Exhibit 23

ludwig Nov 21, 2023 1:10 AM | Deleted on Dec 21, 2023 2:51 AM

sorry for not getting to this today, I was super busy with moving preparation

ludwig Nov 21, 2023 1:10 AM | Deleted on Dec 21, 2023 2:51 AM

from a quick look at the file, it does indeed have a format that's close to what we want

ludwig Nov 21, 2023 1:11 AM | Deleted on Dec 21, 2023 2:51 AM

and I agree that it would be nice to have one part of the pipeline that downloads all sorts of files regardless of type (and regardless of whether they're in our root set of URLs or one hop away)

ludwig Nov 21, 2023 1:12 AM | Deleted on Dec 21, 2023 2:51 AM

do you know where the URL lists are coming from?

ludwig Nov 21, 2023 1:13 AM | Deleted on Dec 21, 2023 2:51 AM

there are a few details I'd like to check, e.g., not all images have alt text in the file - is that because we sometimes don't store it, or because the image had no alt text? (somewhat confusingly, we sometimes store alt text "", i.e., empty string)

ludwig Nov 21, 2023 1:13 AM | Deleted on Dec 21, 2023 2:51 AM

another one is whether we get all image file types

ludwig Nov 21, 2023 1:13 AM | Deleted on Dec 21, 2023 2:51 AM

maybe it could be good if we chat about this today (Tuesday) for a bit. I'd be happy to meet any time 10:30 am PT or later @gyula

gyula Nov 21, 2023 2:11 AM | Deleted on Dec 21, 2023 2:51 AM

do you know where the URL lists are coming from?All of the urls in the [redacted] dataset are coming from CommonCrawl. Every link was collected, not just images (links in a,img,script and style tags).

gyula Nov 21, 2023 2:13 AM | Deleted on Dec 21, 2023 2:51 AM

there are a few details I'd like to check, e.g., not all images have alt text in the file - is that because we sometimes don't store it, or because the image had no alt text? (somewhat confusingly, we sometimes store alt text "", i.e., empty string)If the alt parameter is present, the link was crawled from an img tag. If the alt attribute is empty, then the img had no alt text. If the alt parameter is not present, the url is coming from an other tag type (a, script, style).

gyula Nov 21, 2023 2:14 AM | Deleted on Dec 21, 2023 2:51 AM

another one is whether we get all image file typesWe grabbed links, not file types. The links were not verified as far as I remember. It is easily possible that an url with an alt attribute points to an .exe file for example, if the person who created the website messed up the src tag for one reason or another.

gyula Nov 21, 2023 2:18 AM | Deleted on Dec 21, 2023 2:51 AM

maybe it could be good if we chat about this today (Tuesday) for a bit.Sent you an invite. If the time is not right for you then feel free to move it a little bit later.



Unknown Nov 21, 2023 1:00 PM | Deleted on Dec 23, 2023 2:26 AM

This message was deleted.

P **pranay** Nov 21, 2023 1:00 PM | Deleted on Dec 22, 2023 3:34 AM

[redacted]

G **gyula** Nov 21, 2023 1:57 PM | Deleted on Dec 22, 2023 3:34 AM

[redacted]

Sent on Nov 21, 2023 1:57 PM

[redacted]

pranay Nov 21, 2023 2:36 PM | Deleted on Dec 22, 2023 3:34 AM

Why can't we use nougat for pdf parsing?

pranay Nov 21, 2023 2:36 PM | Deleted on Dec 22, 2023 3:34 AM

😮

pranay Nov 21, 2023 2:37 PM | Deleted on Dec 22, 2023 3:34 AM

cc: @jennifer

jennifer Nov 21, 2023 2:38 PM | Deleted on Dec 22, 2023 3:34 AM

we can we just need the model first

jennifer Nov 21, 2023 2:38 PM | Deleted on Dec 22, 2023 3:34 AM

I mean we probably can, we also would need to figure out a way to do it while keeping security happy

Unknown Nov 21, 2023 2:39 PM | Deleted on Dec 23, 2023 2:26 AM

This message was deleted.

pranay Nov 21, 2023 2:39 PM | Deleted on Dec 22, 2023 3:34 AM

Do they not have a trained model we can use already?

pranay Nov 21, 2023 2:43 PM | Deleted on Dec 22, 2023 3:34 AM

[redacted]

pranay Nov 21, 2023 2:48 PM | Deleted on Dec 22, 2023 3:34 AM

I'll investigate the specifics of how it works to learn 🙂

jennifer Nov 21, 2023 2:50 PM | Deleted on Dec 22, 2023 3:34 AM

oh man last I checked I thought they did not. looks like they have base model weights. we can totally just run it

jennifer Nov 21, 2023 2:51 PM | Deleted on Dec 22, 2023 3:34 AM

from what i remember from their paper, caveats are it only works on english

pranay Nov 21, 2023 3:01 PM | Deleted on Dec 22, 2023 3:34 AM

@jennifer is there a IP/licensing reason we cannot use it?

[redacted]

pranay Nov 21, 2023 3:04 PM | Deleted on Dec 22, 2023 3:34 AM

Wait no that's not okay

pranay Nov 21, 2023 3:04 PM | Deleted on Dec 22, 2023 3:34 AM

Non-commercial?

jennifer Nov 21, 2023 3:04 PM | Deleted on Dec 22, 2023 3:34 AM

wait does it refer to the output?

pranay Nov 21, 2023 3:09 PM | Deleted on Dec 22, 2023 3:34 AM

Presumably? Not an expert

pranay Nov 21, 2023 3:09 PM | Deleted on Dec 22, 2023 3:34 AM

Ow the nc part of the license wouldn't do anything? :log-think:

jennifer Nov 21, 2023 3:09 PM | Deleted on Dec 22, 2023 3:34 AM

not an expert but my immediate thought was oh man of course my original interpretation was really dumb

pranay Nov 21, 2023 3:11 PM | Deleted on Dec 22, 2023 3:34 AM

I guess we could train it ourselves? Not excited about it, but doable. I asked devi in tokens-internal

[redacted]

Unknown Nov 21, 2023 4:53 PM | Deleted on Dec 24, 2023 1:58 AM

This message was deleted.

pranay Nov 21, 2023 4:53 PM | Deleted on Dec 22, 2023 3:34 AM

@jennifer do we have an ETA on the new tokens and/or net-new tokens from the crawler metric?

Edited on Nov 21, 2023 5:01 PM

@jennifer do we have an ETA on the new tokens and/or net-new tokens from the crawler metric?

Sent on Nov 21, 2023 4:53 PM

@jennifer do we have an ETA on the new tokens from the crawler metric?

P **pranay** Nov 21, 2023 4:53 PM | Deleted on Dec 22, 2023 3:34 AM

several people are very excited about this 😄