GSoC’21 Proposal

Dimensionality Reduction for Studying Diffuse Circumgalactic Medium

Kshitij Soni (kshitijsoni@iipe.ac.in)

April 12 2021

Abstract

The goal of this project is to implement the machine learning dimensionality reduction techniques for the dataset of stimulated quasar absorption spectra such that the reduced dimensional data will maintain a high level of accuracy compared to the original dataset. The gas surrounding galaxies outside their disks or ISM and inside their virial radii is known as the circumgalactic medium" (CGM). The CGM is a source for a galaxy's star-forming fuel, the venue for galactic feedback and recycling, and perhaps the key regulator of the galactic gas supply. Observations from all redshifts and from across the electromagnetic spectrum indicate that CGM gas has a key role in galaxy evolution. The history of the CGM can be explored by developing computational and machine learning tools to use quasar absorption lines to determine the gas composition, temperature and density.

1. Background

Significant progress has been made in the past decade or so in observing the cosmic-ray/B-field, as well as various phases of the CGM. [1] Gives the information about how are energy, mass and metal contents spatially distributed and partitioned in the different components. Moreover, how are they linked to properties of host galaxies and their global clustering and intergalactic medium environments. Lastly, what are the origin, state, and life-cycle of the CGM. These questions have been addressed with multi-wavelength observations of the CGM. [2] focus on understanding the physical properties of extended Lyman-alpha emitting gas at high redshift, from the circumgalactic medium (CGM) and intergalactic medium (IGM). [3] gives information about the hydrodynamical simulations of the massive dark matter halos which host luminous quasars under predict the amount of cool gas observed in quasar environs by a large factor, challenging our understanding of how massive galaxies form.

2. Methodology

The proposed methodologies are

1. Principal Component Analysis
2. Generalized Discriminant Analysis
3. Linear Discriminant Analysis
4. t- SNE Algorithm
5. Multidimensional Scaling

The significant characteristic of our data is volume. The main challenge encountered in dimensionality reduction will be that it will not lose much of the vital aspects of the primary dataset. The accuracy of the model frequently improves by using more main components and does not compensate for the temporal task derived from such improvement. During the project timelines, different algorithms will be implemented and it is expected that nearly 70% accuracy will be obtained compared with the primary dataset.

3. Timeline

Community Bonding Period (May 17 - June 6)

- Introduce myself and this project in the ML4Sci and GSoC mailing list.
- Remain constant touch with my mentors using Gmail. Set up the requirement and discuss algorithm details with the mentors.
- Discuss with mentors about the implementation plan.
- Study more research papers, blog, articles, conference papers on Diffuse Circumgalactic Medium.
- Set up the dev environment and my blog page for TODO list and weekly reports.

Official Coding Period (June 7 - August 16)

Week 1 (June 7 - June 14)

- Understanding the dataset and its components
- Starting with PCA Algorithm, will develop the mathematical model and begin coding.
- Split the data into training set, validation set and testing set

Week 2 - 3 (June 15- June 28)

- Code functions for PCA Algorithm
- Implement PCA Algorithm with the training set and then validate it
- Test the Algorithm with testing set, plot ROC Curve, calculate precision and recall
- Bug fixes if any

Week 4 - 5 (June 29 - July 11)

- Implement the t-SNE Algorithm
- Comparing t-SNE and PCA
- Testing both the algorithms by changing the train test ratio for the dataset
- Again comparing the results with the test data

***First Evaluation Period (July 12 - July 16)**

- Deliver the code and test results
- Submit First Evaluation Report
- Update Github Repo

Week 6 - 7 (July 17 - July 31)

- Discuss with the mentors for multidimensional scaling (MDS)
- Code multidimensional scaling and test with the testing dataset
- Visualize correlation matrix for MDS
- Comparing it with PCA and t-SNE. Split the dataset into a new train test ratio. Implement the code, test again and compare all the three

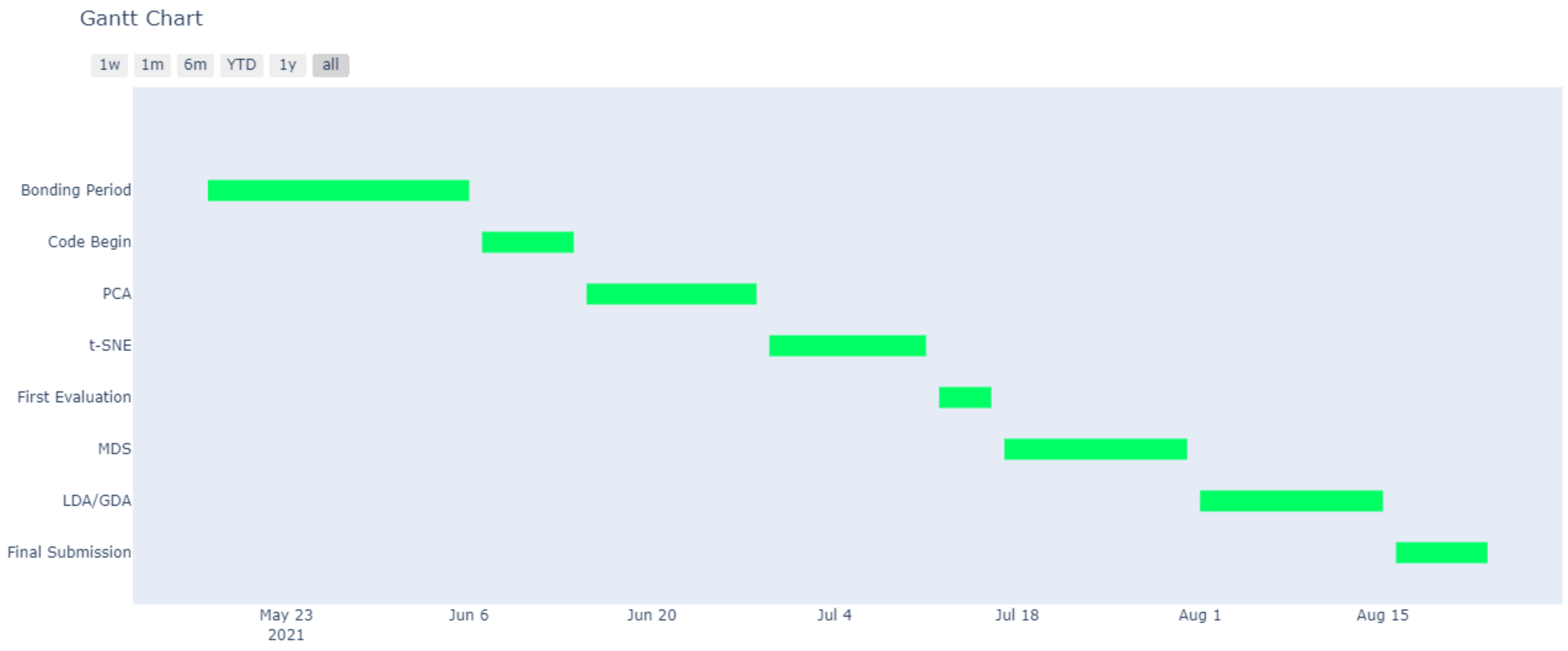
Week 8 - 9 (August 1- August 15)

- Not having previous experience with LDA or GDA, will discuss with the mentors and choose anyone.
- Learn and code the algorithm. Implement it with any other dataset
- Set the environment for the large dataset and implement the same for the given dataset
- Test the code with the testing set
- Compare the PCA, MDS, t-SNE, LDA/GDA
- Change the train-test ratio. Implement all four algorithms, test with the new test data and compare the accuracy, precision and recall for all the four algorithms.
- Pull request for code review and merge
- Buffer time for unexpected delay

***Final Evaluation Period (August 14 - August 22)**

- Deliver all the code, pdf, notebook with results
- Discuss with the mentors and consider the remark provided. Find errors if any and fix them.
- Wrap up the project. Submit the Final Project Report along with the code and results.
- **Submit the final evaluation of my mentors.**

Figure No. 1



4. References

[1]. Q. D. Wang et al., “The panchromatic circumgalactic medium,” *Decadal Survey Science White Paper.*

[2]. Fabrizio, A.B.(2015). *Characterizing the Circumgalactic Medium in Emission.* Ruperto-Carola-University of Heidelberg, Germany.

[3]. Hennawi, J. (2014). The Circumgalactic Medium of Quasars. *Proceedings of the International Astronomical Union 9(S304):355*. https://doi:10.1017/S174392131400430X

5. GSoC Participation

This is my first time to apply for Google Summer of Code. I did not submit a proposal to any other organisation.