PROJECT REPORT :

Multi-Label Classification of Scientific Literature Using the NASA SciX Corpus

OURDANE AYOUB ELSS
IPSA
Ma412 – Mathematical Foundations for Statistical Learning

Table of Contents

Abstract

This project tackles the challenge of **multi-label** text classification applied to scientific literature from the NASA SciX corpus, which includes thousands of research paper titles and abstracts. The aim is to develop a machine learning system capable of automatically assigning multiple relevant **keywords** to each document, thereby enhancing metadata generation and information retrieval in scientific databases. Unlike single-label classification, this approach captures the multidimensional nature of scientific texts more accurately. A key challenge lies in enabling the system to recognize and label topics that vary significantly in **complexity** and **frequency**, requiring models that can generalize well across both common and rare scientific terms. The process begins with comprehensive data preprocessing, including **tokenization**, **stopword** removal, and vectorization using the TF-IDF technique. Labels are binarized to handle the multi-label nature of the task. Several machine learning algorithms are implemented and evaluated using key metrics such as **precision**, **recall**, and **F1-score**. Among the tested models—**Naïve Bayes**, **Neural Networks**, and **Decision Trees**—the Decision Tree classifier, implemented with scikit-learn's MultiOutputClassifier, demonstrated the most balanced performance. The project relies on Python and key libraries including scikit-learn, NLTK, matplotlib, and Hugging Face's datasets library. The final model offers a scalable and interpretable approach to automatic keyword tagging, highlighting both the strengths and limitations of classical algorithms in handling large-scale multi-label classification problems.

Introduction

In today's data-driven world, classification plays a crucial role in organizing and interpreting information to optimize workflows and support informed decision-making. While humans intuitively classify data in everyday life, large-scale problems require sophisticated computational tools capable of handling high volumes and complex relationships within the data. One such challenge in machine learning is **multi-label classification**, where each instance can be assigned **multiple labels** rather than a single category. This approach is especially useful when analyzing scientific literature, as research documents often cover several interconnected topics.

This project focuses on developing a system capable of automatically predicting relevant keywords—such as *solar wind*or *lunar composition*—from the **titles and abstracts** of scientific papers. By applying multi-label text classification techniques, the goal is to train models that can detect associations between a document's content and its corresponding scientific themes. Unlike single-label classification, which restricts documents to one category, this approach enables a richer and more accurate representation of each paper's subject matter.

The dataset used in this project is provided by the **NASA Astrophysics Data System (ADS)** and hosted on **Hugging Face**. It comprises the **SciX corpus**, a collection of publications spanning **astronomy, astrophysics, physics, and earth sciences**, annotated using the **Unified Astronomy Thesaurus (UAT)**. The corpus includes **18,677 training samples** and **3,025 test samples**, each containing a paper's title, abstract, and a set of verified UAT keyword labels.

Automatic keyword labeling is critical for efficient indexing and retrieval of scientific content. It reduces the need for manual curation and enhances the search experience on platforms like SciX. In this project, we implement and compare several machine learning models—including Decision Trees, Naïve Bayes, and Neural Networks—that predict UAT-based keywords from raw text. This report presents the **data exploration**, **preprocessing pipeline**, **model development**, **performance evaluation**, and a discussion of results obtained from these experiments.

Data Analysis

The dataset used in this project consists of five key fields: **bibcode**, **title**, **abstract**, **verified UAT IDs**, and **verified UAT labels**. The training set contains 18,700 entries, while the validation set includes 3,000. For this task, we focus on a simplified yet informative **subset—titles** and their corresponding **UAT labels**—as this is sufficient for keyword prediction.

The dataset is loaded using the **datasets** library from Hugging Face, which provides structured access to the SciX corpus. After loading, we inspect its structure by printing a few samples, focusing on the fields of interest: **title** and **verified UAT labels**.

To better understand the distribution of labels, we extract all keywords from the training samples and compute their frequencies using Python's **Counter** class. A bar chart is then generated to visualize the class distribution, which highlights the **label imbalance** a common challenge in multi-label tasks.

For text preprocessing, we use the **nltk** library to prepare the data for vectorization. Required resources such as the **Punkt** tokenizer and a list of English **stopwords** are downloaded. We define a custom **preprocess**_text function that:

- Converts text to lowercase
- Tokenizes it into words
- Removes non-alphanumeric characters and stopwords
- Rejoins the cleaned tokens into a single string

This function is applied to the titles in the training set, producing a cleaned and normalized text dataset. To transform the text into numerical input features, we use **TfidfVectorizer**, limiting the vocabulary to the 5,000 most relevant terms based on frequency and importance.

Meanwhile, the multi-label targets (UAT labels) are converted into a binary matrix using **MultiLabelBinarizer**, making them compatible with multi-label classification models.

Finally, the dataset is split into training and testing sets using **train_test_split**, with 80% of the data used for training and 20% for testing. The split is performed with a fixed random state to ensure reproducibility. As a final verification step, we print the shapes of the resulting feature and label matrices and display a few preprocessed titles to confirm the success of the preprocessing pipeline.

However, the resulting plot was overly condensed and difficult to interpret due to the large number of distinct labels (fig.1), making it visually cluttered and unreadable. To improve clarity, we limited the visualization to the top 50 most frequent labels (fig.2). This revised plot provides a more comprehensible view of the dataset's core label distribution and helps highlight potential imbalance issues, which are critical for modeling.

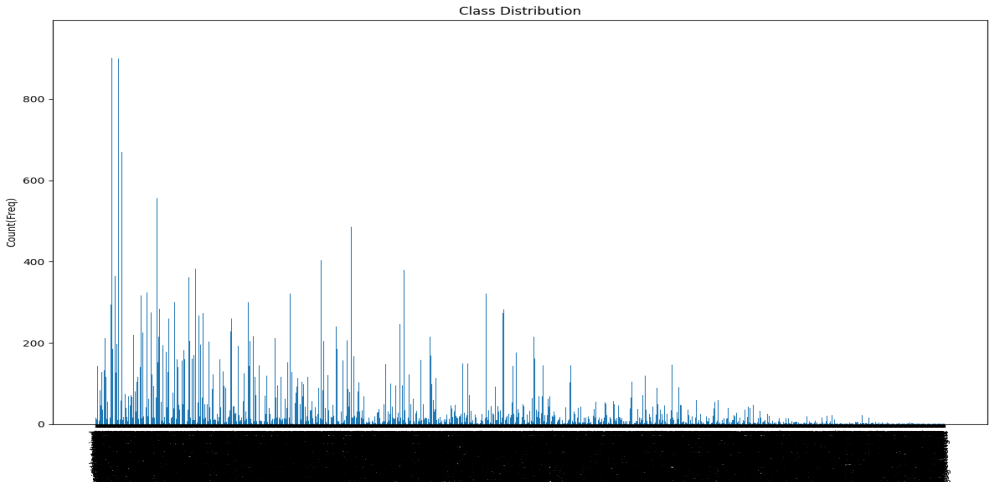


FIG.1 PLOT LABELS FREQUENCY

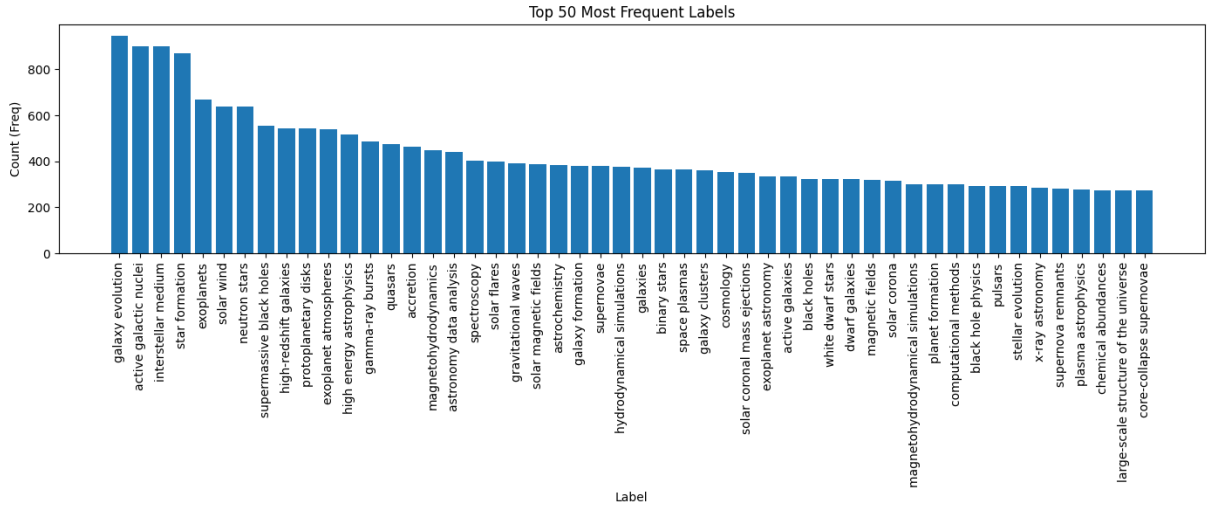


FIG.2 PLOT TOP 50 LABELS FREQUENCY

Possible Solutions

Throughout the project, we considered six different models as candidates for solving the multi-label classification problem. Each model offers its own strengths and trade-offs, depending on dataset characteristics, scalability needs, and computational constraints. Below is a brief overview and critical reflection on each method.

Support Vector Machines (SVM)

Support Vector Machines are well-established supervised learning algorithms, mostly used for binary classification problems. The idea is to find an optimal hyperplane that maximally separates data points from different classes. By maximizing the margin, the model tends to generalize better and reduce the risk of misclassification. In practice, kernels are often applied to project data into higher-dimensional space, which allows the model to handle non-linearly separable cases.

However, SVMs do not scale easily to multi-class or multi-label scenarios. Handling such tasks requires constructing multiple binary classifiers using strategies like one-vs-one or one-vs-all. This introduces significant computational overhead, especially when the number of labels is high, as is the case in this dataset. Given the complexity and cost of training many classifiers simultaneously, we decided not to proceed with this method.

K-Nearest Neighbors (K-NN)

The K-NN algorithm is a non-parametric, instance-based method that classifies new samples based on the majority class of their closest neighbors in the feature space. It is intuitive and can work reasonably well in heterogeneous datasets. The choice of 'k' (number of neighbors) is essential: too small a value can lead to noisy predictions, while too large a value may result in overly generalized classifications.

Despite its simplicity, K-NN faces limitations when applied to high-dimensional data, such as TF-IDF vectors from text. The computational cost of calculating distances grows significantly, making it inefficient on large datasets. Additionally, no learning phase means predictions require access to the full training set at runtime, which is not ideal for scalable deployment.

Neural Networks (NN)

Neural networks are a flexible and powerful family of models capable of capturing non-linear relationships and learning complex patterns in the data. Our setup included an input layer, multiple hidden layers with ReLU activation, and a sigmoid-activated output layer to handle multi-label outputs. The model is trained using a binary cross-entropy loss and optimized with Adam.

Neural networks often yield good results when enough data and compute resources are available. That said, they also require fine-tuning of hyperparameters, are prone to overfitting if not properly regularized, and are relatively slow to train. In our context, the model showed promising performance, although some instability in recall was noticed across rare labels, maybe due to class imbalance.

Decision Trees

Decision Trees build classification rules by recursively splitting the dataset based on selected features that reduce impurity (e.g., Gini index or entropy). They are interpretable, easy to implement, and provide a straightforward path from input to decision.

In this project, we used DecisionTreeClassifier within a MultiOutputClassifier to support multi-label classification. Among all models tested, the Decision Tree approach delivered the most balanced performance across precision, recall, and F1-score. While it's still susceptible to overfitting and doesn't scale as well as ensemble methods, its clarity and reasonable generalization made it the most suitable choice for our task.

Logistic Regression

Logistic regression is another classical method primarily suited for binary classification. It models the probability of class membership using a logistic function and assumes a linear relationship between input features and the log-odds of the outcome. Despite its simplicity and interpretability, this assumption may not hold in high-dimensional, non-linear data such as textual information.

For multi-label settings, logistic regression would also require a one-vs-rest strategy and is generally not favored for problems involving hundreds of labels. Additionally, in our tests, it proved to be slow and not much better than simpler models in terms of prediction quality. So, it was excluded from the final selection.

Naïve Bayes

Naïve Bayes classifiers are based on Bayes' Theorem, assuming conditional independence between features given the label. While this assumption is unrealistic in most real-world settings, it turns out to work surprisingly well for text classification tasks, thanks to the sparsity and structure of word-based representations.

We used the MultinomialNB implementation wrapped with a OneVsRestClassifier to adapt it for multi-label output. The model was extremely fast to train and evaluate, with very high precision on frequent labels. However, its recall was quite low, especially for underrepresented classes. This probably reflects its tendency to make conservative predictions unless it is very confident.

Overall, although Naïve Bayes showed good performance on some axes, it wasn’t as well balanced as the Decision Tree model, which we selected for further development.

Model Selection

Three models were selected for implementation and evaluation, aiming to identify which one delivers the most accurate multi-label keyword predictions. Each model builds on the same preprocessed dataset, which includes TF-IDF vectorized titles and binarized UAT labels.

Neural Networks

The neural network architecture was implemented using TensorFlow's Keras API. It begins with an input layer of 512 nodes using ReLU activation, followed by a dropout layer to help prevent overfitting. A second hidden layer with 256 nodes and another dropout is added before reaching the output layer, which uses a sigmoid activation function to support multi-label prediction across all unique classes.

The model is compiled using the Adam optimizer and binary cross-entropy loss, as it aligns with the binary nature of each label in a multi-label context. Training data is converted to dense arrays, and the model is trained for 10 epochs with a batch size of 32. Validation data is also used during training to monitor model performance.

After training, the model is evaluated on the test set, reporting test loss and accuracy. Predictions are then generated and binarized using a defined threshold. These binary outputs are inversely transformed back into label names for interpretation. Key performance metrics, including accuracy, precision, recall, and F1-score, are computed to assess the model's effectiveness. Additionally, confusion matrices for each label are plotted using Matplotlib for deeper diagnostic analysis.

Naïve Bayes

Before training, the **MultiLabelBinarizer** is adjusted to exclude always-present labels, which could bias the classifier and distort evaluation metrics. For this model, we use the **MultinomialNB** class from scikit-learn, embedded in a OneVsRestClassifier wrapper to support multi-label output.

The model is trained on the TF-IDF-transformed input data and corresponding binarized labels. After training, predictions are generated on the test set and compared to true labels. Standard evaluation metrics—including accuracy, precision, recall, and F1-score—are computed to analyze the model's overall effectiveness and ability to handle the label diversity in the dataset.

Evaluation and Metrics

To assess the performance of each implemented model, we used a combination of standard multi-label evaluation metrics: **accuracy**, **precision**, **recall**, and **F1 score**. These metrics provide insight into different aspects of model quality, particularly important when label imbalance is present.

Neural Network

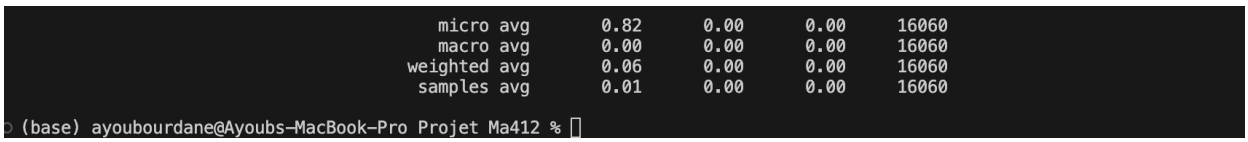
- **Accuracy**: 0.01 → Very poor; only about 1% of labels were predicted correctly.
- **Precision**: 0.60 → Decent; when the model makes a prediction, it is correct 60% of the time.
- **Recall**: 0.04 → Very low; the model identifies only 4% of the actual labels.
- **F1 Score**: 0.09 → Relatively low; indicates poor balance between precision and recall.

Although the neural network achieved the highest **precision**, its **recall** and **F1 score** were notably low. This suggests the model may be focusing on a limited set of frequent labels and failing to generalize across the full label space. Therefore, despite its strong predictive confidence when it does label, it is not an ideal solution due to its poor ability to detect the majority of relevant labels.

Naïve Bayes

- **Accuracy**: 0.001 → Extremely poor; only about 0.1% of the labels were correctly predicted.
- **Precision**: 0.80 → High; when the model does make a prediction, it is correct 80% of the time. However, it rarely predicts anything.
- **Recall**: 0.001 → Almost nonexistent; the model only retrieved 0.1% of the actual labels.
- **F1 Score**: 0.003 → The lowest among all models; reflects a significant imbalance between precision and recall.

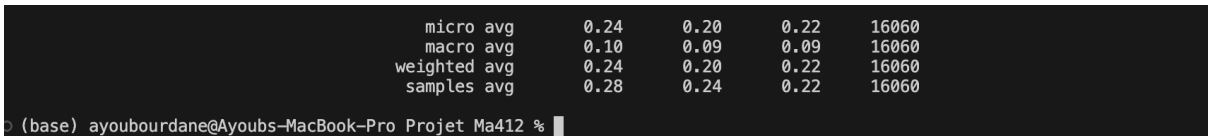
The Naïve Bayes model demonstrates **very high precision** but suffers from **critically low recall**. This means it only makes predictions when it's extremely confident, resulting in a conservative model that fails to generalize across the full label spectrum. Despite its efficiency, this behavior renders it impractical for multi-label classification where identifying diverse themes is key.



Decision Tree

- **Accuracy**: 0.008 → Poor; only 0.8% of the labels were correctly predicted.
- **Precision**: 0.20 → Low; when the model predicts a label, it is correct 20% of the time.
- **Recall**: 0.20 → Relatively higher than the other models; it captures 20% of the actual labels.
- **F1 Score**: 0.20 → The highest among the three models; reflects a more balanced trade-off between precision and recall.

Despite its modest precision, the Decision Tree model achieved the highest **F1 score**, indicating that it manages a better compromise between predicting relevant labels and minimizing false positives. This balanced performance makes it the most effective and reliable model among the ones tested for handling the multi-label classification task in this project.



Results Analysis

The implementation of the Decision Tree model for multi-label classification on the adsabs/SciX_UAT_keywords dataset demonstrates a systematic approach to solving the problem, encompassing data preprocessing, feature extraction, and model evaluation. The pipeline effectively integrates the use of the DecisionTreeClassifier within a MultiOutputClassifier to address the multi-label nature of the task.

Despite this comprehensive setup, the evaluation metrics indicate significant room for improvement:

- **Accuracy**: The model's accuracy (0.008) is very low, suggesting that fewer than 1% of the labels were correctly predicted. This indicates a general difficulty in capturing the full label space.
- **Precision**: With a precision of 0.2, the model makes correct predictions only 20% of the time. While this is better than a random guess, it signals that more refined features or a more sophisticated model might be needed.
- **Recall**: A recall of 0.2 shows that the model identifies 20% of the true labels. Though still low, this is a relative strength compared to the other models, suggesting some potential in refining the decision boundary.
- **F1 Score**: At 0.2, the F1 score—representing the harmonic mean of precision and recall—is the highest achieved among the models. This shows that the Decision Tree model offers the most balanced trade-off, even if the absolute performance is limited.

In summary, while the Decision Tree approach demonstrates comparative advantages in recall and F1 score, the overall effectiveness of the system remains modest. It presents a foundation that can be iteratively improved with further model tuning, better feature engineering, or the adoption of ensemble methods such as Random Forests or Gradient Boosting.

Model Strengths and Challenges

Model Strengths:

The pipeline is structurally sound and well-organized. It leverages TF-IDF vectorization to transform textual data into numerical features, which allows the model to handle the unstructured nature of scientific titles. Simultaneously, the MultiLabelBinarizer is employed to convert the multi-label annotations into a format suitable for supervised learning. These preprocessing steps are essential for making multi-label classification possible.

Moreover, the Decision Tree model brings interpretability to the system. It can be used to trace how specific features contribute to predictions, which is useful when transparency and explainability are valued. For smaller datasets or less complex structures, such models can sometimes outperform more complicated ones.

Challenges:

- **Data Imbalance**: The label frequency distribution revealed a significant class imbalance—some keywords appear in hundreds of documents while others occur only once or twice. This imbalance likely impairs the model's ability to recognize rare classes. Future improvements might include balancing the training set via oversampling underrepresented classes, undersampling frequent ones, or incorporating class-weighted loss functions.
- **Model Complexity**: Decision Trees struggle in high-dimensional spaces like the one formed by the 5,000 TF-IDF features. The model may become overly complex and prone to overfitting. A more effective approach would be to use ensemble methods such as Random Forest or Gradient Boosting, which mitigate overfitting by combining predictions from multiple trees.
- **Evaluation Metrics**: The low values across all evaluation metrics suggest that there is much room for enhancement. Hyperparameter tuning—including maximum tree depth, split criteria, and minimum samples per leaf—could help optimize the Decision Tree's performance. Feature engineering steps, like reducing TF-IDF dimensions or employing techniques like PCA, may also contribute to performance gains.

In conclusion, although the Decision Tree provides a structured and interpretable framework, its performance remains limited in its current form. Improving label balance, experimenting with dimensionality reduction, and trying advanced ensemble techniques are recommended next steps for developing a more accurate and generalizable multi-label classification system.

Conclusion

This project aimed to tackle the complex problem of multi-label classification in scientific literature using the NASA ADS SciX UAT dataset. Through systematic data exploration, careful preprocessing, and the implementation of multiple classification models, we evaluated how well different approaches could predict relevant keywords from paper titles. Among the models tested—Neural Networks, Naïve Bayes, and Decision Trees—the Decision Tree model emerged as the most balanced performer in terms of F1 score, offering a better trade-off between precision and recall.

However, despite the relative success of the Decision Tree, overall performance metrics remained low. The dataset's inherent challenges—such as extreme class imbalance, high feature dimensionality from TF-IDF, and the sparsity of short textual inputs—likely contributed to these limitations. In particular, the model's struggle to generalize across the long tail of infrequent labels highlighted the need for more nuanced methods that can handle imbalance and capture label dependencies.

Looking forward, several improvements could be explored. Incorporating ensemble models like Random Forests or Gradient Boosting could enhance generalization and reduce overfitting. Addressing class imbalance with resampling techniques or custom loss functions may help recover performance on underrepresented labels. Finally, leveraging semantic embeddings or transformer-based models could offer a more context-aware representation of textual data, moving beyond TF-IDF's limitations.

In sum, while this project has laid a solid foundation for multi-label text classification using interpretable models, future iterations should aim for more sophisticated architectures and better handling of dataset complexity to achieve higher performance and robustness.

Annexe

red giant bump	0.00	0.00	0.00	2
red giant clump	0.43	0.75	0.55	4
red giant stars	0.12	0.17	0.14	6
red giant tip	0.50	0.25	0.33	4
red noise	0.00	0.00	0.00	0
red sequence galaxies	0.00	0.00	0.00	1
red straggler stars	0.00	0.00	0.00	1
red supergiant stars	0.21	0.33	0.26	9
reddened stars	0.00	0.00	0.00	1
reddening law	0.00	0.00	0.00	4
redshift surveys	0.18	0.17	0.17	18
redshifted	0.00	0.00	0.00	1
reflecting telescopes	1.00	1.00	1.00	1
reflection nebulae	0.00	0.00	0.00	2
regolith	0.00	0.00	0.00	2
regression	0.00	0.00	0.00	1
reionization	0.43	0.32	0.36	38
reissner-nordström black holes	0.00	0.00	0.00	0
relativistic aberration	0.00	0.00	0.00	0
relativistic binary stars	0.00	0.00	0.00	1
relativistic cosmology	0.00	0.00	0.00	0
relativistic disks	0.00	0.00	0.00	1
relativistic fluid dynamics	0.00	0.00	0.00	4
relativistic jets	0.36	0.26	0.30	38
relativistic mechanics	0.00	0.00	0.00	3
relativistic stars	0.00	0.00	0.00	2

relativity	0.00	0.00	0.00	4
relaxation time	0.00	0.00	0.00	0
remote sensing	0.00	0.00	0.00	7
remote telescope astrophotography	0.00	0.00	0.00	1
resonant kuiper belt objects	0.25	1.00	0.40	1
reverberation mapping	0.53	0.60	0.56	15
rgb photometry	0.00	0.00	0.00	0
rich galaxy clusters	0.00	0.00	0.00	2
ring galaxies	0.00	0.00	0.00	1
ring nebulae	0.00	0.00	0.00	1
ring resonance	0.00	0.00	0.00	0
robust regression	0.00	0.00	0.00	2
roche limit	0.00	0.00	0.00	0
roche lobe	0.00	0.00	0.00	1
roche lobe overflow	1.00	0.50	0.67	2
rockets	0.00	0.00	0.00	0
rotating black holes	0.00	0.00	0.00	3
rotation powered pulsars	0.33	0.12	0.18	8
rotational spectroscopy	0.00	0.00	0.00	1
rovers	0.00	0.00	0.00	3
rr lyrae variable stars	0.67	0.80	0.73	10
rrab variable stars	0.00	0.00	0.00	1
rrc variable stars	0.00	0.00	0.00	0
rrd variable stars	0.00	0.00	0.00	0
rs canum venaticorum variable stars	0.00	0.00	0.00	1
runaway stars	0.00	0.00	0.00	2
rv tauri variable stars	0.00	0.00	0.00	1
s stars	0.00	0.00	0.00	2

s-process 0.00 0.00 0.00 6

sagittarius dwarf spheroidal galaxy 1.00 0.50 0.67 2

satellite microlensing parallax 0.00 0.00 0.00 0

saturn 0.50 0.33 0.40 6

saturnian satellites 0.50 0.44 0.47 9

scalar-tensor-vector gravity 0.00 0.00 0.00 1

scale height 0.00 0.00 0.00 0

scaling relations 0.00 0.00 0.00 12

scattered disk objects 0.20 1.00 0.33 1

schwarzschild black holes 0.00 0.00 0.00 0

schwarzschild metric 0.00 0.00 0.00 0

schwarzschild radius 0.00 0.00 0.00 0

sculptor dwarf elliptical galaxy 0.00 0.00 0.00 1

search for extraterrestrial intelligence 0.75 0.67 0.71 9

seasonal phenomena 0.00 0.00 0.00 2

secondary cosmic rays 0.00 0.00 0.00 1

selenology 0.00 0.00 0.00 0

semi-detached binary stars 0.00 0.00 0.00 1

semi-regular variable stars 0.00 0.00 0.00 1

seyfert galaxies 0.35 0.30 0.32 37

shepherd satellites 0.00 0.00 0.00 0

shocks 0.39 0.25 0.31 51

short period comets 0.00 0.00 0.00 12

short period variable stars 0.00 0.00 0.00 0

sigma8 0.00 0.00 0.00 0

silicate grains 0.00 0.00 0.00 4

silicon burning 0.00 0.00 0.00 0

silicon monoxide masers 0.00 0.00 0.00 2

silicon stars	0.00	0.00	0.00	2
single x-ray stars	0.00	0.00	0.00	4
single-dish antennas	0.00	0.00	0.00	0
single-linkage hierarchical clustering	0.00	0.00	0.00	0
sky brightness	0.00	0.00	0.00	1
sky noise	0.00	0.00	0.00	0
sky surveys	0.15	0.20	0.17	30
sloan photometry	0.00	0.00	0.00	1
slow irregular variable stars	0.00	0.00	0.00	1
slow novae	0.00	0.00	0.00	0
slow solar wind	0.10	0.09	0.10	11
small magellanic cloud	1.00	0.14	0.25	7
small molecules	0.00	0.00	0.00	3
small solar system bodies	0.14	0.11	0.12	27
smoothing	0.00	0.00	0.00	0
sociology of astronomy	0.00	0.00	0.00	2
soft gamma-ray repeaters	0.67	0.33	0.44	6
software available on request	0.00	0.00	0.00	0
software documentation	0.00	0.00	0.00	2
software tutorials	0.00	0.00	0.00	0
solar abundances	0.00	0.00	0.00	11
solar active region filaments	0.00	0.00	0.00	4
solar active region magnetic fields	0.07	0.06	0.06	18
solar active region velocity fields	1.00	0.33	0.50	3
solar active regions	0.18	0.08	0.11	36
solar activity	0.12	0.10	0.11	52
solar analogs	0.00	0.00	0.00	5
solar atmosphere	0.11	0.07	0.08	29

solar atmospheric motions	0.00	0.00	0.00	1
solar chromosphere	0.46	0.49	0.47	37
solar chromospheric heating	0.20	0.50	0.29	2
solar convective zone	0.00	0.00	0.00	5
solar corona	0.22	0.19	0.21	67
solar coronal heating	0.40	0.24	0.30	34
solar coronal holes	0.20	0.17	0.18	12
solar coronal lines	0.00	0.00	0.00	2
solar coronal loops	0.50	0.28	0.36	18
solar coronal mass ejection shocks	0.00	0.00	0.00	12
solar coronal mass ejections	0.58	0.52	0.55	75
solar coronal plumes	0.67	0.67	0.67	3
solar coronal radio emission	0.00	0.00	0.00	9
solar coronal seismology	0.33	0.50	0.40	2
solar coronal streamers	0.17	0.33	0.22	3
solar coronal transients	0.00	0.00	0.00	9
solar coronal waves	0.25	0.12	0.16	25
solar cycle	0.23	0.19	0.20	27
solar differential rotation	0.00	0.00	0.00	3
solar dynamo	0.33	0.17	0.22	6
solar e corona	0.00	0.00	0.00	1
solar eclipses	0.50	0.33	0.40	3
solar electromagnetic emission	0.00	0.00	0.00	2
solar energetic particles	0.57	0.35	0.43	37
solar evolution	0.00	0.00	0.00	2
solar extreme ultraviolet emission	0.11	0.09	0.10	23
solar f corona	0.00	0.00	0.00	1
solar faculae	0.00	0.00	0.00	0

solar fibrils	0.00	0.00	0.00	1
solar filament eruptions	0.40	0.38	0.39	16
solar filaments	0.27	0.18	0.21	17
solar flare spectra	0.17	0.12	0.14	8
solar flares	0.61	0.40	0.49	89
solar gamma-ray emission	0.50	0.20	0.29	5
solar granulation	0.20	0.25	0.22	4
solar granules	0.00	0.00	0.00	1
solar instruments	0.00	0.00	0.00	3
solar interior	0.33	0.33	0.33	6
solar k corona	0.00	0.00	0.00	1
solar magnetic bright points	0.00	0.00	0.00	1
solar magnetic fields	0.28	0.25	0.26	88
solar magnetic flux emergence	0.11	0.25	0.15	4
solar magnetic reconnection	0.48	0.44	0.46	57
solar meridional circulation	0.17	0.50	0.25	2
solar motion	0.00	0.00	0.00	0
solar mottles	0.00	0.00	0.00	1
solar nebulae	0.00	0.00	0.00	2
solar neighborhood	0.12	0.14	0.13	7
solar neutrino problem	0.00	0.00	0.00	0
solar neutrinos	0.00	0.00	0.00	0
solar observatories	0.00	0.00	0.00	0
solar optical telescopes	0.00	0.00	0.00	1
solar oscillations	0.30	0.19	0.23	16
solar particle emission	0.25	0.22	0.24	9
solar photosphere	0.10	0.08	0.09	25
solar physics	0.14	0.13	0.14	45

solar prominences	0.40	0.33	0.36	12
solar radiation	0.00	0.00	0.00	2
solar radiative zone	0.00	0.00	0.00	1
solar radio emission	0.00	0.00	0.00	14
solar radio flares	0.17	0.14	0.15	7
solar radio telescopes	0.00	0.00	0.00	2
solar radius	0.00	0.00	0.00	0
solar rotation	0.20	0.20	0.20	5
solar spectral irradiance	0.00	0.00	0.00	2
solar spicules	0.50	0.33	0.40	6
solar storm	0.00	0.00	0.00	4
solar surface	0.00	0.00	0.00	1
solar system	0.00	0.00	0.00	15
solar system astronomy	0.00	0.00	0.00	6
solar system evolution	0.00	0.00	0.00	2
solar system formation	0.00	0.00	0.00	12
solar system gas giant planets	0.00	0.00	0.00	7
solar system planets	0.14	0.14	0.14	7
solar system terrestrial planets	0.00	0.00	0.00	9
solar telescopes	0.00	0.00	0.00	2
solar transition region	0.14	0.12	0.13	8
solar ultraviolet emission	0.20	0.07	0.10	15
solar white-light flares	0.00	0.00	0.00	1
solar wind	0.67	0.63	0.65	132
solar wind termination	0.00	0.00	0.00	2
solar x-ray emission	0.20	0.22	0.21	9
solar x-ray flares	0.25	0.13	0.17	15
solar-planetary interactions	0.00	0.00	0.00	7

solar-terrestrial interactions	0.20	0.18	0.19	11
solid body tides	0.00	0.00	0.00	0
solid matter physics	0.17	1.00	0.29	1
space astrometry	0.67	0.29	0.40	7
space debris	0.00	0.00	0.00	0
space observatories	0.00	0.00	0.00	4
space plasmas	0.38	0.33	0.35	67
space probes	0.00	0.00	0.00	5
space research	0.00	0.00	0.00	0
space telescopes	0.12	0.06	0.08	16
space vehicle instruments	0.00	0.00	0.00	6
space vehicles	0.00	0.00	0.00	3
space weather	0.09	0.12	0.10	17
spacetime metric	0.00	0.00	0.00	0
spatial point processes	0.00	0.00	0.00	2
special relativity	0.00	0.00	0.00	2
speckle interferometry	0.33	1.00	0.50	1
spectral energy distribution	0.16	0.12	0.14	25
spectral index	0.00	0.00	0.00	5
spectral line identification	0.20	0.10	0.13	10
spectral line lists	0.25	0.25	0.25	4
spectrometers	0.00	0.00	0.00	5
spectrophotometric standards	0.00	0.00	0.00	2
spectrophotometry	0.00	0.00	0.00	7
spectropolarimetry	0.14	0.19	0.16	16
spectroscopic binary stars	0.00	0.00	0.00	15
spectroscopy	0.12	0.08	0.10	97
spin-orbit resonances	0.00	0.00	0.00	0

spiral arms	0.00	0.00	0.00	6
spiral galaxies	0.29	0.24	0.26	17
spiral pitch angle	0.00	0.00	0.00	1
src variable stars	0.00	0.00	0.00	0
standard candles	0.00	0.00	0.00	2
standard stars	0.00	0.00	0.00	0
star atlases	0.00	0.00	0.00	0
star clusters	0.33	0.14	0.20	35
star counts	0.00	0.00	0.00	3
star formation	0.38	0.32	0.35	174
star forming regions	0.04	0.03	0.04	32
star-planet interactions	0.00	0.00	0.00	7
starburst galaxies	0.42	0.47	0.44	43
starlight polarization	0.14	0.11	0.12	9
starspots	0.21	0.16	0.18	19
statistical parallax	0.00	0.00	0.00	1
stellar abundances	0.28	0.22	0.25	50
stellar accretion	0.07	0.07	0.07	15
stellar accretion disks	0.04	0.06	0.05	35
stellar activity	0.19	0.21	0.20	33
stellar ages	0.14	0.17	0.15	18
stellar associations	0.25	0.18	0.21	11
stellar astronomy	0.12	0.07	0.09	15
stellar atmospheres	0.06	0.04	0.04	28
stellar atmospheric opacity	0.14	0.20	0.17	5
stellar bow shocks	0.00	0.00	0.00	1
stellar chromospheres	0.22	0.29	0.25	7
stellar classification	0.00	0.00	0.00	3

stellar colors	0.00	0.00	0.00	2
stellar convection envelopes	0.00	0.00	0.00	3
stellar convective shells	0.00	0.00	0.00	0
stellar convective zones	0.00	0.00	0.00	7
stellar cores	0.00	0.00	0.00	4
stellar coronae	0.00	0.00	0.00	8
stellar coronal dimming	0.00	0.00	0.00	0
stellar coronal lines	0.00	0.00	0.00	0
stellar coronal loops	0.00	0.00	0.00	0
stellar coronal mass ejections	0.00	0.00	0.00	3
stellar diffusion	0.00	0.00	0.00	5
stellar distance	0.00	0.00	0.00	5
stellar dynamics	0.11	0.08	0.09	26
stellar effective temperatures	0.00	0.00	0.00	1
stellar evolution	0.16	0.15	0.16	60
stellar evolutionary models	0.11	0.10	0.11	20
stellar evolutionary tracks	0.00	0.00	0.00	1
stellar evolutionary types	0.00	0.00	0.00	0
stellar faculae	0.00	0.00	0.00	1
stellar feedback	0.20	0.20	0.20	10
stellar flares	0.42	0.28	0.33	18
stellar granulation	0.00	0.00	0.00	1
stellar inner cores	0.00	0.00	0.00	0
stellar interiors	0.18	0.12	0.14	17
stellar jets	0.20	0.14	0.17	21
stellar kinematics	0.05	0.04	0.04	26
stellar luminosities	0.00	0.00	0.00	3
stellar magnetic fields	0.00	0.00	0.00	14

stellar mass black holes	0.12	0.10	0.11	31
stellar mass functions	0.00	0.00	0.00	6
stellar mass loss	0.18	0.14	0.16	21
stellar masses	0.00	0.00	0.00	5
stellar mergers	0.00	0.00	0.00	15
stellar motion	0.00	0.00	0.00	2
stellar nucleosynthesis	0.00	0.00	0.00	7
stellar occultation	0.29	0.40	0.33	5
stellar oscillations	0.14	0.11	0.12	18
stellar parallax	0.00	0.00	0.00	1
stellar phenomena	0.00	0.00	0.00	3
stellar photometry	0.00	0.00	0.00	5
stellar photospheres	0.00	0.00	0.00	2
stellar physics	0.21	0.18	0.19	17
stellar populations	0.00	0.00	0.00	18
stellar processes	0.00	0.00	0.00	2
stellar properties	0.00	0.00	0.00	7
stellar pulsations	0.00	0.00	0.00	12
stellar radii	0.00	0.00	0.00	2
stellar remnants	0.00	0.00	0.00	4
stellar rotation	0.29	0.21	0.25	42
stellar spectral lines	0.08	0.11	0.10	9
stellar spectral types	0.00	0.00	0.00	0
stellar streams	0.50	0.33	0.40	9
stellar structures	0.00	0.00	0.00	4
stellar surfaces	0.00	0.00	0.00	2
stellar types	0.00	0.00	0.00	5
stellar wind bubbles	0.00	0.00	0.00	4

stellar winds	0.12	0.09	0.10	35
stellar x-ray flares	0.00	0.00	0.00	3
stellar-interstellar interactions	0.00	0.00	0.00	2
stratosphere	0.00	0.00	0.00	0
strong gravitational lensing	0.47	0.41	0.44	34
structure determination	0.00	0.00	0.00	0
strömgren photometry	0.00	0.00	0.00	0
strömgren spheres	0.00	0.00	0.00	1
su ursae majoris stars	0.00	0.00	0.00	1
subdwarf stars	0.00	0.00	0.00	2
subgiant stars	0.00	0.00	0.00	2
submillimeter astronomy	0.37	0.22	0.27	50
substellar companion stars	0.00	0.00	0.00	1
sundivers	0.00	0.00	0.00	0
sungrazers	0.00	0.00	0.00	0
sunskirters	0.00	0.00	0.00	0
sunspot cycle	0.33	0.12	0.18	8
sunspot flow	0.00	0.00	0.00	1
sunspot groups	0.00	0.00	0.00	4
sunspot number	0.00	0.00	0.00	2
sunspots	0.35	0.42	0.38	19
sunyaev-zeldovich effect	0.00	0.00	0.00	8
super earths	0.20	0.08	0.12	12
superbubbles	0.00	0.00	0.00	4
superclusters	0.00	0.00	0.00	1
supergiant stars	0.33	0.25	0.29	4
supergranulation	0.00	0.00	0.00	1
supermassive black holes	0.40	0.37	0.38	116

supernova dynamics	0.00	0.00	0.00	5
supernova neutrinos	0.43	0.38	0.40	8
supernova remnants	0.71	0.59	0.64	70
supernovae	0.45	0.39	0.42	76
support vector machine	0.00	0.00	0.00	1
surface composition	0.00	0.00	0.00	3
surface gravity	0.00	0.00	0.00	2
surface ices	0.14	0.10	0.12	10
surface photometry	0.00	0.00	0.00	2
surface processes	0.00	0.00	0.00	4
surface variability	0.00	0.00	0.00	1
surveys	0.13	0.10	0.11	58
survival analysis	0.00	0.00	0.00	0
sx phoenicis variable stars	0.00	0.00	0.00	0
symbiotic binary stars	1.00	0.38	0.55	8
symbiotic novae	0.50	0.33	0.40	3
t associations	0.00	0.00	0.00	1
t dwarfs	0.08	0.09	0.08	11
t subdwarfs	0.00	0.00	0.00	3
t tauri stars	0.33	0.19	0.24	16
tailed radio galaxies	0.00	0.00	0.00	0
technosignatures	0.67	1.00	0.80	6
tectonics	0.00	0.00	0.00	0
telescopes	0.33	0.17	0.22	6
termination shock	0.00	0.00	0.00	3
the moon	0.46	0.35	0.40	17
the sun	0.13	0.14	0.13	29
theoretical models	0.00	0.00	0.00	14

theoretical techniques	0.50	0.20	0.29	5
thermal properties (ice)	0.00	0.00	0.00	0
thermosphere	0.00	0.00	0.00	1
three-body problem	0.50	0.11	0.18	9
tidal disruption	0.73	0.54	0.62	41
tidal distortion	0.00	0.00	0.00	3
tidal friction	0.00	0.00	0.00	2
tidal interaction	0.00	0.00	0.00	3
tidal radius	0.00	0.00	0.00	0
tidal tails	0.00	0.00	0.00	4
tides	0.17	0.50	0.25	2
time domain astronomy	0.13	0.14	0.14	29
time series analysis	0.16	0.12	0.14	25
time-of-flight mass spectrometry	0.00	0.00	0.00	2
timing variation methods	0.00	0.00	0.00	2
titan	0.83	1.00	0.91	5
total eclipses	0.00	0.00	0.00	1
trans-neptunian objects	0.07	0.08	0.07	13
transient detection	0.12	0.29	0.17	7
transient sources	0.00	0.00	0.00	25
transit duration variation method	0.00	0.00	0.00	0
transit instruments	0.00	0.00	0.00	1
transit photometry	0.12	0.06	0.08	35
transit timing variation method	0.00	0.00	0.00	5
transition probabilities	0.25	0.33	0.29	3
transits	0.20	0.11	0.14	18
transmission spectroscopy	0.25	0.20	0.22	15
triangulum galaxy	0.75	1.00	0.86	3

trigonometric parallax	0.00	0.00	0.00	3
trinary stars	0.00	0.00	0.00	1
triple lens microlensing	0.00	0.00	0.00	0
triton	0.00	0.00	0.00	2
trojan asteroids	0.33	1.00	0.50	1
trojan planets	0.00	0.00	0.00	0
turnoff point	0.00	0.00	0.00	0
twilight	0.00	0.00	0.00	1
two-body problem	0.33	0.50	0.40	2
two-color diagrams	0.00	0.00	0.00	1
two-point correlation function	0.00	0.00	0.00	4
twotinos	0.00	0.00	0.00	0
type ia supernovae	0.62	0.67	0.64	39
type ib supernovae	0.33	0.14	0.20	7
type ic supernovae	0.40	0.15	0.22	13
type ii cepheid variable stars	0.00	0.00	0.00	2
type ii supernovae	0.44	0.30	0.36	27
u geminorum stars	0.00	0.00	0.00	2
ultra-high-energy cosmic radiation	0.20	0.09	0.13	11
ultracompact dwarf galaxies	0.00	0.00	0.00	1
ultraluminous infrared galaxies	0.10	0.09	0.10	11
ultraluminous x-ray sources	0.00	0.00	0.00	3
ultraviolet astronomy	0.11	0.11	0.11	19
ultraviolet color	0.00	0.00	0.00	0
ultraviolet extinction	0.00	0.00	0.00	2
ultraviolet observatories	0.00	0.00	0.00	0
ultraviolet photometry	0.00	0.00	0.00	1
ultraviolet sources	0.00	0.00	0.00	6

ultraviolet spectroscopy	0.00	0.00	0.00	10
ultraviolet surveys	0.00	0.00	0.00	4
ultraviolet telescopes	0.00	0.00	0.00	1
ultraviolet transient sources	0.00	0.00	0.00	4
umbra	0.00	0.00	0.00	0
uncertainty bounds	0.00	0.00	0.00	0
upper atmosphere	0.00	0.00	0.00	3
uranian satellites	0.57	1.00	0.73	4
uranus	0.57	1.00	0.73	4
urca process	0.00	0.00	0.00	0
uu herculis stars	0.00	0.00	0.00	1
uv ceti stars	0.00	0.00	0.00	0
van allen radiation belts	0.00	0.00	0.00	2
variable radiation sources	0.00	0.00	0.00	3
variable star period change	0.00	0.00	0.00	0
variable stars	0.06	0.06	0.06	18
venus	0.70	0.70	0.70	10
very large array	0.00	0.00	0.00	0
very large telescope	0.00	0.00	0.00	2
very long baseline interferometers	0.00	0.00	0.00	0
very long baseline interferometry	0.27	0.24	0.25	17
very small grains	0.00	0.00	0.00	1
vesta	0.00	0.00	0.00	3
vibrational spectroscopy	0.00	0.00	0.00	1
virgo cluster	1.00	0.67	0.80	3
virgo supercluster	0.00	0.00	0.00	0
virtual observatories	0.00	0.00	0.00	0
visible sources	0.00	0.00	0.00	1

visual binary stars	0.12	0.20	0.15	5
visual observation	0.00	0.00	0.00	0
voids	1.00	0.57	0.73	7
volcanism	0.00	0.00	0.00	4
volcanoes	0.00	0.00	0.00	1
von zeipel theorem	0.00	0.00	0.00	0
voronoi tessellation	0.00	0.00	0.00	0
vy sculpturis stars	0.00	0.00	0.00	0
w ursae majoris variable stars	0.00	0.00	0.00	1
warm dark matter	0.00	0.00	0.00	2
warm ionized medium	0.12	0.33	0.18	3
warm neutral medium	0.00	0.00	0.00	0
warm-hot intergalactic medium	0.00	0.00	0.00	0
water masers	0.25	0.33	0.29	3
water vapor	0.00	0.00	0.00	6
wavelet analysis	0.00	0.00	0.00	5
wc stars	0.00	0.00	0.00	3
weak gravitational lensing	0.25	0.33	0.29	9
weak-line t tauri stars	0.00	0.00	0.00	0
white dwarf stars	0.53	0.48	0.50	64
wide binary stars	0.25	0.40	0.31	5
wide-field telescopes	0.00	0.00	0.00	0
wn stars	0.00	0.00	0.00	0
wolf-rayet stars	0.20	0.14	0.17	7
wormholes	0.00	0.00	0.00	0
wz sagittae stars	0.00	0.00	0.00	2
x-ray active galactic nuclei	0.10	0.09	0.09	47
x-ray astronomy	0.14	0.10	0.11	61

x-ray binary stars	0.24	0.28	0.26	40
x-ray bright point	0.00	0.00	0.00	2
x-ray bursters	0.00	0.00	0.00	2
x-ray bursts	0.00	0.00	0.00	7
x-ray detectors	0.00	0.00	0.00	2
x-ray identification	0.00	0.00	0.00	2
x-ray novae	0.00	0.00	0.00	0
x-ray observatories	0.00	0.00	0.00	2
x-ray photometry	0.00	0.00	0.00	2
x-ray point sources	0.00	0.00	0.00	3
x-ray quasars	0.09	0.17	0.12	6
x-ray sources	0.05	0.03	0.04	30
x-ray stars	0.33	0.12	0.18	8
x-ray surveys	0.00	0.00	0.00	3
x-ray telescopes	0.17	0.17	0.17	6
x-ray transient sources	0.24	0.17	0.20	23
xallarap effect	0.00	0.00	0.00	0
xenobiology	0.00	0.00	0.00	0
y dwarfs	0.29	0.22	0.25	9
yellow hypergiant stars	0.00	0.00	0.00	2
yellow straggler stars	0.00	0.00	0.00	0
young disk cepheid variable stars	0.00	0.00	0.00	1
young massive clusters	0.00	0.00	0.00	4
young star clusters	0.14	0.09	0.11	22
young stellar objects	0.15	0.15	0.15	41
z camelopardalis stars	0.00	0.00	0.00	0
zero-age main sequence stars	0.00	0.00	0.00	0
zodiacal cloud	0.00	0.00	0.00	5

zz ceti stars	0.00	0.00	0.00	1
micro avg	0.24	0.20	0.22	16060
macro avg	0.10	0.09	0.09	16060
weighted avg	0.24	0.20	0.22	16060
samples avg	0.28	0.24	0.22	16060

(base) ayoubourdane@Ayoubs-MacBook-Pro Projet Ma412 %