar to us, the content of the questions are quite different from one another. One question gets right to the point. It gives a general explanation of AI generated material and explains their inquiry. The other question spends most of their body paragraph explaining what one particular computer generated corpus is and how it is searched. We got confused initially because the bulk of the question body really doesn't have a direct correlation to the actual question being presented. The specifics could easily have been summarized or even omitted. This is the reason we think our model also thought these questions were different.

Also, please note that question 1985 shows up both here and in a correctly predicted pair from above.

4. The questions in question:

ID: 4231

Title: Can computer generated images, text, or other artifacts infringe on a copyright

Body: a few examples: An artificial intelligence draws a picture identical to hello kitty, without having ever seen hello kitty before. an algorithm for generating science fiction novels creates a work nearly identical to some already published work, again without ever having received any of the copyrighted information as input. The reason I ask is because it is likely that this will eventually happen, and the spirit of a copyright is that it is intended to prevent someone from benefiting from another person s legitimate work. These cases strike me as a bit odd primarily because (given sufficient auditing of the generating system) it can be proven that the work (although nearly identical) is also original.

ID: 1985

Title: Is a randomly-generated book a violation of copyright

Body: Let s say I create a computer program that randomly chooses words and records them. Now, this generator spits out an exact word-for-word copy of a non-public-domain book. If I publish this (without knowing it s a copy), would I be in violation of the original book s copyrights What if I could reasonably prove that it was generated randomly

ID: 15279

Title: Can Obama pardon Clinton

Body: Pardons are mentioned in the United States Constitution at Article 2, Section 2. ... The President ... shall have Power to Grant Reprieves and Pardons for Offenses against the United States, except in Cases of Impeachment. Can Obama legally pardon Clinton for any potential crimes she might be convicted of in the future Would such a pardon likely withstand challenge

ID: 15177

Title: Can Clinton be pardoned without being charged or convicted

Body: Is it possible for someone to be pardoned if they haven t been convicted or even charged with a crime (edited)

ID: 3526

Title: What exactly does Copyright © [year] [company] on a website entail

Body: Almost every website has some variation of Copyright © [year] [company] at the bottom. Sometimes they also add All rights reserved . What exactly do those terms entail My biggest concern is this: by writing that, is the company claiming to own everything on the website, even potentially copyrighted user-submitted material

ID: 1837

Title: How to properly maintain website footer copyright notice

Body: There are often 3 copyright notice formats one might come across in the footer of website: Copyright © [Year the site or page was published]-[Year always kept current] Copyright © [Year this page on site was published] Copyright © [Year always kept current] Acknowledging there could be quite a many more possibilities, I hope to just focus on these formats at the moment. Are any of these more preferable in the eyes of the law, or are there situations where one of these may be more preferable Possibly a need to use a combination of techniques

ID: 4043

Title: Is the Library of Babel violating copyright

Body: The Library of Babel is a database of millions of random strings of text. Among this randomness, passages from other sources inevitably appear, such as the book of Genesis or works by Shakespeare, or this question. You can also search the library for specific strings. If a copyrighted work appears in the database, is it breaching copyright If not, can I use that text without breaching copyright (provided I indicate my source is the Library of Babel) You can find the Library of Babel at https: libraryofbabel.info