Dan Bailey

ENG 381

Overall Results

Data set generation was, well, not exactly fun. In retrospect, I should’ve used the API (Application Programmatic Interface) that ChatGPT offers, rather than manually entering each of the prompts and then handling the copy-paste. I could’ve written about 30 minutes of code and had this whole thing done in a matter of another 30 minutes of the script running. As it was, I spent the better part of 8 hours doing this – the tedious part being thinking up new and different prompts for ChatGPT to write about.

**Concerns**

1. The dataset was limited by the Chat GPT prompts I could think of to generate. So this is a very Dan-centric list, comprised of a lot of short pieces. This might be offset by the human-written text, which was copied out of my personal ebook library. Again, it’s a Dan-centric dataset, but the blocks of text are much larger.
2. There may be some interesting stuff, as I’ve recently learned that large language models (LLMs) actually have what could be considered “watermarks” in them. (Goldstein, 2023). This could influence the results of the Zipf analysis, but given the dataset site, this shouldn’t be massive.
3. Dataset size is my other concern – 100,000 words each doesn’t feel like enough, but I don’t think I’ll have a definitive feel for that until I run the first pass of the analysis.

**Learnings**

1. ChatGPT can’t count. No, really. I would frequently ask it for a word count associated with the response, and it would be like, “Certainly! Here’s 2000 words on Topic X!” And then I’d get an 800-word reply. This, as you can see from the prompts document, led to having to create a lot of prompts to get to 100,000 words.
2. When I get bored or frustrated with coming up with prompts, I start writing some really weird ones.
3. If you ask ChatGPT to write on the topic of its choice, it frequently writes about the application of AI in healthcare, and it reads almost like a sales brochure.
4. ChatGPT really doesn’t like it when you ask “Why is Elon Musk such a douchebag?” Whether they’re trying to avoid controversy or if this is just an effect of Musk having been an early investor in OpenAI is something that I’m sure I’ll never have a satisfactory answer to.

**Next Steps**

1. There’s some data normalization that I’ll need to undertake – getting the formatting out of the ChatGPT responses, for example. And stripping punctuation out of everything. And then converting the Word docs to ASCII text. Fortunately, all of this can be done programmatically, so that shouldn’t take much more than an hour.
2. Running first pass analysis with Zipf.
3. Wrap up testing on the Shannon Entropy analysis.
4. Set up Github repository for code, in case anyone else finds it useful.

All in all, getting this all together was fascinating, even if there were times where it where it was frustration, and I’m eager to see what the first pass of Zipf gets me.

**References:**

*Watermarking LLMs to Fight Plagiarism with Tom Goldstein – 621,* YouTube, 20 Mar. 2023, <https://www.youtube.com/watch?v-QC6wXsusHjk>. Accessed 15 Oct. 2023.