

EXHIBIT 6

Woodhouse

EXHIBIT 88

Message

From: Lukas Blecher [REDACTED]@meta.com]
Sent: 5/15/2023 6:56:24 PM
To: Nikolay Bashlykov [REDACTED]@meta.com]; Lukas Blecher [REDACTED]@meta.com]
Subject: Message summary [{"otherUserFbId":100083037918592,"threadFbId":null}]

Lukas Blecher (5/15/2023 05:22:50 PDT):

>Hi Nikolay,
 >I'm unable to locate about 470k pdfs on the faircluster. Some of the numbers just don't add up
 >Do you know how many files are at [REDACTED]? It takes forever to count them

Nikolay Bashlykov (5/15/2023 05:29:05 PDT):

>hey Lukas!

Nikolay Bashlykov (5/15/2023 05:29:07 PDT):

>EPUBs (310k): [REDACTED]
 >111,292 on FAIR Cluster
 >199,145 on RSC
 >PDFs (847k): [REDACTED]
 >647,932 on Fair Cluster
 >199,145 on RSC
 >loaded previously EN PDF/EPUB on fair cluster (~480k):
 [REDACTED]

Nikolay Bashlykov (5/15/2023 05:29:28 PDT):

>ah, the formatting is not good, check it out in the doc:
[https://docs.google.com/document/d/\[REDACTED\]](https://docs.google.com/document/d/[REDACTED])

Lukas Blecher (5/15/2023 05:46:58 PDT):

>I've seen that but that's not what I get when I try to find all of the pdfs.
 >The way I do it:
 >- Start from the [REDACTED] and get the 1.3M md5s we want to process.
 >- Check if it's already processed (480k)
 >- Check if the pdf is downloaded (330k), in the following order
 > - Check if it's at [REDACTED]
 > - Check if it's at [REDACTED]
 >
 >470k md5s match none of these conditions. About 160k of these are known to be corrupted. Another 200k are on RSC
 >That still leaves at least 200k PDFs we don't have, right?

Lukas Blecher (5/15/2023 05:52:43 PDT):

>Ah just found another 70k in your directory that have non-lowered md5 names. Is that possible?

Nikolay Bashlykov (5/15/2023 05:54:39 PDT):

>yes, that could be the case

Nikolay Bashlykov (5/15/2023 05:56:14 PDT):

>yeah, that's why I want to organize it better - to put it all in one folder 😊 but now I am blocked because apparently that's not easy to get data from RSC to faircluster

Lukas Blecher (5/15/2023 05:59:47 PDT):

>Anything I can do to help? Maybe I can help downloading to my home dir if space is an issue.

Nikolay Bashlykov (5/15/2023 06:22:12 PDT):

>that'll be great! that's what I thought:
 >- combine all the raw EPUBs/PDFs here: [REDACTED]
 >- copy all the remaining epub/pdfs from this list [REDACTED]
 [REDACTED] from the Guillaume's folder:
 >- chunk the files in the folders, I used this, but doesn't matter:
 [REDACTED]

Nikolay Bashlykov (5/15/2023 06:22:45 PDT):

>in the meanwhile, I'll try to transfer raw data from RSC to my laptop and then move it to fair cluster

Nikolay Bashlykov (5/15/2023 06:23:10 PDT):

>once we have all together we'll do another audit of what's downloaded

Nikolay Bashlykov (5/15/2023 06:24:29 PDT):

>after that we process the remaining PDFs and I move the scitech folder (pdfs & epub) to s3 bucket to free up the sapce on faircluster

Nikolay Bashlykov (5/15/2023 06:25:08 PDT):
>here is the function I used for PDF/EPUB detection:

>
>
>
>
>
>
>
>
>

Nikolay Bashlykov (5/15/2023 06:25:12 PDT):
>what do you think?

Lukas Blecher (5/15/2023 06:45:15 PDT):
>Ok sounds good. So how about this: I'm going to start re-downloading the corrupted PDFs from Guillaume's folder?

Nikolay Bashlykov (5/15/2023 06:45:46 PDT):
>re-downloading from libgen?

Lukas Blecher (5/15/2023 06:46:04 PDT):
>Yes, using your script

Nikolay Bashlykov (5/15/2023 06:47:14 PDT):
>yep, I have "Instructions to download libgen" in the doc

Nikolay Bashlykov (5/15/2023 06:47:23 PDT):
>the only thing

Nikolay Bashlykov (5/15/2023 06:47:37 PDT):
>do you want to re-download using fair cluster?

Lukas Blecher (5/15/2023 06:48:10 PDT):
>I was planning to do that, yes

Nikolay Bashlykov (5/15/2023 06:48:10 PDT):
>the problem is that we all share the same IP address there: [REDACTED] so since I'm loading already you'll be throttled most likely

Nikolay Bashlykov (5/15/2023 06:48:41 PDT):
>so you can re-download them to your laptop and then scp the data to faircluster

Lukas Blecher (5/15/2023 06:49:52 PDT):
>I didn't know that. Even if I'm on a devmachine?

Nikolay Bashlykov (5/15/2023 06:50:35 PDT):
>you can try "curl ifconfig.me"

Nikolay Bashlykov (5/15/2023 06:50:41 PDT):
>and see what ip you have

Lukas Blecher (5/15/2023 06:50:54 PDT):
>ok you're right ^^

Nikolay Bashlykov (5/15/2023 06:52:29 PDT):
>also you'll need to update in the script:
>MD5_LIST = [REDACTED] to your list of files
>AND this part in the code to select all the MD5 you want:

>
>
>
>
>
>
>
>
>

Lukas Blecher (5/15/2023 06:53:07 PDT):
>I'll see what I can do. I'm currently visiting my parents and they have *very* slow internet

Nikolay Bashlykov (5/15/2023 06:53:24 PDT):
>aah, yeah, that could be an issue ☺

Nikolay Bashlykov (5/15/2023 06:53:49 PDT):
>but you can help with organizing the files though if you want

Nikolay Bashlykov (5/15/2023 06:54:01 PDT):
>this one

Lukas Blecher (5/15/2023 06:55:11 PDT):
>Sure no problem. If we do that in place the current OCR jobs will fail. Not too big a problem since only me and Angela are running some at the moment (as far as I can see)

Lukas Blecher (5/15/2023 06:55:26 PDT):
>Also I'd need write permissions :)

Nikolay Bashlykov (5/15/2023 06:55:41 PDT):
>we can start with EPUBs

Nikolay Bashlykov (5/15/2023 06:55:48 PDT):
>as I've processed them already

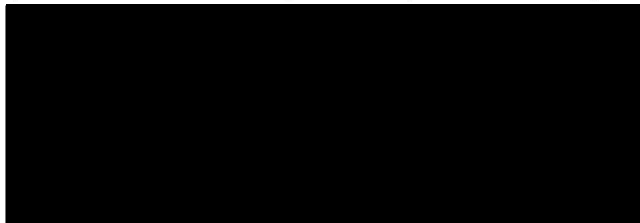
Nikolay Bashlykov (5/15/2023 06:56:07 PDT):
>and I think I've given the maximum permissions to my folder ☺

Nikolay Bashlykov (5/15/2023 06:56:35 PDT):
>you can try creating a tmp.txt file there or smth

Lukas Blecher (5/15/2023 07:15:20 PDT):
>On it :)
>I've just found something, not sure if you're aware: There are md5.None files (instead of .epub suffix) in libgen_epub.
>The ones I checked were corrupted. Is that intentional?

Nikolay Bashlykov (5/15/2023 07:16:13 PDT):
>yes, those are the corrupted ones, where the magic number didn't work:

>
>
>
>
>
>
>
>
>
>



Lukas Blecher (5/15/2023 07:16:58 PDT):
>Ok just wanted to check :)

Nikolay Bashlykov (5/15/2023 07:17:53 PDT):
>we can try re-downloading them again

Nikolay Bashlykov (5/15/2023 07:18:23 PDT):
>btw, which chunks to parse I can give to Naman? sorry for asking again, I think you already told me ☺

Lukas Blecher (5/15/2023 07:18:42 PDT):
>I'll collect all corrupted files and save the md5s

Lukas Blecher (5/15/2023 07:19:50 PDT):
>I think it was 1773-1998
>but if Sergey's chunk is still free Naman can also do that one

Lukas Blecher (5/15/2023 08:11:58 PDT):
>I need write permissions for the libgen_epub dir

Nikolay Bashlykov (5/15/2023 08:33:05 PDT):
>could you try now?

Lukas Blecher (5/15/2023 08:33:25 PDT):
>Yep, thanks

Nikolay Bashlykov (5/15/2023 08:57:50 PDT):
>hey Lukas, updates on the parsing:
>Naman took the part 1773-1998:
>Angela: started the run for indexes for Sergey and her own:

Lukas Blecher (5/15/2023 08:59:29 PDT):
>Did Angela take it all? Bc we can split them as well. I have capacity as well

Nikolay Bashlykov (5/15/2023 08:59:43 PDT):

>I thinks she just launched it all

Nikolay Bashlykov (5/15/2023 08:59:58 PDT):

>we'll use your capacity for the next chunk, which I'm loading from RSC

Lukas Blecher (5/15/2023 09:01:07 PDT):

>Ok, do you have an eta? I can also start with the PDFs that you loaded since the last time

Nikolay Bashlykov (5/15/2023 09:03:09 PDT):

>the first chunk would be in an hour or so, so yeah, you can start with the new ones, there are ~50k or so new

Nikolay Bashlykov (5/15/2023 09:35:57 PDT):

>hey Lukas, could you add status for parsing in the doc similar to this for downloads (no rush, we can do it tmr!) to have an overall picture of progress:

>Fiction EN download status (5pm 15.05.2023): (33% done of 1,159,720)

>EPUBs (338,797):

>PDFs (44,109) :

Nikolay Bashlykov (5/15/2023 09:37:40 PDT):

>*I'll add the status for EPUBs processing, so only PDFs from your side

Nikolay Bashlykov (5/15/2023 09:48:43 PDT):

>Number of Scitech EN PDFs: 1,272,655

>Number of Scitech EN EPUBs: 454,064

Lukas Blecher (5/15/2023 09:49:44 PDT):

>So 34% overall

Nikolay Bashlykov (5/15/2023 09:50:14 PDT):

>yep

Lukas Blecher (5/15/2023 09:56:56 PDT):

>here a vl list of all corrupt epub. There are some that slipped through your header check I noticed.

>Restructuring is in process

Nikolay Bashlykov (5/15/2023 09:57:38 PDT):

>we would need to re-download them, right?

Nikolay Bashlykov (5/15/2023 09:57:47 PDT):

>I'll do that

Lukas Blecher (5/15/2023 10:01:53 PDT):

>I think so, yes

Nikolay Bashlykov (5/15/2023 10:04:44 PDT):

>could you estimate how much GPU hours we would need for sci-mag? it's roughly same size as scitech and I assume all PDFs

Nikolay Bashlykov (5/15/2023 10:05:12 PDT):

>we can give a ball park number similar to what we've used for scitech

Nikolay Bashlykov (5/15/2023 10:06:27 PDT):

>~2k GPUs on FAIR Cluster for 2 weeks to parse that?

Lukas Blecher (5/15/2023 10:08:51 PDT):

>I think the overhead will be bigger for sci-mag bc there are fewer pages per document. A paper might be only 2 pages long and that wastes VRAM. I think I could implement a better batching

>A similar time frame is probably a good guess. A bit longer than sci-tech for sure though

Lukas Blecher (5/15/2023 10:09:19 PDT):

>I'll do some quick calculations

Nikolay Bashlykov (5/15/2023 10:10:35 PDT):

>there are 82M documents with median 6 pages per doc. BUT I don't have the share of english, so we can assume 50% (similar to scitech)

Lukas Blecher (5/15/2023 10:30:51 PDT):

>2.6 weeks is what I get with a rather large uncertainty.

>I could imagine that the conversion speed is not the same because books tend to be less information dense than papers

Nikolay Bashlykov (5/15/2023 10:31:44 PDT):

>ok, so we need 2.6k to fit in 2 weeks

Lukas Blecher (5/15/2023 10:32:52 PDT):

>Ah wait I missed something

Lukas Blecher (5/15/2023 10:34:24 PDT):

>1 week will probably be enough. I'll write down the calculations real quick

Nikolay Bashlykov (5/15/2023 10:45:54 PDT):

>btw, I will put th PDFs from my devserver here (200k new PDFs to parse):
[REDACTED]

Lukas Blecher (5/15/2023 10:47:19 PDT):

>Ok nice! Since I was missing the pdfs with capital letters in the name before I have enough to do at the moment. If anybody needs new pdfs, I have plenty splits ready

Nikolay Bashlykov (5/15/2023 10:48:05 PDT):

>aah, yeah, I think, we'll need to lowercase them all, to avoid confusion 😊

Nikolay Bashlykov (5/15/2023 10:48:19 PDT):

>ok, so I'll ask for 2 weeks just in case

Lukas Blecher (5/15/2023 11:10:37 PDT):

>I did that for the epub's (it's still collecting but should be done soon) :) Don't know what to do with the .None files though

Lukas Blecher (5/15/2023 11:14:58 PDT):

>eta 1h

Nikolay Bashlykov (5/15/2023 11:56:24 PDT):

>@silent the ".None" - they'll go to the re-download list