Une image contenant Police, logo, texte, Graphique

Description générée automatiquementQuentin Chardon du Ranquet

Clément Chandezon

IPSA 4A – Embedded Systems

Outils mathématiques avancés pour la science des données et la prise de décision

Une image contenant cercle, art, léger

Description générée automatiquement

Table des matières

[Introduction: 2](#_Toc187342204)

[Code explication: 3](#_Toc187342205)

[Methods: 4](#_Toc187342206)

[Results: 4](#_Toc187342207)

# Introduction:

Dans un contexte où la recherche scientifique est en constante expansion, la nécessité d'outils efficaces pour la classification et l'exploration de données devient primordiale. Le projet présenté ici se concentre sur la classification multi-étiquette des documents scientifiques en utilisant le corpus NASA SciX, ensemble de publications couvrant des domaines tels que l'astronomie, la physique, et les sciences planétaires.

La classification multi-étiquette permet d'assigner plusieurs mots-clés pertinents à chaque document, facilitant ainsi une catégorisation précise et une recherche efficace dans des bases de données interdisciplinaires. Ce projet vise à concevoir et entraîner un modèle d'apprentissage automatique capable de prédire ces mots-clés à partir des titres et résumés des articles. Les résultats attendus incluent un gain de temps dans l'indexation manuelle des documents et une optimisation des recherches.

À travers une exploration rigoureuse des données, une sélection judicieuse des modèles, et une analyse des performances, ce projet s'inscrit dans une démarche scientifique visant à renforcer les outils de gestion et d'accès à la littérature scientifique.

# Code explication:

We decided to use multiple methods to answer the classification of the text. We used the LinearSVC, combined with the OneVsRestClassifier and we used the method of the SGDClassifier, combined with the OneVsRestClassifier.

We decided mainly to focus on the speed of the classification, with these methods that allow us a fast classification of the input text.

The use of the OneVsRestClassifier is needed, because the model needs to understand there are more than two different classes. The use of OneVsRestClassifier allows this and this is why we combine it with our classifier.

We decided to train our model on the smaller dataset, named *val-00000-of-00001-66ce8665444026dc*, to fasten the training.

To train our models, we decided to combine the title and the abstract together to have one long string of characters to be understood by our models.

We modify the text string created by using the *TfidfVectorizer* function in sklearn to create a matrix of frequences of each word in the string of characters.

There are 3 different files in the GitHub:

The file **"***First draft.ipynb***"** was the first file I created for text classification.

The file **"***Pipeline implementation.ipynb***"** implements the code found in the sklearn documentation using the dataset provided to us.

The file **"***Optimization Pipeline implementation.ipynb***"** enhances the code from the documentation with various techniques.

For each classifier, we use a random\_state equal to 42, to have the same results every time, so we can compare every result with the precedent.

## Methods:

* LinearSVC:

This method is mainly used in multi-classification problems, with the One-Vs-Rest strategy, to find which class fits more with the given data.

* SGDClassifier:

The SGDClassifier is performant for larger characteristics, and in our case is useful, because we have a lot of data to analyze, because some abstracts are very long.

It can also use regression loss to avoid overfitting, and we use the log\_loss parameter to use this regression loss.

## Results:

For the last file created and the most optimized named *Optimization Pipeline implementation.ipynb,* we obtain the next results.

For the **SGDClassifier**, we have:

Accuracy: 0.006622516556291391

Precision: 0.2701458173445894

Recall: 0.2558139534883721

F1 Score: 0.2627846211272863

Hamming Loss: 0.008123894533338817

For **LinearSVC**, we have:

Accuracy: 0.019867549668874173

Precision: 0.24635456638526476

Recall: 0.4713656387665198

F1 Score: 0.32358870967774194

Hamming Loss: 0.005520134918349719

We can see that the LinearSVC method gets better results than the SGDClassifier method, because even if the precision is lower for this method, all other metrics are greater than the other method. The hamming loss must be near to 0, so it gets better result for the LinearSVC method. It also has a lower running time.

By printing the number of ones in the results, we should obtain approximately the same number between the predicted values and the trained ones.

By printing these, we obtain:

SGDClassifier: predicted\_values: 2752 / true\_values: 2606

LinearSVC: predicted\_values: 1362 / true\_values: 2606

We can see that the number of predicted values for the LinearSVC method is very different from the true values because it is missing half of the values instead of the SGDClassifier even if it has a higher precision than the precedent method.

# **Conclusion**

Ce projet a permis d'explorer différentes méthodes de classification multi-étiquette appliquées à un corpus scientifique riche et varié. À travers l'utilisation des algorithmes LinearSVC et SGDClassifier, combinés avec le OneVsRestClassifier, nous avons pu comparer leurs performances selon des métriques variées telles que la précision, le rappel, le F1-score, et la perte de Hamming.

Les résultats obtenus mettent en évidence une supériorité globale de la méthode LinearSVC en termes de précision et de rapidité, bien que son faible nombre de prédictions reste une limitation à prendre en compte. À l'inverse, le SGDClassifier, malgré des performances légèrement inférieures sur certaines métriques, offre un nombre de prédictions plus proche des valeurs réelles.

Ce travail ouvre la voie à des optimisations futures, notamment en affinant les prétraitements ou en explorant d'autres modèles d'apprentissage. L'objectif final reste de fournir un outil performant et fiable pour faciliter la gestion et l'accès aux données scientifiques, répondant ainsi à un besoin croissant dans le domaine de la recherche