- Name of activity (Homework 3: Python and Web Scraper)
- Your name: Dima Mikhaylov
- Your UVA computing IDs: agp7dp@virginia.edu

Web Scraper Report

This is in support of a web scraper implementation, submitted separately, hosted on https://github.com/allaccountstaken/best_of_Python/blob/main/M4/WebScraper.ipynb.

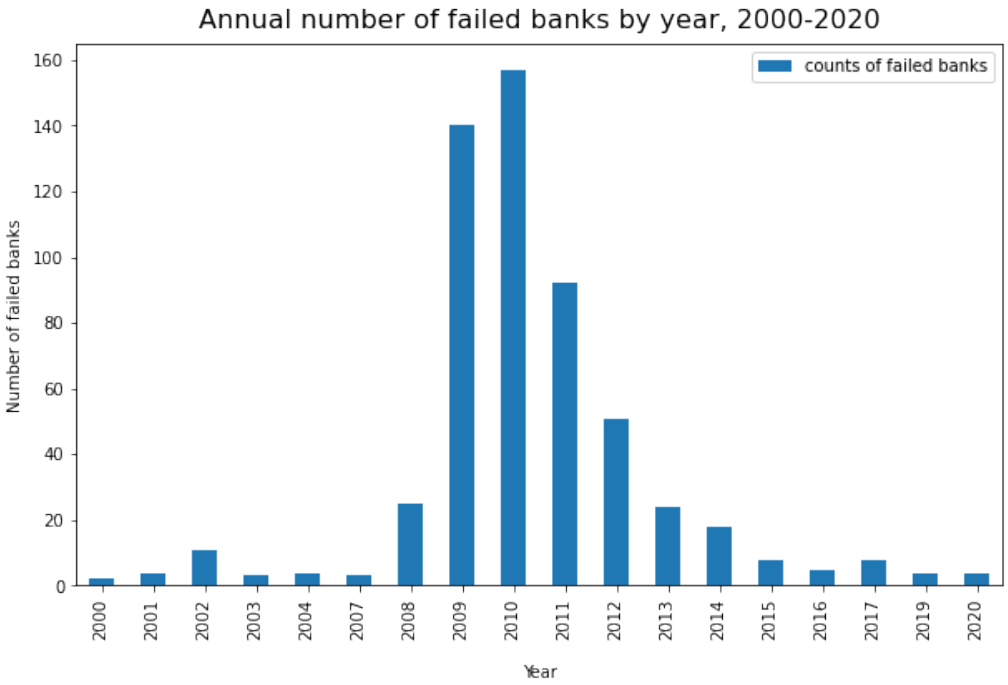
In short, the scraper is implemented in Python 3 on a local Jyputer Notebook using standard **requests** library. HTML web pages are parsed with a **beautifulsoup** parser to obtain tags object that is later used to extract rows and columns from a target table in a list of lists format. The target table is stored in **pandas** `DataFrame` for additional manipulations and analysis. First, it is analyzed for data consistency. Once the acceptable data quality is achieved, the data are written to a back-up `csv` file. Finally, the file is read into a `DataFrame` and used for visual exploratory data analysis with **matplotlib** library; `pyplot` object.

The source of the data is maintained by Federal Deposit Insurance Corporation (FDIC) and made available to public on the following address https://www.fdic.gov/resources/resolutions/bank-failures/failed-bank-list/. This page contains numerous hyperlinks to the agency resources, as well as an HTML table with information about failed banks, i.e. the target table. This table contains information about failed regulated banks going back to the year of 2000. Columns include bank's name, city, state, certificate number, acquiring institution, date, and regulator's fund used to support bank's liquidation. The table has 563 rows, each for one institution failed in the last 20 years.

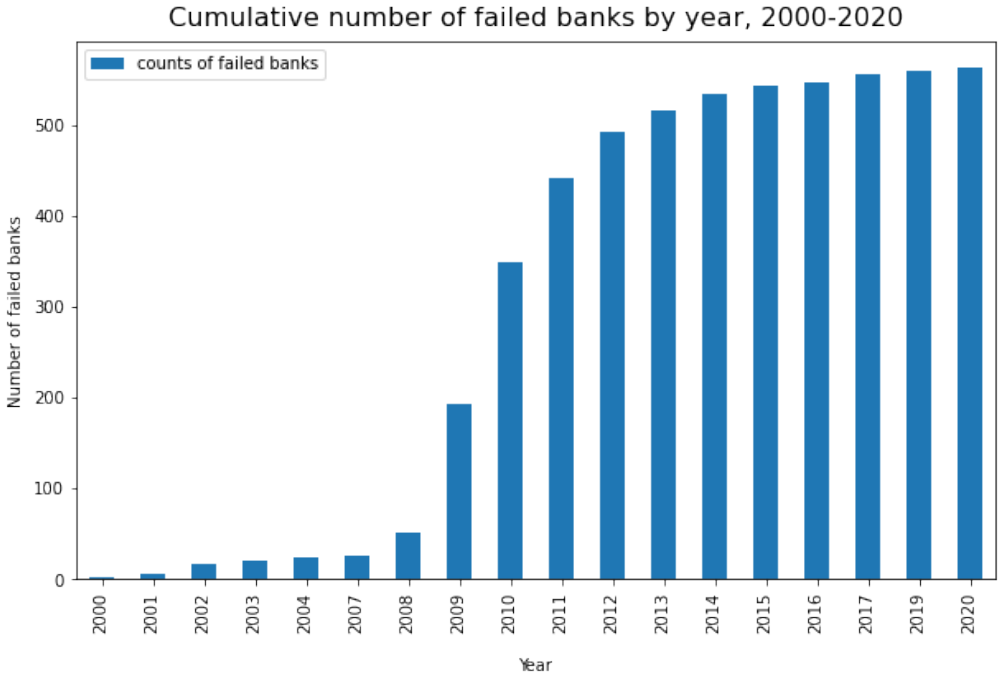
In order to obtain these data programmatically, the scraper connects to the source using **requests** library. Server response is received in a text format and parsed using **beautifulsoup** parser. From a visual examination of the web page source code it appears that the target table is referred to as "dataTables-content-main". Therefore, `soup` find method was used to search for this string. Resulting `match` object is of type `tags` and it can be used to rebuild the table where `<tr>` tags row level records and `<td>` tags columns. Looping through rows first and columns afterwards produces the list of lists, temporary stored in a structure named `data`.

`DataFrame` object can be built in **pandas** directly from the list of lists by specifying corresponding column names. The resulting `DataFrame` has the expected 563 rows of non-null object type records and uses approximately 31 KB of memory. The data can be written to file using `to_csv` method and providing the file name, "dataTables-content-main.csv". When reading the file back into an `analisis_set`, an index is set to the previously created index with `index_col=0`. Furthermore, column date needs to be converted to appropriate `datetime` format using generic pandas method.

The ultimate value of the dataset is that it can be used to produce insights into how many banks failed and when. Original FDIC table is spread along several web pages and does not allow for a big picture view. For example, grouping banks by year of failure and taking counts produces the following chart below. This shows that something truly extraordinary happened in 2009 - 2012.



Alternatively, one can use an additional cumulative summation of annual counts to see how the total number of failed institutions accumulated through the last 20 years. From the chart below, it is obvious that most of these 563 instances come from the years after 2009 but before 2013 when the speed of accumulations slows down significantly.



Additional insights could be produced by scraping other FDIC web pages. For instance, there are tables with resolution costs breakdown that show how much money is being spent on different restructuring transactions. These data, if merged by bank name or certificate number, can be further analyzed, for example, on quarterly or even monthly level.

Generally speaking, the banking industry is going through a rapid consolidation and the overall number of banks has decreased from 8,500 down to 5,000 in the last 20 years. From the scraped data it appears that only roughly 500 banks actually failed and most of these failures were recorded in the aftermath of a severe economic downturn due to subprime mortgage crisis. This could be a starting point for in-depth analysis of what happened to other 3,500 institutions that have not officially failed, but yet are not present today due to charter changes, voluntary liquidations, mergers and acquisitions. For example, one possible research question could be if these 3,500 institutions share common features with 500 failed banks? Were they likely to fail in 2009 - 2012, but survived? Did they receive bailouts or experienced significant market share decline during these years?