1. Team Name: Max Headroom 1
2. Team Members: Kaya Borlase, Joseph Helbing, Lynette Dang, Isabella Duan + Class Section: Section 2
3. GitHub Repository Link: https://github.com/cs-ssa-w22/final-project-max-headroom-1
4. Description: According to the University of Chicago, “the faculty members of UChicago College are distinguished scholars and scientists renowned for shaping and defining entire fields of study. The faculty members are also infused with a strong sense of pride in their teaching and a profound commitment that their work as educators will have a dramatic impact in defending the fundamental values and the style of intellectual life that defines the University” (UChicago, 2022). How does the university seek out individuals who fit into these distinguished roles? What factors might the university consider “distinguished”? Could this be connected to educational background? The goal of this project is to investigate the faculty hiring practices of the University of Chicago, to answer the question: Where do faculty at the University of Chicago get their education, specifically their doctorate degrees (PhD, MD, JD, etc), before entering their roles at the university? We plan on analyzing the educational background of current faculty at University of Chicago. We will cross-reference this university rankings to see if University of Chicago hires only faculty from certain tiers of educational institutions. We will then compare this metric across two different divisions to determine if our results could potentially be consistent across the university. We will visualize these results using geospatial analysis as well as frequency and network analysis. Using University of Chicago as an example, this project will ultimately explore networks in careers in academia.
5. Data Sources:
    a. UChicago Faculty PhD Background
        i. Scraped educational data from uchicago.edu department websites for the following departments/divisions[1]:
            1. Social Science Division (for sure):

               https://socialsciences.uchicago.edu/directory/all/all/all/faculty

[1] We will choose one department/division out of Division of Biological Sciences or Harris School of Public Policy (depending on whether we want to compare two similar or two different departments). We are leaning towards Biological Sciences Division because this will show the scope of University hirings better

2. AND Harris School of Public Policy (tentatively):

   https://harris.uchicago.edu/directory

3. OR Division of Biological Sciences (tentatively):

   https://biologicalsciences.uchicago.edu/our-faculty

ii. Missing Data[2]: Professors who don't have education listed on their site at all; people who only studied through their Master's. Our plan for the missing data:

1. We check the proportion of missing data compared to collected data.

2. We select a random sample of missing data, manually collect data on what schools professors graduate from and check whether the distribution of missing education data systematically differs from collected data. If the distribution of missing data does not match that of the scraped data, we will manually gather this information. If it does, missing data will be dropped from the dataset.

3. If the missing data is not readily available publicly, due to ethical concerns, those rows will be dropped from the dataset.

4. We put professors who never obtained a doctorate degree but currently hold teaching positions at UChicago, we will create a sub-dataset for them, if time permits.

iii. Intended Sample Size: Around 150~200 faculty per department

iv. Data cleaning: described below in Task Three

b. Worldwide University Rankings, from 2012. Because the faculty at UChicago were hired over the course of many years, we don't believe that having a ranking list from 2012 is an issue[3]. We actually believe that this might be more representative of educational rankings when all faculty were hired rather than just looking closer to the present.

[2] We haven't dealt with missing data yet, the below description is just a tentative plan

[3] While individual schools' rankings might fluctuate, the tier of schools stays stable across years. So we have decided to stick with the 2012 rankings.

6. Tasks/Steps[4]:
    a. Task One: Build a crawler to visit and pull out links where faculty data can be found. This crawler should crawl a department's faculty page and return A dictionary mapping of the Name, Directory Page, Individual Page, and Education background (initialized to nothing to start)
        i. Python Packages Needed: bs4, json, queue, requests[5]
        ii. Team Member(s) Responsible: Joe
    b. Task Two: Crawl each of the gathered links to pull out faculty education data if it is on either the individual page or the directory page. Update the education values in the dictionary with unclean text gathered from these pages
        i. Python Packages Needed: json, requests, bs4
        ii. Team Member(s) Responsible: Kaya
    c. Task Three: Clean scraped educational data. There will be two cases here. Case one will be if the education was in a list form on the website, easily accessible. The second case will be if the education data is found within a paragraph (i.e. "This person received their Master's from Univ1 and their PhD from Univ2). Within this task, we will make sure that the education level is cleaned to say "PhD" rather than all variations (such as "phd", "Ph.D.", "doctorate", etc)
        i. Python Packages Needed: regex, pandas, queue, csv
        ii. Team Member(s) Responsible: Kaya (Case One) and Lynette (Case Two)
    d. Task Four: Gather and Clean University Ranking Data by pulling out only ranking column and University name
        i. Python Packages Needed: regex and json
        ii. Team Member(s) Responsible: Joe

[4] The following steps apply to all departments across UChicago. So far we have completed up to the initial data analysis (task 1-5) for the social science division, the other departments are to be completed.
[5] We have borrowed util.py from

e. Task Five: Perform Initial Data Analysis on the University data, looking at frequency of education backgrounds to get an idea of the distribution of universities where UChicago faculty get their doctorate from
    i. Python Packages Needed: matplotlib, seaborn, graph-tool
    ii. Team Member(s) Responsible: Lynette and Isabella
f. Task Six: Complete analysis of the data: network analysis, etc. (spatial analysis, network analysis, location-based analysis). We would love to conduct network analysis, create interactive plotting (world map VS US map) that allows the audience to play around with the education criteria (tier of school) and output the number of professors having that specific criteria of educational background (frequency), and a map visualization.
    i. Python Packages Needed: graph-tool, geopandas, geoplot, bokeh
    ii. Team Member(s) Responsible: Lynette and Isabella
g. Task Seven: Writeup and Presentation
    i. Python Packages Needed: N/A
    ii. Team Member(s) Responsible: All

We plan to complete crawling and scraping for the other department by the end of this weekend (February 20), data analysis and visualization by the end of next weekend (February 27). Starting from Week 8 (February 28), we will wrap up our analysis, clean up missing data, start preparing for our presentation.