EXHIBIT 155

danny Jun 4, 2021 12:53 PM

One of my [redacted] has failed a few times, and I'm trying to figure out how to best debug it. Do the [redacted] somewhere I can search for errors? It seems to get stuck at stage 3 or 4, and to not give errors I can see tmuxing into the orphaned consoles (just warnings).

danny Replied on Jun 4, 2021 12:54 PM

command

[redacted]

sam-pling Replied on Jun 4, 2021 12:54 PM

you could try [redacted]

danny Replied on Jun 4, 2021 12:54 PM

[redacted]

sam-pling Replied on Jun 4, 2021 12:55 PM

coo cloudwatch yourjobname will also show logs

sam-pling Replied on Jun 4, 2021 12:55 PM | Edited on Jun 4, 2021 12:56 PM

oh I see what you're saying though, [redacted]

sam-pling Replied on Jun 4, 2021 12:56 PM

do you have one running now? does it stop, or just hang forever?

danny Replied on Jun 4, 2021 12:56 PM | Edited on Jun 4, 2021 12:56 PM

I see a bunch of these, and [redacted] by one of the running tasks) Reason: Remote RPC client disassociated. Likely due to containers exceeding thresholds, or network issues. Check driver logs for WARN messages.

danny Replied on Jun 4, 2021 12:58 PM

Often [redacted] recovers from some warnings like that during training, but keeps getting stuck at the same stage (3 and 4 now).

I don't have many ideas besides make a new box

sam-pling Replied on Jun 4, 2021 12:58 PM

what's the job name?

danny Replied on Jun 4, 2021 12:59 PM

I'm currently running jobs directly on the box, since I couldn't get monitoring commands like [redacted]

Exhibits#:
Mann 205
11/20/2025
exhibitsticker.com

danny Replied on Jun 4, 2021 1:00 PM

Maybe that's part of current cluster issues though

(base) danny@Dannys-MBP anthropic [redacted]

jared Replied on Jun 4, 2021 1:03 PM

I can pair on this right now if you like?

(Just got back from an errand)

jared Replied on Jun 4, 2021 1:04 PM

It sounds like if it worked with a lower [redacted]

Naively it probably shouldn't, given that you have [redacted] but.... 🤷

jared Replied on Jun 4, 2021 1:05 PM

@ben and I usually need [redacted], though that's significantly larger

One thing you can do is use [redacted] to look a bit more at what [redacted] I can show you how. I think it's also in the [redacted]

ben Replied on Jun 4, 2021 1:07 PM | Edited on Jun 4, 2021 1:07 PM

if you get an error like this [redacted] send me your pod name and i'll fix it. have to [redacted]

ben Replied on Jun 4, 2021 1:08 PM

usually when you have Lost Task it's because one of the workers OOM'd or got killed

ben Replied on Jun 4, 2021 1:09 PM

i'm available for call if needed

danny Replied on Jun 4, 2021 1:22 PM

It looks like a lot of tasks during stage 3 and 4 failed. My current plan is to [redacted], in case it's an issue of [redacted], but if that doesn't seem promising ben a call would be helpful

danny Replied on Jun 4, 2021 1:26 PM

Got a lot of warnings like this also

[redacted]

[redacted]

ben Replied on Jun 4, 2021 1:26 PM

👍

ben Replied on Jun 4, 2021 1:26 PM | Edited on Jun 4, 2021 1:27 PM

[redacted]

danny Replied on Jun 4, 2021 1:27 PM

[redacted]

ben Replied on Jun 4, 2021 1:27 PM

you should have [redacted] probably

ben Replied on Jun 4, 2021 1:27 PM | Edited on Jun 4, 2021 1:28 PM

each box has [redacted]

danny Replied on Jun 4, 2021 1:28 PM

Hmm. Pair for a bit then?

ben Jun 4, 2021 3:06 PM

Done benchmarking extractors on [redacted] test documents. In summary, readability is [redacted] slower than the others, so I'll ban it for now. newspaper, which was what we used in GPT-2, had an error rate of [redacted], however most of the other extractors output nothing in those cases anyway. Based on this, I'm pretty comfortable sticking with just newspaper

+1 2

jared Replied on Jun 4, 2021 3:08 PM | Edited on Jun 4, 2021 3:08 PM

Crazy thought — use newspaper, train a filter, then go back and use readability (if it's much better) on the docs that we decide have passed the filter.

sam-pling Replied on Jun 4, 2021 3:08 PM

how does it end up improving vs. justext?

➕ **1**

sam-pling Replied on Jun 4, 2021 3:09 PM | Edited on Jun 4, 2021 3:09 PM

if you happen to have any comparisons to look at, I'd be interested

jared Replied on Jun 4, 2021 3:09 PM | Edited on Jun 4, 2021 3:09 PM

An even crazier thought is to just use the WET files to train a filter, and then go back and get and extract the docs that are rated as high quality. Sounds like that would be too far off distribution and the filter wouldn't work well.

ben Replied on Jun 4, 2021 3:10 PM

i will have comparisons to share very soon!

rolled_up_newspaper 1 **1**

little cleanup needed

justext is about [redacted] faster, but it includes a ton of boilerplate

+1 1

even if you set the no_boilerplate flag

1

sam-pling Replied on Jun 4, 2021 3:11 PM

boilerplate = useless junk in addition to real text?

ben Replied on Jun 4, 2021 3:29 PM

eg headers, footers, navbars, cookie policies, etc

1

ben Replied on Jun 4, 2021 3:38 PM | Edited on Jun 4, 2021 3:40 PM

Small sample:

[redacted]

ben Replied on Jun 4, 2021 11:19 PM | Edited on Jun 4, 2021 11:20 PM

I'm starting to think @jared's WET idea isn't that bad… it'll be orders of [redacted], and hopefully enough of it's good that the model will learn to ignore the boilerplate. I'll get us a sample by end of weekend.

+1 1

sam-pling Replied on Jun 4, 2021 11:24 PM | Edited on Jun 4, 2021 11:26 PM

scrolling through those samples — ~~justext~~ newspaper has a lot more "error": "'NoneType' object has no attribute 'xpath'" — this is the error you were talking about before?

ben Replied on Jun 4, 2021 11:26 PM

WET actually looks ok! Lots of junk but good extraction. I think they improved the extractor quality after I complained ~1y ago [redacted]

1

S **sam-pling**

justext:
About
Standards
History
Other Documents
Home/Westminster Larger Catechism/2. How doth it appear that there is a God?
2. How doth it appear that there is a God?
388 views
The very light of nature in man, and the works of God, declare plainly that there is a God; 7Rom. 1:19–20, Ps. 19:1–3, Acts 17:28 but his word and Spirit only do sufficiently and effectually reveal him unto men for their salvation. 81 Cor. 2:9–10, 2 Tim. 3:15–17, Isa. 59:21
Westminster Larger Catechism
1Manuals
6Standards
33Westminster Confession of Faith
196Westminster Larger Catechism
107Westminster Shorter Catechism
Flatbase by Nice Themes © 2019. — Powered by WordPress.
Footnotes [+]
1.
↑
Rom. 1:19–20, Ps. 19:1–3, Acts 17:28
2.
↑
1 Cor. 2:9–10, 2 Tim. 3:15–17, Isa. 59:21
3.
↑
Rom. 1:19–20, Ps. 19:1–3, Acts 17:28
4.
↑
1 Cor. 2:9–10, 2 Tim. 3:15–17, Isa. 59:21
5.
↑
Rom. 1:19–20, Ps. 19:1–3, Acts 17:28
6.
↑
1 Cor. 2:9–10, 2 Tim. 3:15–17, Isa. 59:21
7.
↑
Rom. 1:19–20, Ps. 19:1–3, Acts 17:28
8.
↑
1 Cor. 2:9–10, 2 Tim. 3:15–17, Isa. 59:21

newspaper:
The very light of nature in man, and the works of God, declare plainly that there is a God; but his word and Spirit only do sufficiently and effectually reveal him unto men for their salvation.

that's a significant improvement

+1 1

B **ben**

[redacted]
@sam-pling newspaper has the xpath error i'm pretty sure

jared Replied on Jun 4, 2021 11:26 PM

Do we know what WET looks like as a comparison? Also @ben if they only improved their extraction in the last year, does that mean only very recent data is good, or did they re-do it all?

sam-pling Replied on Jun 4, 2021 11:26 PM

oops yeah I meant newspaper

ben Replied on Jun 4, 2021 11:26 PM

they didn't redo it

@jared [redacted]

sam-pling Replied on Jun 4, 2021 11:26 PM

not sure if it's representative, but the amount of garbage in here is quite striking

ben Replied on Jun 4, 2021 11:28 PM

it is representative

there's a boatload of garbage in CC

that's why filtering matters so much

+1 2

ben Replied on Jun 4, 2021 11:29 PM

Here's your 2013 comparison, maybe even more garbage-y [redacted]

[redacted]

+1 1

jared Replied on Jun 4, 2021 11:29 PM

[redacted]

sam-pling Replied on Jun 4, 2021 11:29 PM | Edited on Jun 4, 2021 11:29 PM

(I'm just looking at your earlier samples so far (newspaper vs justext))

ben Replied on Jun 4, 2021 11:30 PM

goose is pretty good too, and gives really nice rich metadata

👀 1 🙌 1

ben Replied on Jun 4, 2021 11:31 PM

Their WET files have encoding issues much less frequently than my extractions. I didn't want to spend the compute detecting the proper encoding and the one provided is frequently wrong.

sam-pling Replied on Jun 4, 2021 11:32 PM

goose does seem to have a lot of stuff that newspaper missed?

just doing a very-unsophisticated scroll through though

ben Replied on Jun 4, 2021 11:32 PM

yea, goose is a pure-python port of readability.js

so i expect it to be better

less special-cased to news

+1 2

ben Replied on Jun 4, 2021 11:33 PM

It's about [redacted], but that's negligible

🙌 2 🚀 1

sam-pling Replied on Jun 4, 2021 11:33 PM

it also looks to support a lot more languages

ben Replied on Jun 4, 2021 11:34 PM

it does claim to, [redacted]

in fact, we may want to mostly english filter for now to save model capacity

sam-pling Replied on Jun 4, 2021 11:37 PM

scrolling through WET – is it way way more complete?

sam-pling Replied on Jun 4, 2021 11:38 PM

there's far more text in there (or maybe it's just a different sample?)

ben Replied on Jun 4, 2021 11:51 PM

it does seem like it's very complete

the extractors are all very flaky for some reason... could be due to the encoding errors i mentioned up thread

ben Replied on Jun 4, 2021 11:52 PM

wow that was easy! [redacted]

2

ben Replied on Jun 4, 2021 11:53 PM

(above is a heavily subsampled)

sam-pling Replied on Jun 4, 2021 11:56 PM

do you have a sense of what WET misses, but the other extractors catch?

ben Replied on Jun 4, 2021 11:58 PM

i'd expect it to be things like worse formatting, more boilerplate in WET

+1 1

sam-pling Replied on Jun 4, 2021 11:59 PM | Edited on Jun 4, 2021 11:59 PM

[redacted]

these are meant to have the same data? or different

I assume different

ben Replied on Jun 5, 2021 12:01 AM

[redacted]

sam-pling Replied on Jun 5, 2021 12:02 AM | Edited on Jun 5, 2021 12:03 AM

cool – looking at [redacted]

[redacted]

ben Replied on Jun 5, 2021 12:03 AM

so the corresponding WET should be [redacted]

ben Replied on Jun 5, 2021 12:03 AM | Edited on Jun 5, 2021 12:03 AM

yep beat me to it

✅ 1

ben Replied on Jun 5, 2021 12:04 AM

if you look at it like this you can [redacted]

ben Replied on Jun 5, 2021 12:04 AM | Edited on Jun 5, 2021 12:04 AM

and if you do this you can make the [redacted]

ben Replied on Jun 5, 2021 12:06 AM | Edited on Jun 5, 2021 12:06 AM

i set up a [redacted]

ben Replied on Jun 5, 2021 12:06 AM

3:06am 😱 nighty night

👌 2

sam-pling Replied on Jun 5, 2021 12:10 AM

WET definitely contains a ton of garbage in addition to the meat we want

just tons and tons of boilerplate stuff

sam-pling Replied on Jun 5, 2021 12:11 AM

on the other hand, it includes some great stuff that e.g. goose just drops (left is WET)

Deleted on 06/04/2024 12:47 AM

image.png 2 MB

🙌 1

(search for Notice that most teams are very close to neutral risk)

justext (right) does keep it , but mangles

Deleted on 06/04/2024 12:47 AM

image.png 1 MB

sam-pling Replied on Jun 5, 2021 12:12 AM

looks like goose gives duplicates of the text above? also odd.

ben Replied on Jun 5, 2021 12:16 AM

[redacted]

Deleted on 06/04/2024 12:47 AM

Screen Shot 2021-06-05639 KB

+1 1

sam-pling Replied on Jun 5, 2021 12:16 AM

oooh even better

ben Replied on Jun 5, 2021 12:16 AM

yea it's md

1

B **ben** Replied on Jun 5, 2021 @ 1:5 AM

did a little polishing on [redacted]

[redacted]

+1 1