Distinguishing real and artificial information

Python Software Foundation (Xbitinfo)

About Me

- Name - Ishaan Jain
- GitHub - Ishaanj18
- Email - mailto:ishaan454@gmail.com
- University - Manipal University Jaipur.
- Program - Bachelor of Technology in Information Technology
- Year - 3rd year
- Timezone - Indian Standard Time (GMT +5:30)

Background

As someone who loves solving problems, I have developed a passion for computer science and software development. I am particularly interested in data structures and algorithms, which I believe are essential for developing efficient and scalable software systems.

In addition to my interest in data structures and algorithms, I have also developed a keen interest in web development. I believe that web development is an exciting field that offers endless possibilities for innovation and creativity. Over the past year, I have gained expertise in various web development skills, including HTML, CSS, JavaScript, and the MERN stack. I have successfully built several full-stack web applications.

In addition to my web development skills, I have developed Python proficiency through my coursework and personal projects. As a tech enthusiast, I am constantly exploring new ways to improve my skills and expand my knowledge. I am excited about the opportunities that lie ahead in the field of technology, and I am eager to contribute to innovative projects that make a meaningful impact.

Logistics

- ➔ The GSoC timeline is in sync with my university's summer break (June end) and thus it will give me enough time to work on this project. I understand that GSoC is equivalent to a full-time program and hence I plan to devote at least 30-35 hours a week to this project.
- ➔ My university reopens in mid-August and even if some part of the timeline coincides there will be no exams in that period, allowing me to devote ample time to work on and complete this project in the stipulated time frame. I am excited to spend my summer working on this project!

Issues Resolved

S No.	Contribution	Issue	PR
1	ci: change linters to use specific ubuntu instead of unbuntu-latest	#2877	#2879
2	Doc: pooch install for xarray dataset	#184	#185

Project Information

- Organisation Name - Python Software Foundation.
- Sub-Organisation Name - Xbitinfo
- Mentor - Milan Klöwer , Hauke Schulz
- Project Size - Large (350 hrs)

Project Abstract

- This project aims to develop an information theoretic approach to filter out artificial information from real information in geospatial datasets. The bitinformation framework is used to distinguish between real and false information, where real information is defined as the mutual information between adjacent bits, and false information is the difference between entropy and real information. The proposed algorithm will filter out artificial information, which is a consequence of prior compression, from real information, which reflects the underlying signal. The project will involve a theoretical review, test case generation, algorithm development, evaluation, and integration into xbitinfo, a software package that implements the bitinformation framework. The resulting artificial information filter will provide a valuable tool for users who may not have access to high-precision/uncompressed data, and it will enhance the accuracy and reliability of geospatial data analysis.

Detailed Objectives

➔ The implementation of the proposed information theoretic approach to filter out artificial information from real information in geospatial datasets would involve a combination of theoretical analysis, algorithm development, and software integration. The resulting filter would be a valuable addition to xbitinfo, enabling researchers and practitioners to distinguish real and artificial information accurately and efficiently in geospatial datasets.

- Theoretical Review: The first step would be to conduct a theoretical review of the existing approaches to distinguish real and artificial information in bitstreams. This would involve a thorough examination of the bitinformation framework and related concepts such as mutual information, entropy, and lossy compression. The review would also explore the challenges and limitations of existing approaches in the context of geospatial datasets.
- Test Cases: The next step would be to develop simple test cases to evaluate the performance of the bitinformation framework in distinguishing real and artificial information in geospatial datasets. These test cases would involve generating synthetic geospatial datasets with varying levels of lossy compression and quantization. The bitstreams generated from these datasets would then be analyzed using the bitinformation framework, and the accuracy of the framework in distinguishing real and artificial information would be evaluated.
- Algorithm Development: Based on the theoretical review and test results, we would then develop new algorithms or refine existing ones to filter out artificial information from real information in geospatial datasets. These algorithms would be based on information theoretic principles and aim to optimize the trade-off between filtering accuracy and computational efficiency. Possible algorithmic approaches include:
  - Threshold-based filtering: This approach would involve setting a threshold value for the mutual information between adjacent bits below which the information would be classified as artificial.
  - Spectral analysis-based filtering: This approach would involve analyzing the frequency spectrum of the bitstream and identifying frequency components that correspond to artificial information.
  - Machine learning-based filtering: This approach would involve training a machine learning model on a set of labeled bitstreams to distinguish between real and artificial information.
- Evaluation: To evaluate the performance of the proposed algorithms, we would use a range of geospatial datasets with varying levels of lossy compression and

quantization. The datasets would be preprocessed to extract the bitstreams, which would then be filtered using the proposed algorithms. The filtering accuracy and computational efficiency of the algorithms would be compared with the existing approaches, including the bitinformation framework. The evaluation would also include a sensitivity analysis to examine the impact of different parameters and assumptions on the filtering accuracy.

- ➔ Integration: Finally, we would integrate the proposed artificial information filter into xbitinfo, a software package that implements the bitinformation framework. The filter would be switched on by default, and users would have the option to turn it off if needed. We would also provide user documentation and a tutorial on how to use the new filter. The integration would involve modifying the existing xbitinfo codebase to incorporate the new filter and testing the modified software thoroughly.

Weekly Timeline

Timeline	Work
Pre GSoC	➔ Contribute to the project by working on issues to further improve my understanding of the codebase. ➔ Interact with the mentors and other contributors.
Community Bonding Period (May 4- May 28)	➔ Bonding with the community actively. ➔ Discussing and refining the project idea with the help of the community and the mentors.

1st Week and 2nd Week (May 29- June 11)	➔ Conduct a thorough review of the bitinformation framework and related concepts such as mutual information, entropy, and lossy compression. ➔ Examine the challenges and limitations of existing approaches in the context of geospatial datasets. ➔ Generate synthetic geospatial datasets with varying levels of lossy compression and quantization.
3rd Week and 4th Week (June 12- June 25)	➔ Develop a plan for the test cases. ➔ Analyze the bitstreams generated from the datasets using the bitinformation framework. ➔ Evaluate the accuracy of the framework in distinguishing real and artificial information.
5th Week and 6th Week (June 26- July 9)	➔ Develop new algorithms or refine existing ones to filter out artificial information from real information in geospatial datasets. ➔ Optimize the trade-off between filtering accuracy and computational efficiency. ➔ Possible algorithmic approaches include threshold-based filtering, spectral analysis-based filtering, and machine learning-based filtering.

7th Week and 8th Week (July 10- July 23)	➔ Use a range of geospatial datasets with varying levels of lossy compression and quantization to evaluate the performance of the proposed algorithms. ➔ Compare the filtering accuracy and computational efficiency of the algorithms with the existing approaches, including the bitinformation framework. ➔ Conduct a sensitivity analysis to examine the impact of different parameters and assumptions on the filtering accuracy.
(July 10- July 14) Mid-term evaluation	➔ Submitting mid-term evaluations.
9th Week and 10th Week (July 24 - August 6)	➔ Integrate the proposed artificial information filter into xbitinfo, a software package that implements the bitinformation framework. ➔ Modify the existing xbitinfo codebase to incorporate the new filter and test the modified software thoroughly. ➔ Provide user documentation and a tutorial on how to use the new filter.
11th Week and 12th Week (August 7- August 20)	➔ Finalize the proposed algorithms and the integration of the new filter into xbitinfo.

	➔ Prepare a report summarizing the theoretical review, test cases, algorithm development, evaluation, and integration. ➔ Prepare a presentation summarizing the key findings and recommendations.
Final Week (August 21 - August 28)	➔ Buffer week. ➔ Fix bugs if there are any. ➔ Improve documentation where they seem to be lacking.
Post GSoC	➔ Continue contributing to the project by fixing issues and adding new enhancements.

Eligibility

➔ Yes, I'm eligible for Google Summer of Code 2023

Are you applying for other Projects?

➔ No, I am applying only for this project.

Communication

➔ You can reach me via email, or phone calls.
➔ I will provide frequent updates to my mentor(s) regarding the project's advancement and any obstacles I may face.
➔ Weekly blog posts will be written by me to document my project's progress.