Topic: Midjourney prompt data from (some) public facing Discord channels.

https://bridges.monash.edu/articles/dataset/Midjourney_2023_Dataset/25038404
This is one of the few datasets I saw floating around. I chose this one in particular to share due to it being relatively small in size and includes the prompts themselves.

Based on the detailed information further below, I think the broad question would be “what type of trends occurred during this period of time from images generated on the Midjourney Discord?”

I feel like some observations that could be taken from the data and be interesting in a visual format would be: Average prompt length, frequency of words, frequency of prompt generation during particular days or times (the dataset states that the timestamp is from when the prompt was entered, but it’s more likely to be when the images were generated and returned by the MJ bot). Other questions could be "average strong/soft variation rolls per unique prompt?”
This particular dataset also includes links to the resulting image generations. I’m not particularly interested in that data, but I could see it being of interest to others.

There could also be a few questions like “Who prompted the most in this time frame and what were their most popular topics?” But that feels more creepy and invasive than anything.

Overall I have in mind: a dashboard with a few different populating charts based on word and/or date input. The resulting presentation could have my own generated charts based on my curiosities, but also the dashboard for others to explore their own as well.
The first big issue that comes to mind is run time for searching through strings.

As a side tangent, it would be neat to take my own personal Midjourney bot data and see the trends within it and perhaps do some comparisons but: pretty sure bot scraping is against Discord’s TOS and I like being not banned, I pay $60/mo so I can have stealth 🤔 why would I air out my cringe? Requesting data from Discord takes up to 30 days so it’s possible I won’t get it in time to tidy and fiddle with it regardless. It could be fun to look at personally anyway.

Initial questions / notes about the dataset(s): When Midjourney renders an image set, it renders a quad of images and considers this ‘usage’ of quad of images as 1 image. Is this data considering “upscales” as individual prompts as well? The dataset states “the upscale dataset contains prompts that were selected for upscaling only,” but breaking an image out of a quad is still called “upscaling” (left over from v4) while the actual upscaling occurs after breaking an image out of the quad and selecting the desired upscale style. I’d assume they mean the latter method but still something to keep note of.

The linked dataset has 500,000+ prompts. Based on the time stamps in this dataset, the prompts are from 10/24/23-11/9/23. Seeing how many unique prompts (vs rerolls or variation rolls) are within this data set would require some filtering and pandas magic.

Based on my knowledge of how the Midjourney Discord works, these prompts would only be from the categories available to the scraping bot at the time. So notably lacking would be prompts from the newbie rooms that it wasn’t assigned to and presumably the 25k-create channels. It, of course, lacks the data from images generated outside of the Discord, directly messaged to the Midjourney Bot, and data from the Midjourney website itself, which is currently in Alpha stage for website image generation.

Since the data has links to the images generated, it lacks the rejection responses including “request cancelled due to image filters” and “banned prompt detected”. Though, after double checking the latter shows up as ephemeral messages anyway and wouldn’t be public facing. Unfortunate, as I’d love to see what weird stuff people are trying to generate on a public facing image rendering channel.

Audience
Intended audience would be myself, and anyone curious of various frequencies of prompt data points from Midjourney during the span of time the data was collected. I imagine anyone interested in ‘image prompt analysis’ could have interest.

Milestone 1 (Apr 15)
Basic layout mocked up, efficient string search algorithms researched, appropriate graph types researched, similar projects researched.
Some basic implementing of the data into graphs should be feasible.

Milestone 2 (Apr 22)
Graphs laid out in Jupyter, data/search filtering, make sure graphs and data are interacting and behaving as they should. Additional research as needed.
Bugfixing. Address feedback. Make sure data isn’t hardcoded in so that switching in datasets is least painful as possible.

Milestone 3 (Apr 29)
Dataset finalization - I presume to have this done earlier, but I’m putting it here in the event a better / more recent dataset comes out in the time period. Existential crisis.
Bugfixing, address feedback, start making the visualization pretty with CSS (I’m lying to myself making it pretty is going to be the first thing I do).
If I’m feeling spicy I might try to implement the dashboard onto a website.

Milestone 1

From the additional meeting time one of the things I'm interested in showing is how many times a prompt is rerun or edited slightly to change the output. Though the reason someone decides to rerun a prompt is likely to be incredibly subjective, I think showcasing the data behind this will illustrate that prompting for an image isn't necessarily "easy". A single run of a prompt wouldn't necessarily indicate a success either, especially considering the public nature of the Discord and how quickly the channels rapid-fire bot responses.

One thing I think I'm getting slightly stuck on is the idea that since I'm looking at data that is primarily text based, that the visualizations need to be, too.

A word cloud feels like it could be an interesting introductory visualization, but I feel like I definitely would want to make it interactive in a manner of 'ok let's take a close look and see why some of these words have such high frequencies'.

Seeing the websites listed on datasketch.es was particularly useful in seeing that it's reasonable and fair to give insight and explanation into the presented data. I've been pondering a decent bit on: "okay, how do I show how Midjourney works so people understand the decisions made, and can see the differences between it and the rather different, but often more familiar ChatGPT text input process?" Turns out I can probably just… explain it. Wild.

I don't have tangible layouts made yet but I think after today's class I definitely have a better mindset for how I might want to lay things out. Existential crisis has been moved to milestone 1 instead of 3. That said, I felt like maybe I was lagging behind pretty hard but I think I have a pretty solid dataset and good directions to explore. Sitting down and playing around with the data is really what I need to do, instead of trying to conceptualize everything in my mind before putting pen to (Jupyter) notebook.

String/Text Search
https://towardsdatascience.com/5-methods-for-filtering-strings-with-python-pandas-ebe4746dcc74

https://www.adventuresinmachinelearning.com/efficiently-searching-for-strings-in-pandas-dataframe/

https://pandas.pydata.org/pandas-docs/stable/user_guide/text.html

General text analysis / visualizations
https://towardsdatascience.com/a-complete-exploratory-data-analysis-and-visualization-for-text-data-29fb1b96fb6a

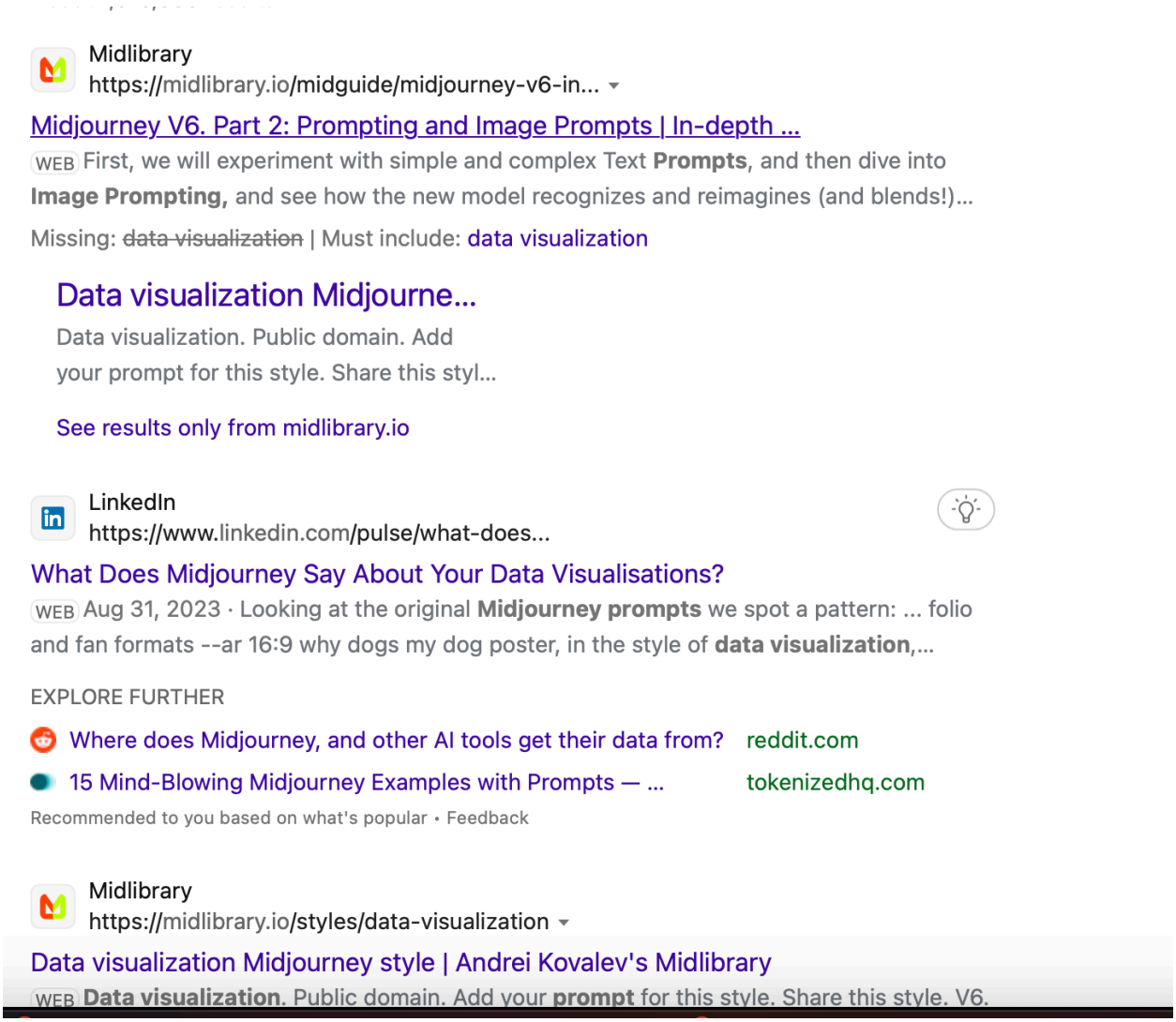
https://levelup.gitconnected.com/crafting-clean-narratives-with-nlp-a-deep-dive-into-text-data-preprocessing-with-python-pandas-37374949603b

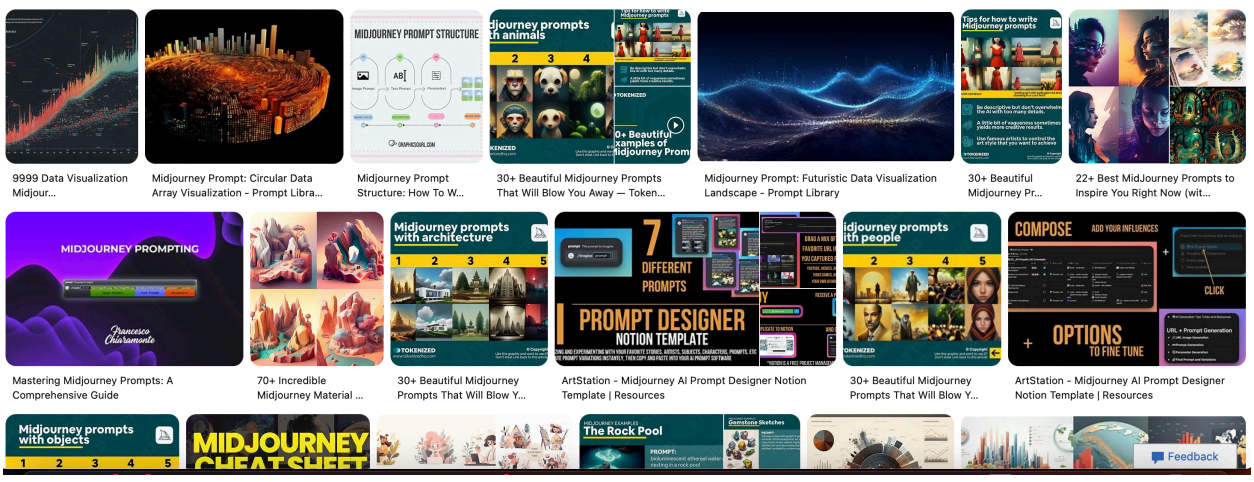
https://towardsdatascience.com/advanced-visualisations-for-text-data-analysis-fc8add8796e2

https://towardsdatascience.com/data-storytelling-with-animated-word-clouds-1889fdeb97b8

Scattertext Visualization
https://guides.library.upenn.edu/penntdm/python/scattertext

I searched for “data visualization of midjourney prompts” and I don’t know what I expected.





Similar Projects

Finding similar projects or even any interesting information is turning out to be a bit of a chore as most results are websites that want to inform people on how to “prompt better” or list prompts for sale.

https://www.kaggle.com/code/succinctlyai/midjourney-prompt-analysis

Milestone 2

The built in pandas methods for string searching seem to work really well without really needing to fuss about optimization. I’m pretty sure I’m still traumatized from trying to take 15-110 at CMU with the early semester focus on sorting methods, recursion and iteration. Probably says something that I even remember the course number 11 years later… ANYWAY.

I’ve been pondering on exploring Natural Language Processors and how I’d potentially want to interact with the data using those versus not. Textblob is allegedly beginner friendly. Based off what I’ve seen it shouldn’t be too troublesome to look into.

Sitting down and just playing with the data definitely helps and definitely helps with thinking about how particular numbers may correlate to others.

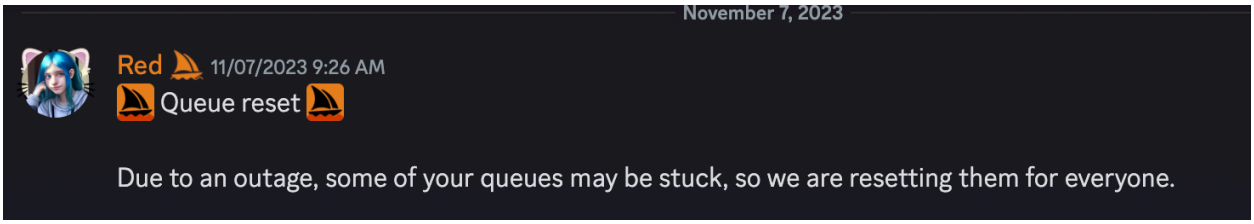
I’m still lagging behind a bit of where I’d like to be (let’s run a workshop right at the end of the semester I said, it’ll be fun, I said), but at least it’s with a feeling of ‘I’m looking forward to sitting back down with this so I can get these questions answered’ and a sense of direction. Importantly, I commented in the questions so I’ll actually remember them. I don’t think necessarily all of them will be relevant to final visualizations; some of them are from pure curiosity on my end. Important part is to not spend overly much time on those.

I was surprised to see how many prompts some individuals generated over the course of time this dataset covers. I’m tempted to use the user ID and creep on their galleries on the main website, but that feels overly creepy. I almost feel like I’d prefer to randomize the user id’s with new unique id’s so it feels less personal.

Milestone 3

Running some additional questions made me realize I probably should’ve thought of some more basic questions baseline. Such as: Are any days missing? Turns out there seems to be no data for Halloween (criminal) and none for 11/7.

Of note on the days that are completely missing or have single or double digit prompts:



Makes me wonder how to include information such as this, if at all. I at least think it's personally interesting to have a starting point for speculation as to why the bot(s) may have failed to scrape on particularly dates/times. I'm not entirely sure on how Discord bots work across the board, but I suppose it's possible that some of the outages could've impacted larger sets of bots. Though, the scraping bots were likely to be "self" bots, or bots acting as users, whereas otherwise Discord bots are labeled as such. The former is against Discord's TOS, as is scraping, of course.

Which brings me to: https://discordstatus.com/uptime?page=2
Notably there was an API outage on Nov 7.

—

In regards to the milestone I set for myself, I feel like I'm definitely going to be going with this dataset as nothing else that I've seen has popped up, and I'm perfectly fine with that.

I'm still a bit behind where I'd prefer to be but I'm happy it's at least not due to lack of content or direction. I'll have to think of things in terms of minimum viable product to hopefully keep track of the main idea. There's definitely an interesting balance with wanting to be transparent about what the data has / what has been removed, etc etc and keeping things simple + communicating clearly. Data viz seems like it would be a really interesting area for user testing to see how successful the story telling actually is.

And, of course, I had a brief thought of 'I'm not going to get distracted by that,' but I've gone and already found a link to read later :^) https://3iap.com/key-questions-for-user-testing-data-visualizations-5vJ8JychRVGIGWq-TpFIIg/ (that hero image was definitely AI generated). And to be fair to myself, UX/UI is one of my key areas of interest.

https://github.com/unzippedzebra/indatasp24Final

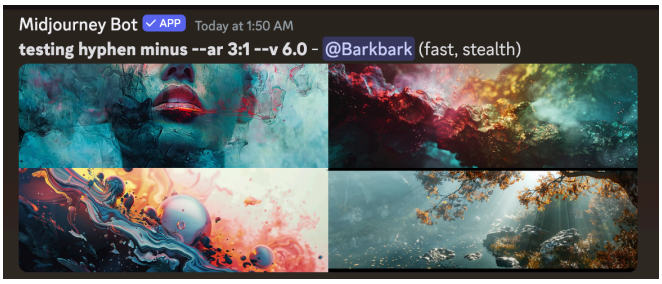
Final aaaah'ing:

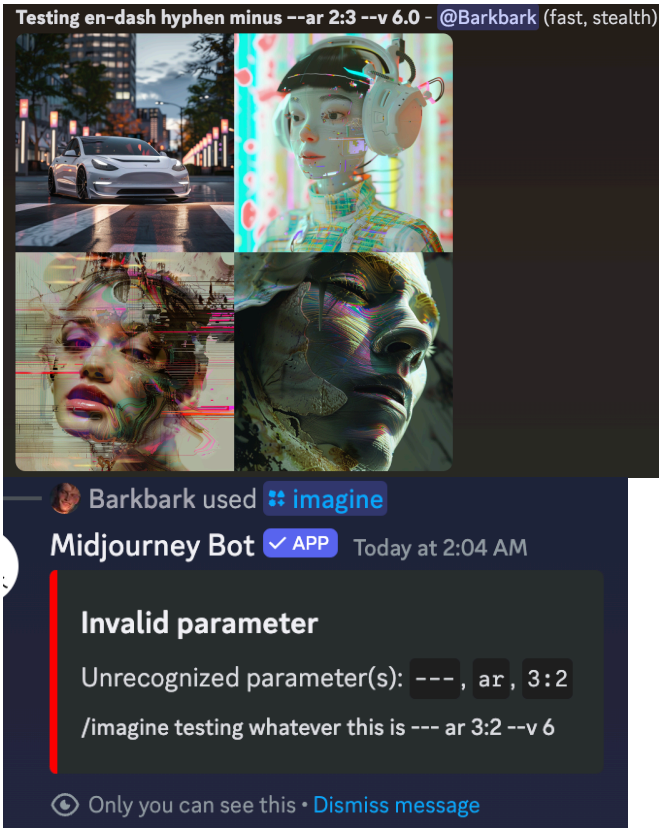
I keep thinking of / remembering little quirks that probably throw off the data filtering and cleaning. A big one is the modifiers. You can use double dash *or* the en dash — . And of course pandas is only going to filter for the double dash when that's what you tell it…
The — is more likely to come from Apple users which could be an interesting little data point on its own.
I remembered this detail when only 5 unique modifiers returned: niji, tile, upbeta and uplight. What's really bizarre is the lack of ar (aspect ratio) and that all the prompts that have modifiers only have one? I wonder if that's a mistake in my filtering or a mistake in the data collection.

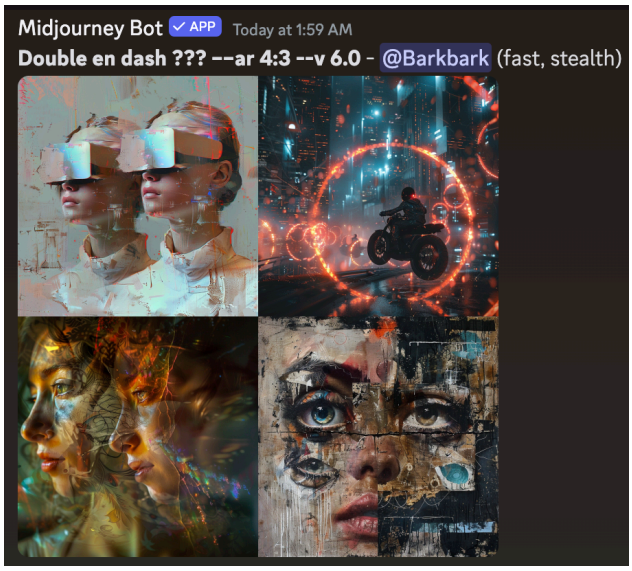
Ohh my god this is so much more painful than I expected. Filtering the parameters ended up being turbo hell for various reasons.

As per Midjourney documentation: Many Apple devices automatically change double hyphens (--) to an em-dash (—). Midjourney accepts both!

…But they also accept wonky things like –- ⸗ . I ended up running a few through mj itself instead of fighting unicode directly







After figuring that out, figuring out the proper regex to extract those strings was a mixture of proper hell, bargaining, ritual sacrifice, checking ChatGPT 4 / Co-Pilot's bullshit on https://regex101.com/ , more bargaining and a little extra existential crisis.

I wondered at some point how often people made mistakes putting in parameters. In the best scientific terms: it's A LOT and I hate them. To be fair, since it was picking up single -'s as well, there's a ton of noise in there as well; a lot seems to be from date ranges.

I've done a lot of 'hmm'ing at numbers based off my own use of Midjourney, though I'm sure anyone with a more stats-y inclination could have the same suspicions, but I'm very surprised that apparently literally not a soul correctly used the —no parameter in this dataset. Which makes me suspicious yet that I still fudged the filtering. But, anecdotally it is the one parameter I've seen people try to use the most and screw up, so that tracks. Midjourney also doesn't throw errors for params with single -'s. It just reads them as a normal token with the rest of the prompt.

```
—no letters font
——seed3388
—1950, scientific illustrations,
—no shadow
—no text
—q 1
—no words
—no humans
—ar 5:4 shading
—Roman wars in 66—135 CE, colorful, with army banner flying in background, highly realisttic,ultra detailed,photo realistic,8k,16k,
—present, rtx, scientific illustrations, realistic marine paintings
—1950
—present raw
—Habsburg wars: The Battle of Lepanto was fought near the Gulf of Corinth, a significant setback for the Ottoman Empire and the last major
naval battle fought entirely with galleys.
——repeat5
—1969, solarizing master, gray and crimson, spherical sculptures. vivid viridian forest green background. mix of vibrant, red, black and wh
ite colours for the shapes. Not so much black
——ar 3:4
—present, ivory, leica cl, chinapunk
—1939 (interwar), devilcore
—present, photo taken with provia, sonian, sculpted, les nabis
——————— green teal black, blue—black and 50 different shades of gray + green, illustration by Bruce Pennington raw
—1950s, full body, light brown and navy, candid, mottled, neo—traditionalist, rounded
—1950s, mottled, natural fibers, elegantly formal, sun—soaked colours, restored and repurposed, photographic weavings
—1950, scientific illustrations, golden sky, beautiful green mountains
—1000 ce, lovecraftian, bentwood, found object
—ar 9:16, Pixar Style, Disney Style
—present, colorful sidewalk scenes, light red and blue, child—like innocence
—present
—ar 3:1
—v 5. 1
—q 3
—1939 (interwar), #vfxfriday, realist detail, anglo gothic
```

This is my hell.

Show Random Prompt

```
Digital painting of a single creature inspired by D&D creature art in full—body, it is a medium sized hypogryph with a mastiff body, eagle h
ead and beak, and lion tail, with bright feathered collar in hues of blue, purple or green, on a white background, in the style of Etrian Od
yssey or hearthstone wings — @DeadbeatHero (Waiting to start)
```

It looks like the scraping bot(s) did pick up some of the in-progress Midjourney bot outputs

Accessible data viz link notes:

https://fionabaudner.medium.com/designing-accessible-charts-39ab0ff546b6

https://www.smashingmagazine.com/2022/07/accessibility-first-approach-chart-visual-design/

https://medium.com/google-design/redefining-data-visualization-at-google-9bdcf2e447c6

https://m2.material.io/design/communication/data-visualization.html