WOODHOUSE EXHIBIT 9

EXHIBIT H

Message

From: David Esiobu [redacted]@meta.com]
Sent: 4/20/2023 9:59:49 PM
To: Thibaut Lavril [redacted]@meta.com]; Melanie Kambadur [redacted]@meta.com]; Nikolay Bashlykov [redacted]@meta.com]; David Esiobu [redacted]@meta.com]; Jacob Xu [redacted]@meta.com]; Moya Chen [redacted]@meta.com]; Marie-Anne Lachaux [redacted]@meta.com]; Todor Mihaylov [redacted]@meta.com]
Subject: Message summary [{"otherUserFbId":null,"threadFbId":5997651113664213}]
Attachments: 339303581_1135728351153971_7606665540056163370_n.png; 342063938_738019914471161_8889390450152634049_n.png; 341302717_149787621385799_5187928533367492841_n.png

Melanie Kambadur (4/20/2023 07:51:02 PDT):
>Hey all! The lawyers/product team/and I [redacted]

Redacted

Redacted - Privilege

Marie-Anne Lachaux (4/20/2023 08:04:44 PDT):
>Hi Melanie, very good news !!
>What I know:

Redacted

Marie-Anne Lachaux (4/20/2023 08:16:21 PDT):
>with what we have*

Melanie Kambadur (4/20/2023 08:42:43 PDT):
>yeah this makes sense to me!

Melanie Kambadur (4/20/2023 08:42:54 PDT):
>thanks for weighing in here @Marie-Anne

Melanie Kambadur (4/20/2023 08:43:18 PDT):
>the thing that I want visibility into is if we are going to need extra GPUs for processing

Marie-Anne Lachaux (4/20/2023 08:43:57 PDT):
>for epubs no need :) only for pdf, so not the expert here !

Melanie Kambadur (4/20/2023 08:43:57 PDT):
>and also if we think we will still need to buy [redacted]. can we try to get answers to those questions soon-ish?

David Esiobu (4/20/2023 08:47:48 PDT):
>re: OCR i started testing out tesseract yesterday on the NYPL data. overall it doesn't appear to need GPUs but it's hit or miss depending on source quality. also starting to ask around to see if there are other tools/pipelines people have used successfully (happy to take suggestions here)
>
>from Tuesday though it sounded like only a small portion will require OCR though?

Nikolay Bashlykov (4/20/2023 08:52:33 PDT):
>yes, makes sense! to make a decision on [redacted] we can do the following:
>1. check % of [redacted] in libgen epub only format
>2. process epubs with the html script and do cross check VS [redacted] for similar books to compare
[@Marie-Anne could you help here?]

Marie-Anne Lachaux (4/20/2023 08:53:19 PDT):
>of course, I'll start today :)

Marie-Anne Lachaux (4/20/2023 08:53:51 PDT):
>with your help ofc ;)

Nikolay Bashlykov (4/20/2023 08:54:00 PDT):
>yep, we need OCR for scanned documents and the ones containing scientific formulas, so should be a smaller portion

Melanie Kambadur (4/20/2023 08:56:24 PDT):
>apologies if this is already what you're suggesting, but can we OCR a portion of the [redacted] books from libgen that match what we already have to compare? Then while we're doing that, measure time/GPU usage so we can project an estimate for GPU needs overall?

David Esiobu (4/20/2023 09:04:45 PDT):
>fun fact Yann was one of the inventors of the djvu format 😀 https://en.wikipedia.org/wiki/DjVu#History
>
>maybe he can help us with that

Nikolay Bashlykov (4/20/2023 09:06:47 PDT):
>yes, sure!

David Esiobu (4/20/2023 09:19:08 PDT):
>hey @Nikolay re: title matching, have you considered using ISBN? it seems the libgen site supports searching by that (although the python client might not)

Nikolay Bashlykov (4/20/2023 09:27:13 PDT):
>I tried that too, the main limitation is that ISBN is different for the same book if it has different edition, format (hardcover, paperback), so I was checking by title and the author match

David Esiobu (4/20/2023 09:55:38 PDT):
>also do we have the extracted/processed [redacted] sample anywhere? i'm not seeing it on hdfs [redacted]

Nikolay Bashlykov (4/20/2023 09:58:41 PDT):
>yeah, it's on RSC, since I ran ablations there: [redacted]
>but let me put it to S3, I think Marie-Anne was also looking for it

David Esiobu (4/20/2023 09:58:59 PDT):
>great, thanks!

David Esiobu (4/20/2023 10:03:27 PDT):
>oh i was hoping to find book metadata as well (e.g. title, author). would i need to go back to the xml for that?

David Esiobu (4/20/2023 10:03:57 PDT):
>or are the ids ISBNs?

Nikolay Bashlykov (4/20/2023 10:06:18 PDT):

>I think, I've included extensive metadata there, let me check

Nikolay Bashlykov (4/20/2023 10:10:40 PDT):
>there is DOI, like this: "10.1007/s00031-023-09793-5"

Nikolay Bashlykov (4/20/2023 10:47:52 PDT):
>here it is: [redacted]

Todor Mihaylov (4/20/2023 13:34:06 PDT):
>epub html is usually very good since it is optimized for e-readers. It does not have boilerplate etc. so we do not need a specific parser. Literally bs4(epub) should be good.

Marie-Anne Lachaux (4/20/2023 13:34:36 PDT):
>I think we can use my code to convert to markdown no ?

Marie-Anne Lachaux (4/20/2023 13:34:49 PDT):
>I know that Guillaume tested rapidly on some epubs at the time and it worked well

Marie-Anne Lachaux (4/20/2023 13:35:29 PDT):
>bs4 u mean beautiful soup ?

Marie-Anne Lachaux (4/20/2023 13:35:39 PDT):
>Yes that's what we do

Marie-Anne Lachaux (4/20/2023 13:35:47 PDT):
>we agree i guess then

Todor Mihaylov (4/20/2023 13:36:17 PDT):
>Yes, your code on top of the beautifulsoup should work well.

Marie-Anne Lachaux (4/20/2023 13:36:27 PDT):
>(sorry a bit tired today)

Marie-Anne Lachaux (4/20/2023 13:37:56 PDT):
>(bs4 is already is the code so i think it's gonna be straightforward)

Todor Mihaylov (4/20/2023 13:42:35 PDT):
>Btw, have we looked at https://sci-hub[redacted]

Todor Mihaylov (4/20/2023 13:43:07 PDT):
>A friend of mine has found a lot of paywalled publications here.

Melanie Kambadur (4/20/2023 13:44:45 PDT):
Redacted - Privilege

Marie-Anne Lachaux (4/20/2023 13:46:59 PDT):
>it's not included in libgen?

Marie-Anne Lachaux (4/20/2023 13:47:14 PDT):

shared: 339303581_1135728351153971_7606665540056163370_n.png

Marie-Anne Lachaux (4/20/2023 13:47:27 PDT):
>```
>You can download the torrents here. Torrents are provided by the Library Genesis project.
>```

Todor Mihaylov (4/20/2023 13:47:53 PDT):
>Libgen is actually powered by scihub https://www.vice.com/en/article/pa7jxb/archivists-are-trying-to-make-sure-a-pirate-bay-of-science-never-goes-down

Todor Mihaylov (4/20/2023 13:47:54 PDT):

shared: 342063938_738019914471161_8889390450152634049_n.png

Todor Mihaylov (4/20/2023 13:48:20 PDT):
>Although scihub mentions 88M documents and libgen states 900k ~1%?

Todor Mihaylov (4/20/2023 13:51:07 PDT):
>Again, based on my friends experience in torrenting books, I think that most of the books that anyone has found useful, are already in torrents which probably got to LibGen. Based on the crowd interest, I would assume that LibGen has the most interesting stuff and I would not rely on [redacted] has a lot of low quality publications..

Nikolay Bashlykov (4/20/2023 13:51:08 PDT):
>interesting, I saw estimates of >4M books + 80M articles in LibGen: http://freeread.org/torrents/

Todor Mihaylov (4/20/2023 13:53:34 PDT):

>Yes, probably I found the info of 900k from an old source

Marie-Anne Lachaux (4/20/2023 13:53:38 PDT):
>Sorry I am not sure I understand, how can we get the 99% remaining then ?

Todor Mihaylov (4/20/2023 13:53:49 PDT):

shared: 341302717_149787621385799_5187928533367492841_n.png

Todor Mihaylov (4/20/2023 13:55:01 PDT):
>The 80M science articles match the SciHub numbers so they probably have everything. I don't see how [redacted] alone can bring value.

Nikolay Bashlykov (4/20/2023 13:55:14 PDT):
>yes, that's about right

Marie-Anne Lachaux (4/20/2023 13:57:49 PDT):
>ok so we agree that if have libgen we have scihub right ?

Todor Mihaylov (4/20/2023 14:00:12 PDT):
>Right

Marie-Anne Lachaux (4/20/2023 14:00:22 PDT):
>cool :)

Nikolay Bashlykov (4/20/2023 14:05:30 PDT):
>re [redacted] that depends on the quality that we can get from parsing pdfs/epubs from libgen. I reckon not all of the [redacted] in libgen would have the easily parsed epub format, so we need to double check the text quality that we can extract

Todor Mihaylov (4/20/2023 14:08:21 PDT):
>Yes, I agree that it needs to be verified. I assumed that the quality of parsing would be matching the one from [redacted] but you have more experience with that.

Todor Mihaylov (4/20/2023 14:09:18 PDT):
>Btw, it would not be trivial to download libgen if everything is in torrents.
https://www.quora.com/What-is-the-probability-of-getting-arrested-for-using-torrents-in-the-USA

Nikolay Bashlykov (4/20/2023 14:09:33 PDT):
>for [redacted] we have the xml format - i.e. the ground truth so to say of the book. All the pdfs/epubs are then created from this xml layout

Todor Mihaylov (4/20/2023 14:09:35 PDT):
>I doubt that AWS would allow torrenting

Todor Mihaylov (4/20/2023 14:11:01 PDT):
>My assumption is that most of the scihub and libgen documents are uploaded from someone who had access to the actual pdf.

Todor Mihaylov (4/20/2023 14:11:04 PDT):
>or epub

Nikolay Bashlykov (4/20/2023 14:12:38 PDT):
>yes, but then we would need to convert pdf back to text (including tables, formulas, etc) and that's where we will might loose the quality

Todor Mihaylov (4/20/2023 14:49:45 PDT):
>It seems that commoncrawl has 2389008 ebooks. I will download them to S3. I have done .mobi parsing before so I have idea how to do that. @Nikolay do you have code for epub?

Nikolay Bashlykov (4/20/2023 14:51:40 PDT):
>I think either Marie-Anne or Guillaume have the code for epub

David Esiobu (4/20/2023 14:59:27 PDT):
>interesting, is this based on counting by content type?

Todor Mihaylov (4/20/2023 14:59:49 PDT):
>Yes