Capstone Project Plan Preliminary

Last updated: November 3, 2023

Stakeholder Names and Roles

Stakeholder	Role
Noah McIntire	*Team member*
Ryan Lipps	*Team member*
Brendan Puglisi	*Team member*
Arnav Boppudi	*Team member*
Antonios Mamalakis	*Mentor*
Adele Plunkett	*Sponsor Role*

Abstract

The ALMA Observatory in Chile is the largest astronomical project in existence, and open to anyone in the world to use. The aim of this project is to streamline the proposal process for astronomers who use ALMA. The Capstone team will develop an algorithm that suggests the optimal technical setup based on the text that an astronomer specifies in the proposal, specifically the title and abstract. More than 4000 projects have been undertaken in the first 10 years of ALMA operations, and all projects are visible (with all metadata, and about 90% of the scientific data available) via the user interface of the ALMA Science Archive: almascience.nrao.edu/aq/ The metadata for these projects can be used to train and test the model. Students selecting this project will develop skills in NLP and data mining.

Outline of the Project

The aim of this project is to streamline the proposal process for astronomers who use ALMA. The Capstone team will develop an algorithm that suggests the optimal technical setup based on the text that an astronomer specifies in the proposal, specifically the title and abstract. More than 4000 projects have been undertaken in the first 10 years of ALMA operations, and all projects are visible (with all metadata, and about 90% of the scientific data available) via the user interface of the ALMA Science Archive: almascience.nrao.edu/aq/ The metadata for these projects can be used to train and test the model. **The primary goal will be to optimize the frequency ranges to be observed simultaneously in a project. Additional technical setup requirements can be explored, if there is time and interest among the group.**

As astronomers can spend weeks preparing these proposals, and we suspect that the technical setup is not always optimized, important outcomes of this project include an immense time-savings for astronomers around the world, as well as a more efficient usage of the ALMA telescope for observations of the Universe.

The stakeholders in this project include sponsor Adele Plunkett, the NRAO, ALMA Observatory, astronomers utilizing ALMA, UVA capstone students.

Success Criteria

SC1	An algorithm that suggests the optimal technical setup based on the text that an astronomer specifies in the proposal, specifically the title and abstract. The primary goal will be to optimize the frequency ranges to be observed simultaneously in a project.
SC2	Github Repo housing code
SC3	Algorithm and supporting code organized into python package
SC4	**Reach goal** Add to or slot into existing ALminer functionality
SC5	**Reach goal** for algorithm — suggest alternative data to collect within recommended frequency band and array settings

Assumptions and Limitations

Identifier	**Description**
[L]	There is no set or efficient way to measure the accuracy or success of our model, unless we do not train it on all available data.
[L]	Text and abstract are not sufficient for model training, this may require full proposal text for better accuracy, however, proposal text is not publicly available.
[A]	The project data can be "labeled" in order to cluster/develop a model to suggest frequency setups.
[A]	Additional metadata, e.g. keyword, can be used to assist model performance, keyword data only available at proposal submission. This is to say we will not have these keywords immediately, but maybe we can ask for them during proposal submission.
[A]	There is not one single "correct" tuning per proposal — multiple tunings could work depending on researchers' preferences.

Potential Background Literature and Resources

List readings / repositories / etc. that might help with the project.

- ALMA Science Archive
- ALMA Proposal Preparation (start at Slide 36 for spectral setup)
- ALMA Proposer's Guide: Spectral capabilities
- ALMA Science Archive Users Manual
- ALMiner Tool (for programmatic access of ALMA data/metadata)
- End-to-End Multiclass Text Classification (NLP)

Brief Outline of the Data

The data has been collected over the course of 10 years, from the use of the ALMA observatory. It has been sourced from previous studies utilizing the antennas. This includes the abstract of the experiments and possibly the full text if needed. Along with the positioning of the antennas (ex: right ascension, declination, and band focus).

Target variables:

- Document feature extraction
- Array settings and frequency ranges

Our model will categorize research proposals in two steps:

- A language model which extracts features from the research proposal
- A clustering model to group similar features from the language model and recommend optimal or frequently used “settings” for accurate positioning of antennas and to optimize frequency support

Raw Data:

- Abstract and titles of studies, keywords, Positioning of Antennas (Right Ascension, Declination, Band, Continuum Sensitivity, Frequency Support, etc.)

Potential predictor variables:

- Project Title and abstract (possibly full project text)

Possible challenges:

- Somewhat small sample of project abstracts/ titles, limiting train/test opportunities

Brief Plan of How the Data will be Modeled and Processed

Project Preliminary Timeline

Project Stage	Due Date	Owner(s)
Literature review	January Semester Start?	All capstone members
Presentation/tutorial on data and use from Dr. Plunkett	November?	
Initial NLP pipeline (cleaning, tokenizing, lemmaization, etc.)	2/7/24	
Initial NLP model	2/7/24	

Initial classification model	2/7/24	
SIEDS proposal	2/7/24	
Model improvements and other architecture exploration	3/31/24	
SIEDS paper due	4/7/24	
Prototype for next ALMA research proposal deadline	3/15/24 - 4/15/24	
Comprehensive Testing/ Evaluation of model(s)	3/15/24 - 4/15/24	
SIEDS	4/27 - 4/29	
Python Packaging and Github Repo	4/30/24	

All the data is easily accessible via https://almascience.nrao.edu/aq/, and can be queried into a jupyter notebook or python script using the ALminer package (https://alminer.readthedocs.io/en/latest/).

Brief Plan of Modeling Approaches

The project will require implementing and deploying a language model, using simple python for initial models with the possibility of moving on to more complicated packages and implementations later on. This is a new topic for all of us, so this will require a review of current literature. The second half of our project will involve using NLP as a sort of dimensionality reduction to certain features, which we can then use in a clustering type approach to develop ranges for our response variables (i.e. *Right Ascension, Declination, Band, Continuum Sensitivity, Frequency Support, etc.)*.

Based on modeling success, we may have to consider spectral tunings as a group versus individual rank-ordered or preference recommendations.

Potential Concerns [C] and Blockers [B]

Identifier	Description
[C]	No immediate experience or knowledge of NLP
[C]	Technical knowledge of frequency setups, general knowledge of data and proposals.

[B]	Archives go down periodically. There are three mirrors of the archives, so we may have to change which server we are sourcing data from. This amounts to changing the URL used in ALminer or other data accessing platform. Max expected downtime is 1 hour.