Tasks of the program:

1. scrape the site with Selenium for .pdf statements of Polish nonprofits
2. use OCRmyPDF to extract text from the PDFs
3. use string manipulation, find and store the interesting values
4. put the data on Excel and format it
5. write a information to people how should they interpret the data
6. put the Excel online

Prerequisits:

I strongly suggest that you download a package manager and use that. I use Chocolatey installed on a cleans system, and my problems with managing packages are since gone.

If you canot install it with you package manager, then use standard -pip install command.

You need to install:

* Chocolatey package manager (or any other package manager)
* Selenium
* download and use right verion of Selenium Web Driver
* OCRmyPDF (for one of its dependencies I had to go with "pip install ocrmypdf --user" to install that)
* Polish language pack from tesseract library, on which OCRmyPDF operates

Future enhancements:

1. The whole proces takes around 17 days to complete. Currently, it takes:
   * 2 days to download the .pdfs (around 40 000). This is mostly because of how the script is written, me using Selenium and the sheer volume of files to be downloaded.
   * 15 days to extract the text from the .pdfs. Is it that I am doing this on a laptop, or can it be done faster?
2. What other columns can I add?
   * I could show the biggest monthly salary for the employee / GEB member
   * Total spent on GEB vs total spent on employees
3. Get the name of the nonprofit using its KRS number with <https://ekrs.ms.gov.pl/>
4. Fix the issues with the column name:
   * Remove duplicates from the name of non-profit
   * Check you way of dealing with quote mark
   * One name for non-profit that has 2-3 different letters in the name

What I learnt:

1. use the right tools for the job. Initially I wanted to look for words with regex; that would do the trick, since being more powerful it also has bigger lerning curve. As a consequence, I stumbled for some time before coming back to native python function, which did the job (find(), isalpha(),isnumeric())
2. consider your options. I started knew about Beautiful Soup, and decided to go with Selenium not understanding the implications of my decision. I managed to succeed with Selenium – but perhaps if I used BS I could spend less time on it?
3. Use small amounts of code to test things, when things don’t work as you expect them check online for solutions. Consider what are your different options that you can use to achive this goal. When something does not work with one tool, don’t be afraid to change the tool.
4. StackOverflow is your friend – and internet in general too. Check it out! When you stumble upon something (and again and again), post a question online. Rule is – the better described the question the more likely you will get a satisfactory answer.

Program specifics:

Whenever you see the KRS number instead of the name of the nonprofit, that means the name of the nonprofit was not extracted from the .pdf file by the OCRmyPDF. There are several things that I could try here – use the KRS numer to get the name from the registry of KRS numbers, use another OCR tool, fill in the missing names manually.