**Monday 2/5/2024**

* I wrote Python scripts this weekend to get the nonprofit name and mission statement from each Guidestar profile website. The script works only up to a certain point because Guidestar uses what I believe to be anti-scraping measures from Cloudflare that restrict my program from getting their data. Also, one important note is that not all nonprofits actually provide a mission statement to the Guidestar profile.
* This week, I will look into bypassing anti-scraping measures for Guidestar and start writing scripts to get mission statements from other websites. Options include searching the non-US nonprofits (like the Australia or UK ones).
* If I cannot web scrape from Guidestar, buying the data from them may be a good path too.
* Dr. Price found an IRS database of tax-exempt organizations with EINs that will be very useful in future web-scraping scripts. I will focus on scraping data from Charity Navigator first, and I will also try to write a script to go to each actual nonprofit website and get the mission statement from the source.
* <https://www.irs.gov/charities-non-profits/exempt-organizations-business-master-file-extract-eo-bmf>
* This link above leads to a list of all nonprofit EINs and other useful data in spreadsheets.

**Sunday 2/11/2024**

* I figured out how to scrape without being blocked by the website's servers; I made my program slower so that I do not send too many requests at once and overload their servers. I am now able to scrape from GuideStar and Charity Navigator, but at a rate of only about 1 website/second.
* I have the 51 spreadsheets listing all tax-exempt organizations from the states and DC, and I chose to scrape the Wyoming nonprofits first with guidestar\_scraper.py because it only has about 5500 nonprofits (5500 seconds = 1.5 hours). Important to note is that over 3K of the 5.5K nonprofits did not provide GuideStar with a mission statement. I am going to write another Python script to have Charity Navigator fill in any possible mission statement gaps.
* Dr. Price suggests that many of the nonprofits with no mission statements will be smaller or inactive causes.
* He also suggests that I focus on getting data from states with less nonprofits, for now. Some states have over 200K charities (200K seconds = 50+ hours) and I cannot keep my computer on for that long to scrape all of that, I will look into bots, virtual machines, or other computers/options to run the script.

**Monday 2/19/**

* I have been running my web scraper script whenever I have a chance, and I have gathered the data from 9 US states. I added more research to my literature review; this time I added research about linguistic structures and keywords affecting sales/fundraising in marketing. This research can help explain why researching language of mission statements is important to increasing charitable giving. I also applied to The University of Alabama's URCA, an undergraduate research conference; I will present a poster about my research if I am accepted.
* I have not yet looked into getting a virtual machine because I have been busy. I will get one when I get a chance; this will help me to run the script uninterrupted and gather the data faster.
* I am learning NLP techniques with several Python libraries, so far including NLTK, Regex, and spaCy. This will be important for when I start using NLP on the textual data.
* I have scraped the information from tax-exempt organizations from 9 states so far.

**Tuesday 2/20/2024**

* I checked the website with the IRS spreadsheets containing information about the tax-exempt organizations, and it was apparently updated on February 13, 2024, which is after I got the spreadsheets initially.
* I changed my scraper script to get the revenue from each organization from the IRS spreadsheet.
* Because I have updated spreadsheets to scrape from and my script can now get valuable revenue information, I will restart the process of scraping to get all my data again. Basically, I am back at 0 states' tax-exempt organizations scraped.

**Wednesday 3/6/2024**

* I met with Dr. Price yesterday to discuss methodology for the project. We need to find a way to characterize the mission statements as altruistic or donor-serving. Some missions statements use language like 'help you' or 'member-focused' (donor-serving, impure altruism) while others talk about how they will help society (society-serving, pure altruism). I will look at linguistics research papers to find ways to characterize mission statement text as pure or impure altruism. Hopefully there will be a quantifiable metric that I can get for each mission statement, which I can then correlate to revenue and income. Basically, the next step is linguistics research. Dr. Price also suggested that I meet with a Linguistics professor from UA to get insights.

**Friday 3/8/2024**

* Raeed and I met with Dr. Price today. We agreed that we would not worry about getting nonprofit mission statement data from other countries for now; we will process the American data first. Also, I shared the only relevant linguistics research paper that I could find. We will set up a meeting with a linguistics professor to gain more insight after Spring Break. Spring Break is this next week (3/11 to 3/15).

**Wednesday 3/13/2024**

* I created a master spreadsheet with the states that I have so far to run some preliminary data analysis, so I can present some preliminary results on my poster at URCA. The spreadsheet contains the data from Alaska, Alabama, Arkansas, Arizona, Colorado, Connecticut, Delaware, Hawaii, Iowa, Idaho, Kansas, Kentucky, Louisiana, Maine, Minnesota, Mississippi, Montana, North Dakota, Nebraska, New Hampshire, New Mexico, Nevada, Oklahoma, Oregon, Rhode Island, South Caroline, South Dakota, Utah, Vermont, Wisconsin, West Virginia, and Wyoming. After creating the master spreadsheet, I removed rows with no mission statements and no revenues. Then, I will use the Quanteda R package to measure the linguistic richness and linguistic readability of each mission statement, allowing me to find a relationship between those and the revenues. This is not related to the goal of the research, which will ultimately characterize the mission statements as more donor-serving or society-serving. However, it will be good preliminary data to show at URCA to indicate that I have done something.

**Tuesday 3/19/2024**

* I bought and set up a Raspberry Pi, and I am able to access it while my computer is on the same network. For some reason, Cloudflare stops me from running the web scraper on that, so I have to use NordVPN on the Pi. However, as soon as I connect to a VPN, I lose access to the Pi and am disconnected because my device and the Pi are no longer on the same network. I started the Python script, and then I ran a command to connect to a VPN, pause the terminal for 24 hours, and then disconnect the VPN, after which Cloudflare will start to block the webscraping again, but I will be able to access the Pi again to check progress and redo the terminal commands. I will see how that script is going tomorrow.

**Tuesday 3/26/2024**

* I presented my research at URCA, which went well. The judges did not grill me too hard, but I answered their questions well. It was a fun experience.
* My Raspberry Pi has some issues with web scraping, but I will continue to troubleshoot that to fix issues. That will let me generate data faster.
* I compared machine learning frameworks that I could use. There are many useful Python libraries, such as TensorFlow, PyTorch, Keras, MXNet, CNTK, etc. The best things that I found were IBM Watson Natural Language Understanding and the Python libraries TensorFlow and PyTorch. I decided not to use IBM's NLU product because it is only free for about 30K NLU items a month, and we will have at least 500K mission statements to analyze. I will learn TensorFlow first (starting today) and see if that is useful for our purposes before considering PyTorch.