Add taxonomic name resolution to the EcoData Retriever to facilitate data science approaches to ecology

Abstract

The EcoData Retriever is an automated software that is used for finding, downloading, and cleaning up ecological data files, and then organizes the whole data in a local database of your choice. The program cleans and restructures datasets in normalized form and formats these datasets according to the suitable standards of the database management systems.

This project deals with handling the problem of resolving scientific names that are stored in appropriate manner across various datasets. This software would enable it's users to conveniently combine different datasets that will be helpful in answering scientific queries.

Technical Details

EcoData Retriever is python based tool for downloading, cleaning up, and restructuring ecological data. The information about the repositories are stored in .script files which contain URLs to the files containing the datasets.

The Challenges which we are facing currently is that, the name of species are frequently redefined which makes it inappropriate to maintain consistency across the datasets and so it becomes difficult to combine multiple datasets together.

This project would include a system for sending a species name to one or more of the taxonomic name resolution services and determine the best name to appear in the generated databases. The project is composed of the following main components:

- Accessing the resolution services.
- Determining the selection strategy.
- Updating the data model.
- The user interface and the control flow.

The download operation is performed through Engine class, which uses urllib library to fetch the dataset file, and store it locally. This operation happens inside download_file() function in ./lib/engine.py file.

The general problem that we will be dealing would be to maintain the accuracy of datasets for which the names of species are redefined that raises the difficulty in combining the multiple datasets together.

Schedule of Deliverables

This time line has been made keeping in mind that it will require work at least 40 hours per week. I am not planning on taking any vacations during the summer and would dedicatedly work on the project and also, I will continue to contribute afterwards.

Before May 25th

- I would utilize this time to interact with the EcoData Retriever Community and get more familiar with the Retriever codebase and do sample tests.
- Determine possible advancements, selection strategy, discuss possibilities and issues.

May 25th - June 7th

Handling Resource Description Framework Schema (RDFs) for testing the XML communications and also test different Taxonomic name resolution services.

June 8th - June 21th

Add functions and writing advance scripts to access more name resolution services in pytaxize.

June 22nd - July 5th

- Test added functions in pytaxize and scripts.
- Write and test the module in Retriever to query names through database.
- Build the control flow for taxonomic name resolution

July 6th - July 19th

- Test on some datasets to evaluate the results with some updates of the selection strategy.
- Test for the control flow for taxonomic name resolution for the desired output format.

July 20th - August 2nd

Perform final tests and check if the current implementations works for all the dataset, introduce more parameters if required

August 3rd - August 16th

- Work on open issues in the retriever like adding new datasets and fixing bugs.
- Work on pytaxize to make it more robust.
- Make any changes needed in Graphical User Interface (GUI) to incorporate this utility. Finalize the code and take feedback from the mentor.

August 17th - August 21th 19:00 UTC

- Week to scrub code, write tests, improve documentation, etc.
- Code polish and merge back to the master.
- Submit final evaluation.

Future works

- I will continue to work on the Retriever Project and finish some feature requests and I will be involved in fixing some bugs in Retriever.
- Adding more reliable name resolution APIs.
- Improving the accuracy of the best name deciding algorithm.
- Bug fixing the existing and adding more engines for different formats of data. For example - JSON, XML etc.

Open Source Development Experience

I have worked on few open source projects like Google's tensorflow https://github.com/tensorflow/tensorflow (https://github.com/tensorflow/tensorflow) which involves computation using data flow graphs for scalable machine learning. I also have contribution in fixing bugs in Mozilla community at bugzilla and also contributed as Firefox Student Ambassador as well as on Free Code Camp https://github.com/FreeCodeCamp/FreeCodeCamp/issues/6126#issuecomment-182170155 (https://github.com/FreeCodeCamp/FreeCodeCamp/issues/6126#issuecomment-182170155) .

I have also worked with a start-up from IIT Kanpur, SIDBI Incubation Center, (India) which involved the development of Dynamic UI graph using plot.ly API for Python and Java Script for testing and plotting the large experimental data sets.

I have been the member of NumFOCUS community and supporting it as a volunteer.

Academic Experience

I am currently a 3rd year (B.Tech) Computer Science and Engineering student at University Institute of Engineering and Technology, Kanpur, India. I have undertaken courses in Data Structures, Algorithm Design and Analysis, Database Management Systems (DBMS), Discrete Mathematics, Compiler Design, Software Engineering, Machine Learning and other varied interests. I have done some academic projects on database management systems. I am a Free Open Source Software (FOSS) enthusiast and like to contribute, collaborate for the same.

This project deals with implementing and automating the process of accessing the data with the taxonomic names resolved. The project requires the knowledge of Python and database systems like MySql, SqLite. I have worked on similar projects that had the same requirements and skill set. https://github.com/raj-maurya/Database-Frontend-backend-Mysql-Project (https://github.com/raj-maurya/Database-Frontend-backend-Mysql-Project)

I also have interned at some research organizations like:

- **Indian Institute of Science Education Research (IISER) Kolkata:**

  I have been actively involved in research and developmental activities since my freshman and sophomore year and have worked as a researcher at Indian Institute of Science Education and Research (IISER) Kolkata, India (2015). At IISER, I have worked on research module of "Time Series Analysis and Prediction", it included the study of various mathematical tools and time series models like ARCH, GARC, TAR and ANN for the analysis of financial as well as the biological experimental data consisting of structured, unstructured data and can be suitable for a range of varied applications like monitoring health of patients, daily price variation, stock market volatility, weather prediction, growth rate of agricultural sector and their market prediction. The analysis of non- linear models including Artificial Neural Network (ANN), Haar functions and Genetic Algorithms for a better understanding and good approach for the non-linear time series and solutions for minimizing the risk of economic bubble burst. Also worked on "Epilepsy Prediction" and "Wavelet analysis".

- **United Nations Educational, Scientific and Cultural Organization (UNESCO):**

  At UNESCO, I was exposed to various approaches and methods to resolve the issues and associate the solutions for the government using cloud computing resources available so, as it becomes easier for the government to interact with its citizen for the smooth running of the whole organization.

Apart from these I have also interned at start-up companies which includes:

- **Absentia Virtual Reality:**

  Data Analytics Internship.

- **Kanopy Techno Solutions:**

  Software Developer Internship.

- **Pollinate Energy:**

  Actively involved with product research and development team.

- **Pariksha.co:**

  E-Learning content developer/curator using LaTex.

Why this project?

While going through the list of organizations and their ideas pages, I was actually looking for the projects which involves the extensive use of database systems, manipulating the large datasets accordingly and I found this project more suitable as well as favorable to me as it involves the usage of almost 80% of the skill sets and technologies that I am already familiar with. Also, my academic as well as non academic projects involved the processing of large data sets using sophisticated database systems.

I want to be involved in this project because it could help scientists and researchers to do their research and will save their precious time. I will be glad to contribute to this project and help others through this open source projects. This project will be a stepping stone for expanding my experience, skills and abilities as well.

Appendix

Contributions and ideas

Concurrent access of the SQLite database: https://github.com/numfocus/gsoc/issues/120 (https://github.com/numfocus/gsoc/issues/120)

References

- https://github.com/weecology/retriever (https://github.com/weecology/retriever)
- http://retriever.readthedocs.org/en/latest/scripts.html (http://retriever.readthedocs.org/en/latest/scripts.html)

Contact Details

- Name: Raj Kumar Maurya
- Email: rajkmaurya111@gmail.com
- GitHub: raj-maurya
- Location: India
- Web Presence: https://rajkmauryablog.wordpress.com/ (https://rajkmauryablog.wordpress.com/)
- Mentors: @ethanwhite @henrysenyondo @sckott