

Projet 5

Catégorisez automatiquement des questions

Loridan Adrien



student machine learning
engineer



medium.com/@adrien.loridan



linkedin.com/in/adrien-loridan



github.com/loridan



Tu vois, le monde se divise en deux catégories...

ceux qui ont la classe et les autres



Tu vois, le monde se divise en deux catégories...

ceux qui ont la classe et les autres

Présentation

Stack Overflow est un site qui permet à ses utilisateurs de créer des questions en rapport avec le développement informatique.

Lors de la création d'une question, il faut associer une ou plusieurs étiquettes qui décrivent le sujet.



Tu vois, le monde se divise en deux catégories...

ceux qui ont la classe et les autres

Présentation

Ces étiquettes sont des catégories ou classes à priori sans hiérarchie.

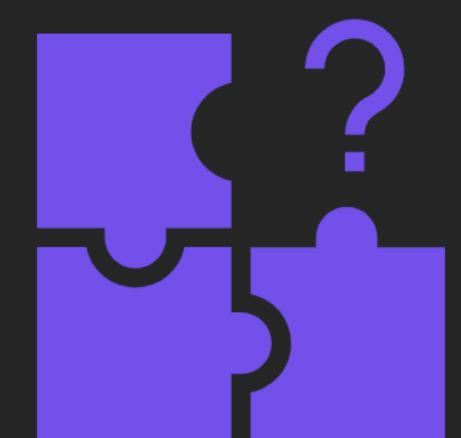
Elles permettent un regroupement facile des questions lors d'une recherche ultérieure par l'utilisateur.

Tu vois, le monde se divise en deux catégories...
ceux qui ont la classe et les autres

Problème

L'utilisateur doit réfléchir et trouver les étiquettes les plus pertinentes.

C'est un effort cognitif supplémentaire désagréable pour l'expérience utilisateur.



Tu vois, le monde se divise en deux catégories...

ceux qui ont la classe et les autres

Problème

Peut-on réaliser un système de suggestion d'étiquettes ?

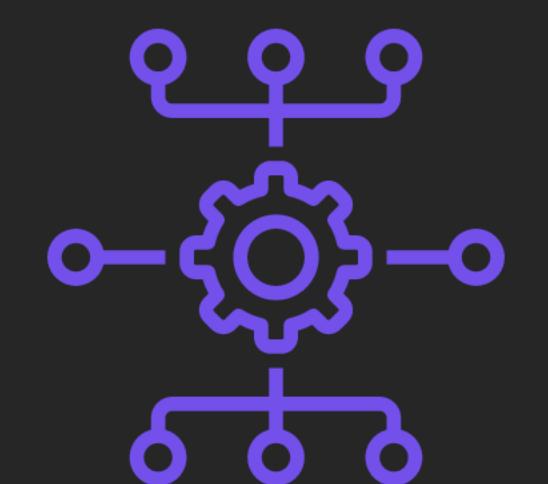


Tu vois, le monde se divise en deux catégories...

ceux qui ont la classe et les autres

Une solution envisagée...

est d'utiliser un algorithme d'apprentissage automatique pour suggérer des étiquettes à partir des données.



Tu vois, le monde se divise en deux catégories...

ceux qui ont la classe et les autres

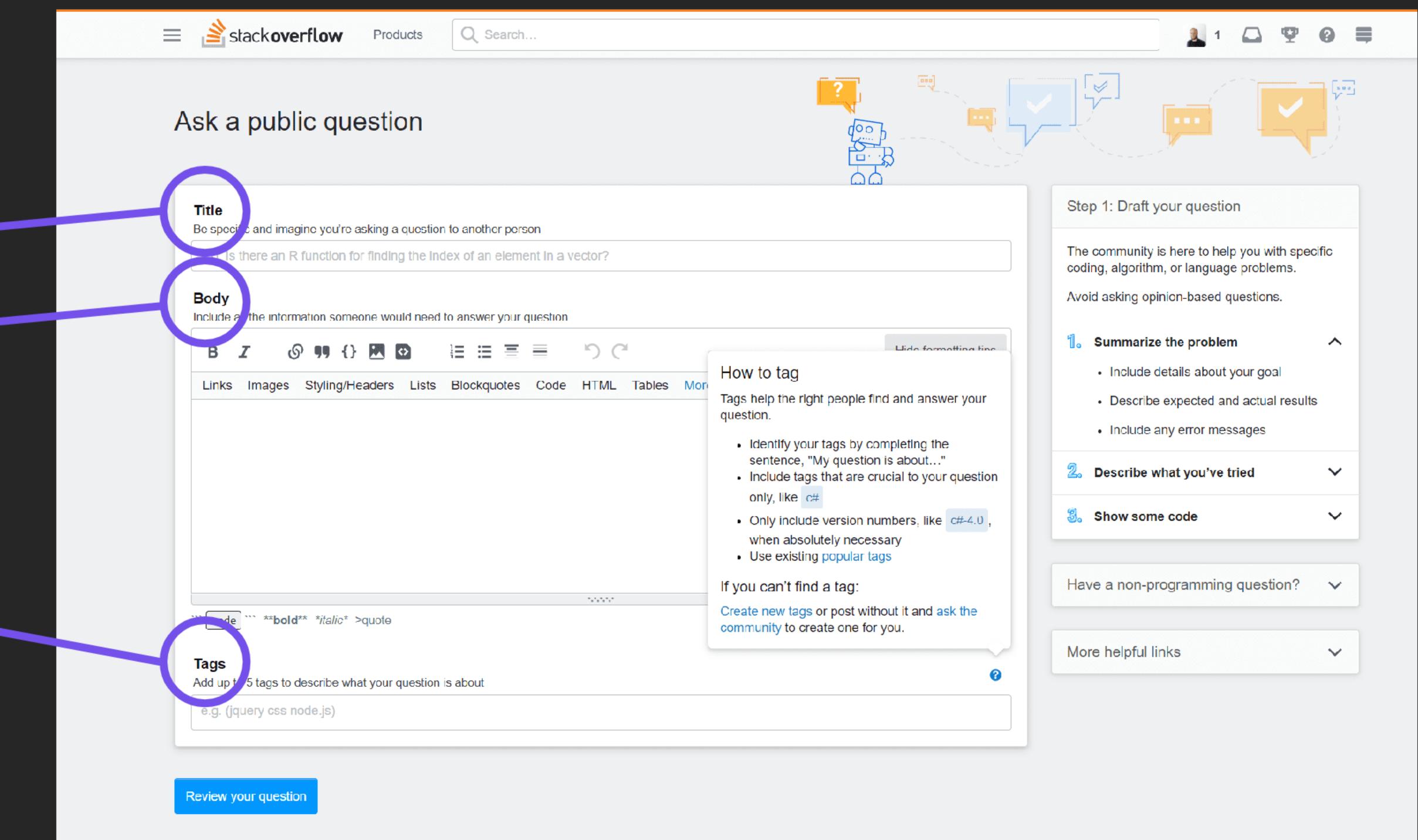
deux classes d'algorithmes d'apprentissage automatique à expérimenter

- non supervisé (modélisation de sujets sans étiquettes)
- supervisé (classification avec étiquettes)

Tu vois, le monde se divise en deux catégories...
ceux qui ont la classe et les autres

sur des données "textuelles"

- Title
- Body
- Tags

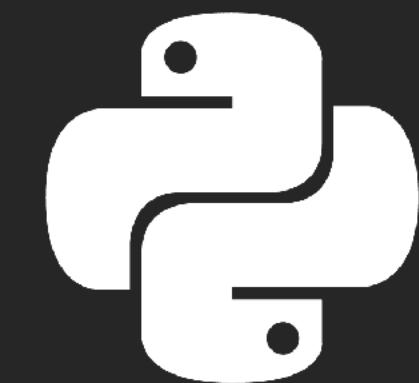


Tu vois, le monde se divise en deux catégories...

ceux qui ont la classe et les autres

sur des données "textuelles"

On va manipuler et traiter ces données à l'aide
d'outils spécifique à la langue, la librairie **NLTK**



Tu vois, le monde se divise en deux catégories...

ceux qui ont la classe et les autres

avec une contrainte spécifique que l'on s'impose

- utiliser un logiciel de gestion de versions

Tu vois, le monde se divise en deux catégories...

ceux qui ont la classe et les autres

git

pour le logiciel de gestion de versions



GitHub

pour la résilience et les nombreuses
fonctionnalités



github.com/Loridan/openclassrooms-iml-projects



Ingénieur Machine Learning

Data Scientist spécialisé dans les algorithmes d'apprentissage automatiques

Ce dépôt contient un ensemble de 8 projets professionnalisants autour des compétences clés de l'ingénieur machine learning

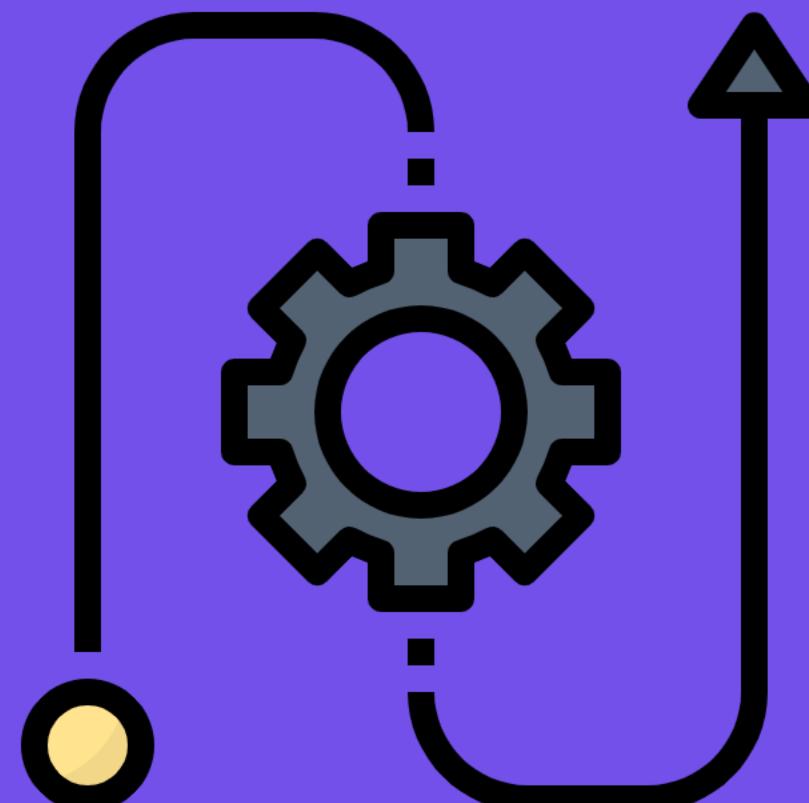
- PROJET 1**
Définissez votre stratégie d'apprentissage
- PROJET 2**
Concevez une application au service de la santé publique
- PROJET 3**
Anticipez les besoins en consommation électrique de bâtiments
- PROJET 4**
Segmentez des clients d'un site e-commerce
- PROJET 5**
Catégorisez automatiquement des questions
- PROJET 6**
Classez des images à l'aide d'algorithmes de Deep Learning
- PROJET 7**
Développez une preuve de concept
- PROJET 8**
Participez à une compétition Kaggle !

Parcours en partenariat avec CentraleSupélec

main	6 branches	0 tags	Go to file	Add file	Code
 Loridan	Loridan feat : update readme better gif	014f8c3 on 21 Oct 2020	4 commits		
	project1	feat: add folder structure	5 months ago		
	project2	feat: add folder structure	5 months ago		
	project3	feat: add folder structure	5 months ago		
	project4	feat: add folder structure	5 months ago		
	project5	feat: add folder structure	5 months ago		
	project6	feat: add folder structure	5 months ago		
	project7	feat: add folder structure	5 months ago		
	project8	feat: add folder structure	5 months ago		
	README.md	feat : update readme better gif	5 months ago		

La science des données

Expérimentation



- 1 Collecter des données
- 2 Préparer et explorer les données
- 3 Modéliser
- 4 Comprendre



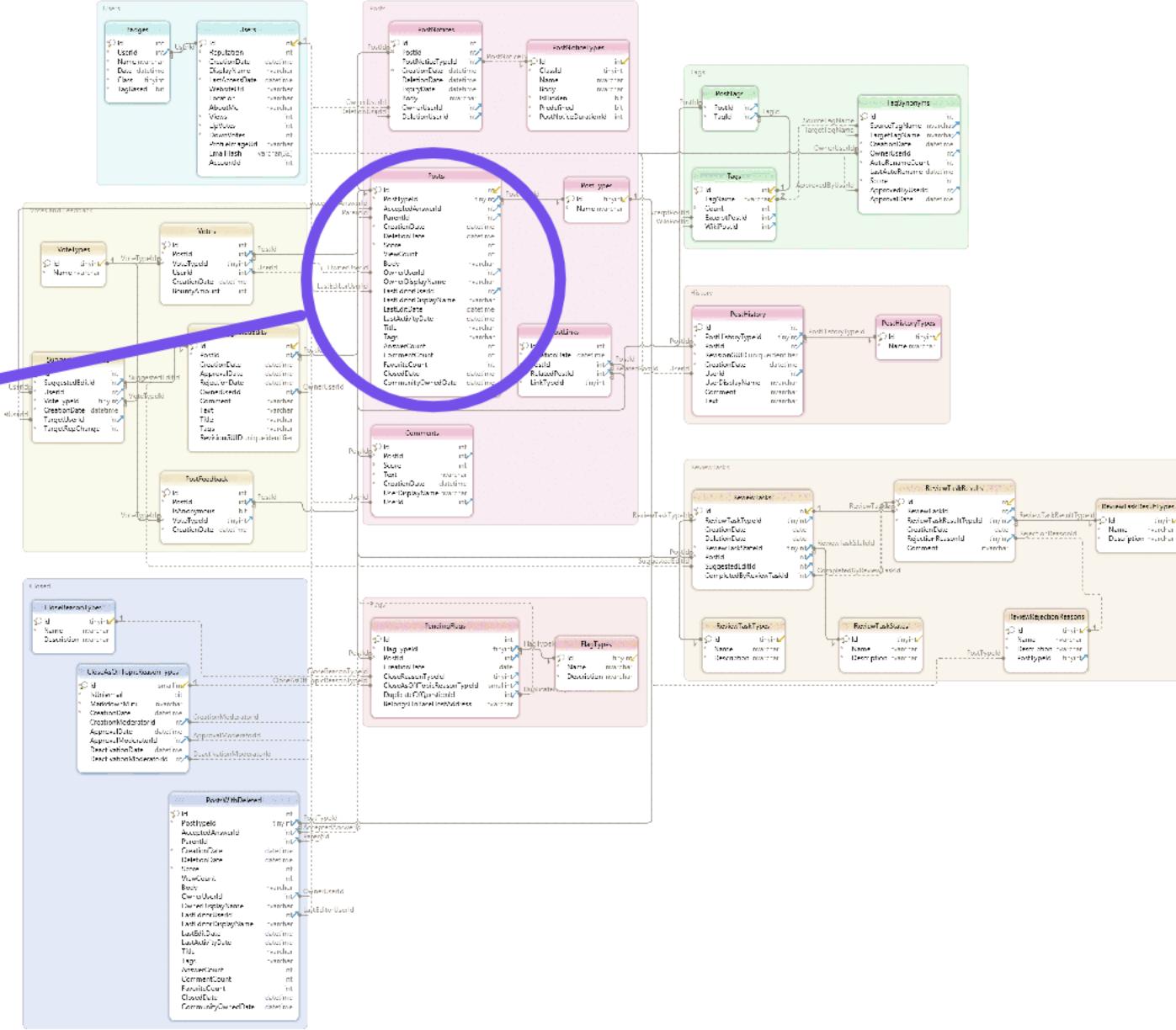
Lien de l'interface d'accès à la base de données

<https://data.stackexchange.com/stackoverflow/query/new>

Schéma de la base de données

On identifie les champs à récupérer sur nos données disponibles lors de la création d'une question.

Ce sont les "Posts" qui n'ont pas de "ParentId".





Lien de l'interface d'accès à la base de données
<https://data.stackexchange.com/stackoverflow/query/new>

Requête SQL

```
SELECT Id,Body,Title,Tags
FROM posts
WHERE Id < 1000000
AND ParentId IS NULL
```

Téléchargement

50000 rows au format .csv

The screenshot shows the StackExchange Data Explorer interface. In the center, there's an 'Editing Query' window with a purple circle highlighting the query text:

```
Project_5
call description
1 SELECT Id,Body,Title,Tags
2 FROM posts
3 WHERE Id < 1000000
4 AND
5 ParentId IS NULL
```

To the right, a 'Database Schema' pane lists the 'Posts' table with its columns: Id (int), PostTypeId (tinyint), AcceptedAnswerId (int), ParentId (int), CreationDate (datetime), DeletionDate (datetime), Score (int), ViewCount (int), Body (nvarchar(max)), and Revisions (table). Below the schema, a list of revisions is shown, with the first four highlighted in yellow. At the bottom, there are buttons for 'Run Query', 'Cancel', and 'Options' (with checkboxes for 'Text only results' and 'Include execution plan').

The screenshot shows the results of the executed query. The table has columns: Id, Body, Title, and Tags. The 'Tags' column is highlighted with a purple circle. The data includes several posts with their respective bodies, titles, and tag lists.

ID	Body	Title	Tags
365473	<p>I want to handle different types of docs like...	generic type dependency injection. How to inj...	inversion of control, container, generic, type, dependency, injection, how, to, inj...
365475	<p>In a php application, I use the following re...	Url rewrite with mod_jk	apache, url, rewrite, mod_jk
365477	<p>In Silverlight (and I guess WPF) why are t...	In Silverlight why are some properties prefixe...	wpf, xml, silverlight
365480	<p>I've got a Core Data application that has ...	Can you Bind to the timeInterval attribute of a...	objective-c, coredata, core-data, osx-leopard
365492	<p>I have read the following properties from ...	I how to read terminalservices IADs IISUser...	ce, map, ca-2.0, terminal-services, iads, userex

Dépasser la limite

Au lieu d'utiliser l'interface de stackoverflow, il est possible de récupérer directement dans un dataframe avec de multiples requêtes python sur l'api afin de dépasser la limite de 50 000 rows, puis d'exporter pour traitement et modélisation.



```
import requests
import time
import pandas as pd

#resoudre captcha sur site avant et recuper id sinon error 500 'https://data.stackexchange.com/stackoverflow'
def run_query_stackoverflow(sql_query, id_captcha):

    url = f'https://data.stackexchange.com/query/save/1/{id_captcha}'
    data = {"title": "P05", "description": "", "sql" : sql_query, "g-recaptcha-response" : ""}

    response = requests.post(url, data)
    job_id = response.json().get('job_id')

    if(job_id is not None) :
        url = f'https://data.stackexchange.com/query/job/{job_id}'
        time.sleep(3)
```

**Quelle est la réglementation sur la protection de ces données ?**

Faible, CC BY-SA 3.0 <https://stackoverflow.com/legal/terms-of-service/free>

Où sont stockées les données ?

Elles sont sur les serveurs distants publiques et sur notre ordinateur en sdd local.

Elles sont également disponible sur dépôt kaggle.

<https://www.kaggle.com/stackoverflow/stackoverflow>

Quel est le type des données ?

structurée

Nettoyage des données textuelles

Importer et observer les données



Identifier et traiter les doublons

Identifier et traiter les valeurs manquantes

Body et Title

Appliquer une fonction de nettoyage :

- supprimer les balises html
- convertir le type de casse à 'lower'
- supprimer à l'aide d'une regex les chiffres et ponctuations
- supprimer les mots courants anglais à l'aide de nltk
- supprimer les mots inutiles identifiés*

Tags

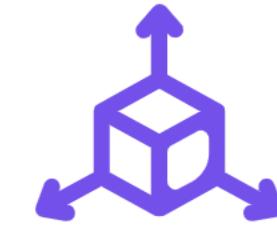
- retirer les chevrons
- séparer sous forme de liste



P05_01_notebook_nettoyage_exploration.ipynb



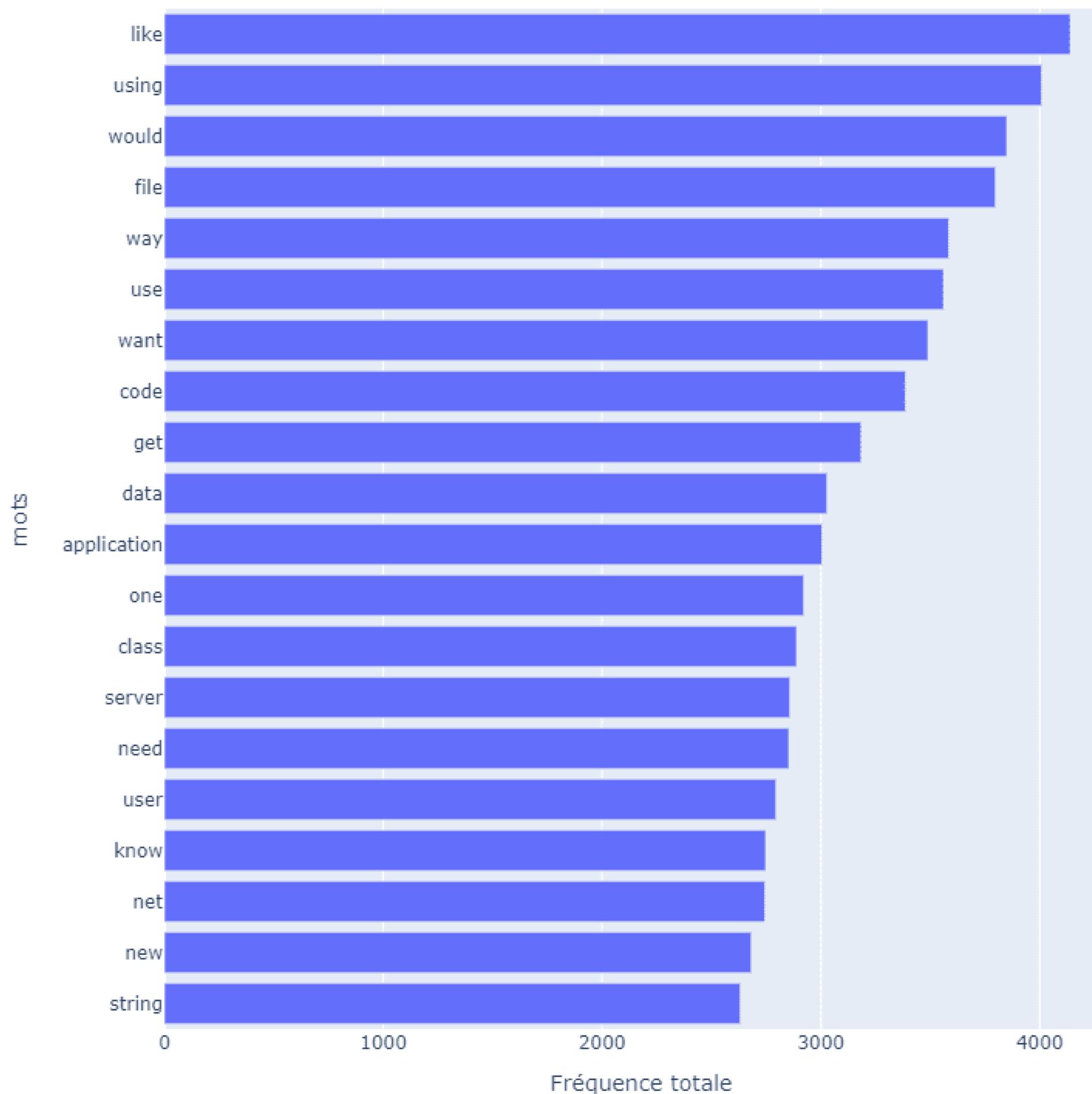
Techniques d'extraction de caractéristiques



- Sac de mots (Bag-of-words model)
- TF-IDF (Term Frequency - Inverse Document Frequency)

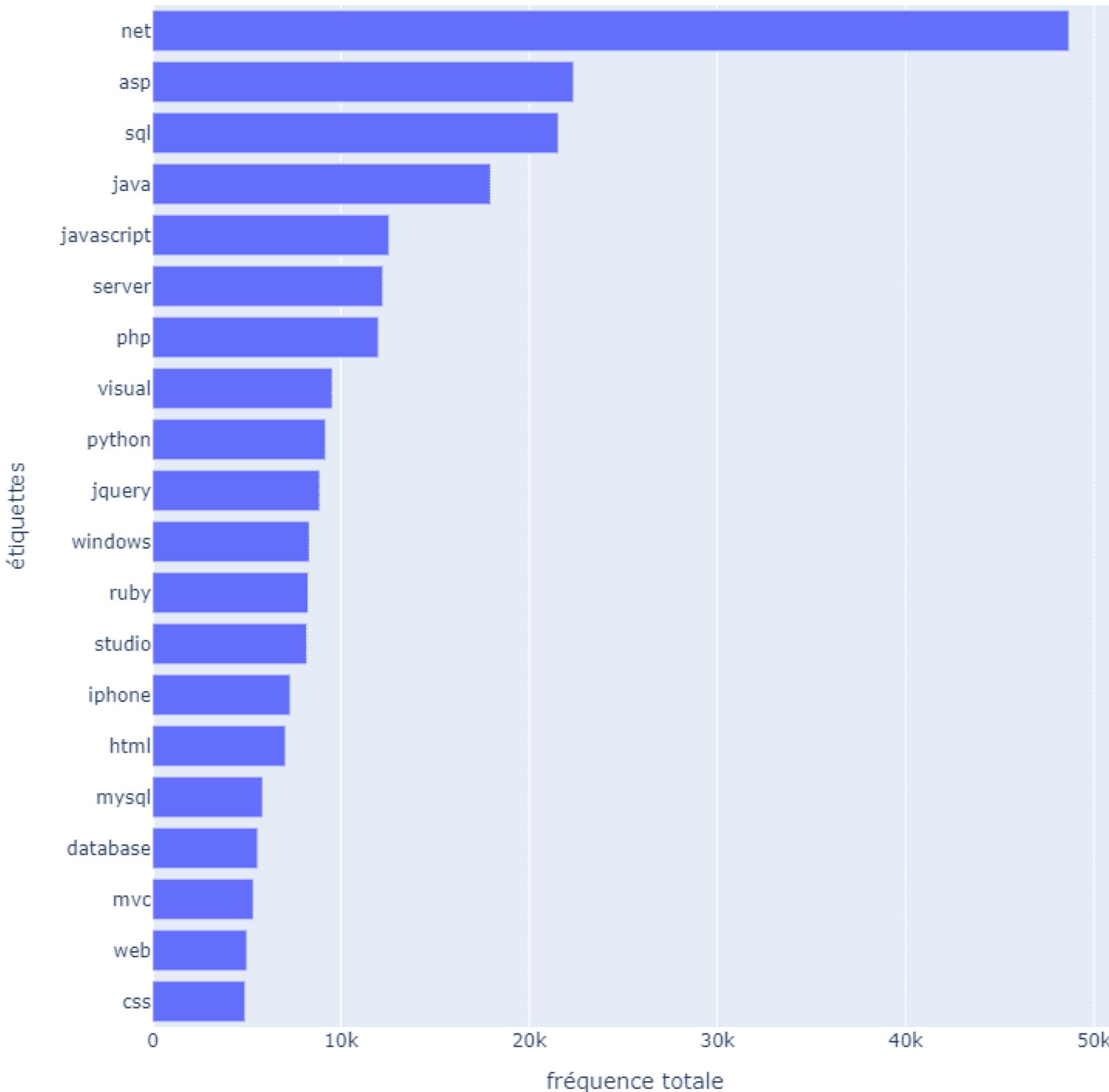


Analyser les $X=20$ mots les plus fréquents



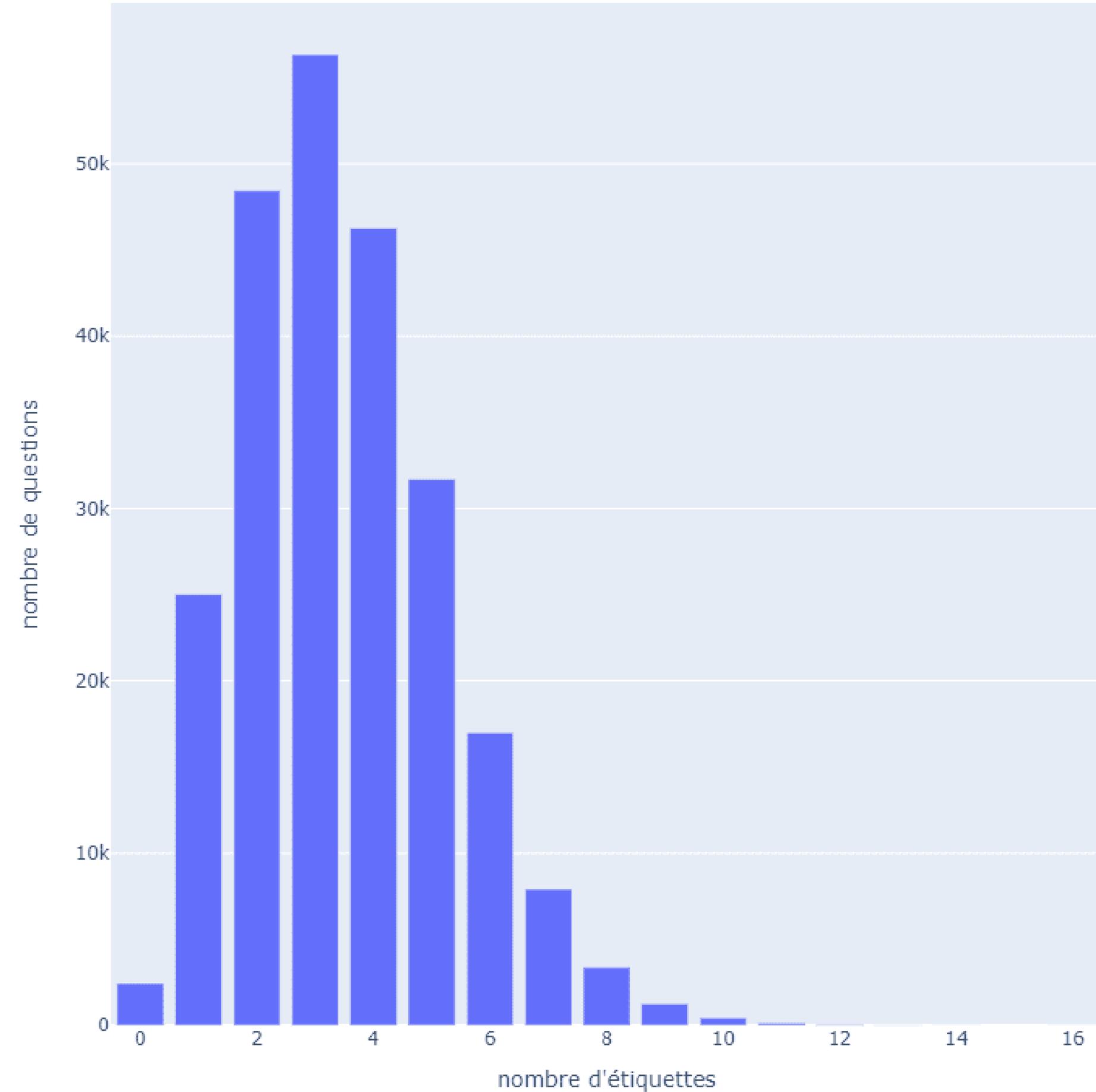


Analyser les X=20 étiquettes les plus fréquentes





Analyser le nombre de questions / nombre d'étiquettes



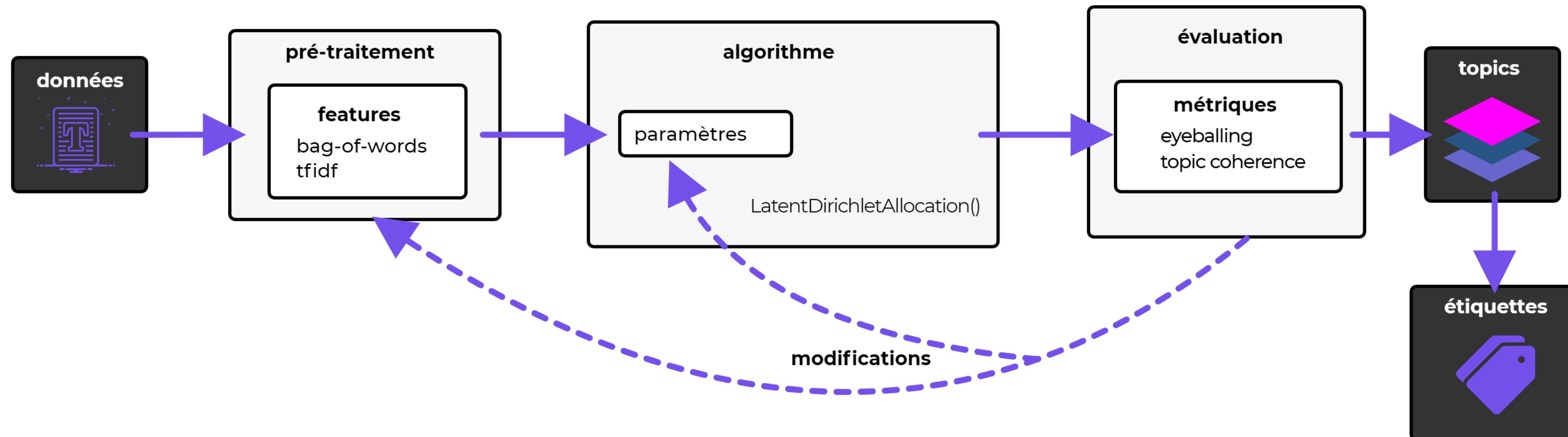


Nettoyage : identifier des mots inutiles

```
: my_stopwords = ['like', 'would', 'want', 'using', 'need', 'thanks']
my_stopwords = pd.Series(my_stopwords)
my_stopwords.to_csv("data/my_stopwords.csv", index=False)
```



Process de modélisation non supervisée (dév)





Latent Dirichlet Allocation

Entrée [8]: `display_topics(lda, lda_posts_body_feature_names, 10)`

Topic 0:

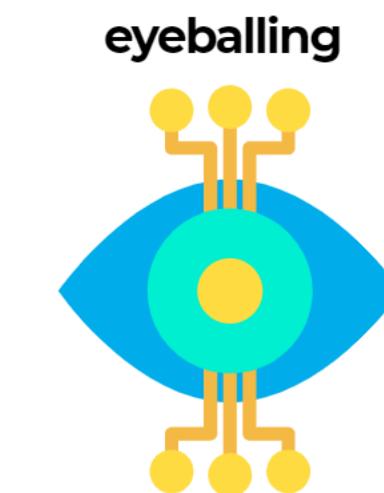
user session login application password email page pdf web authentication

Topic 1:

database data sql table tables db entity server use way

Topic 2:

use good know one net code web looking question application

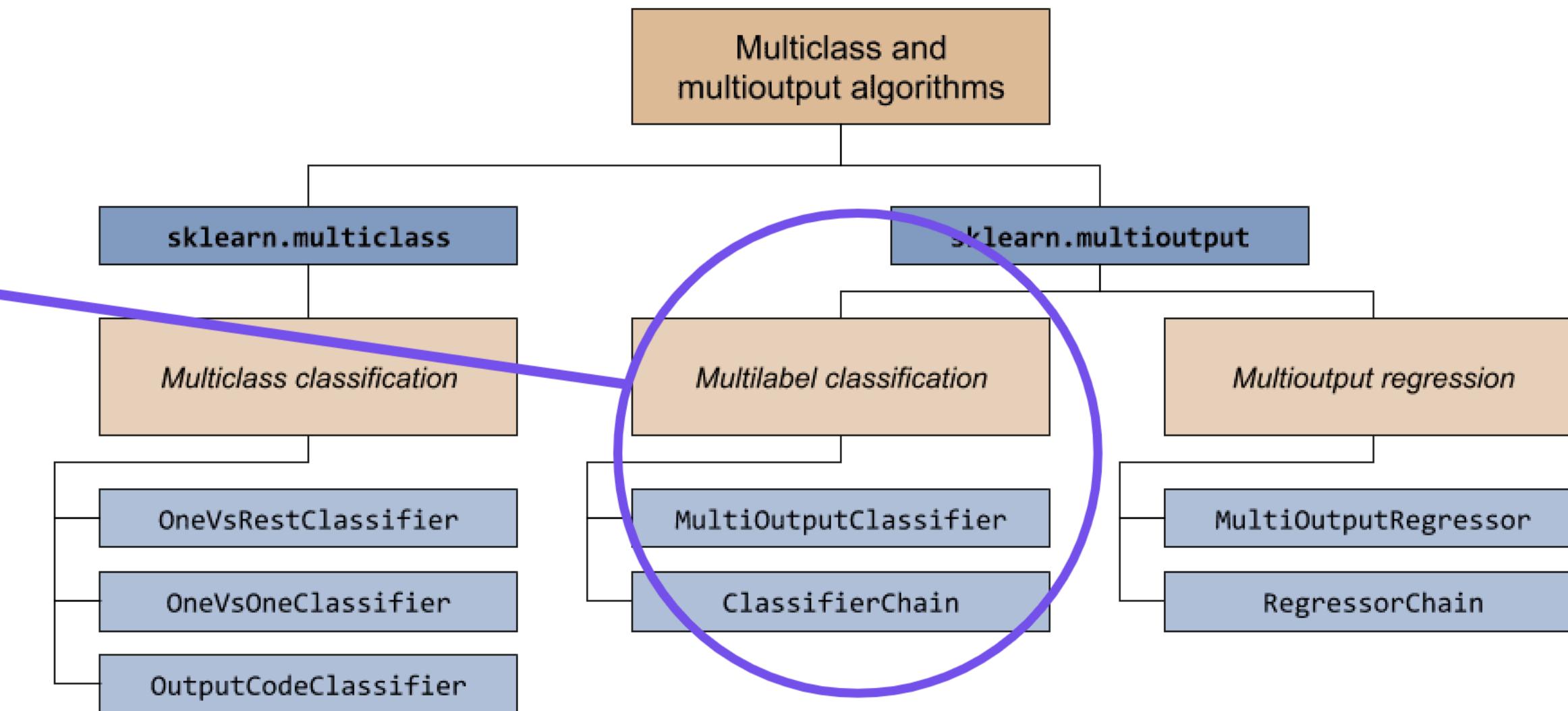




Process de modélisation supervisée (dév)

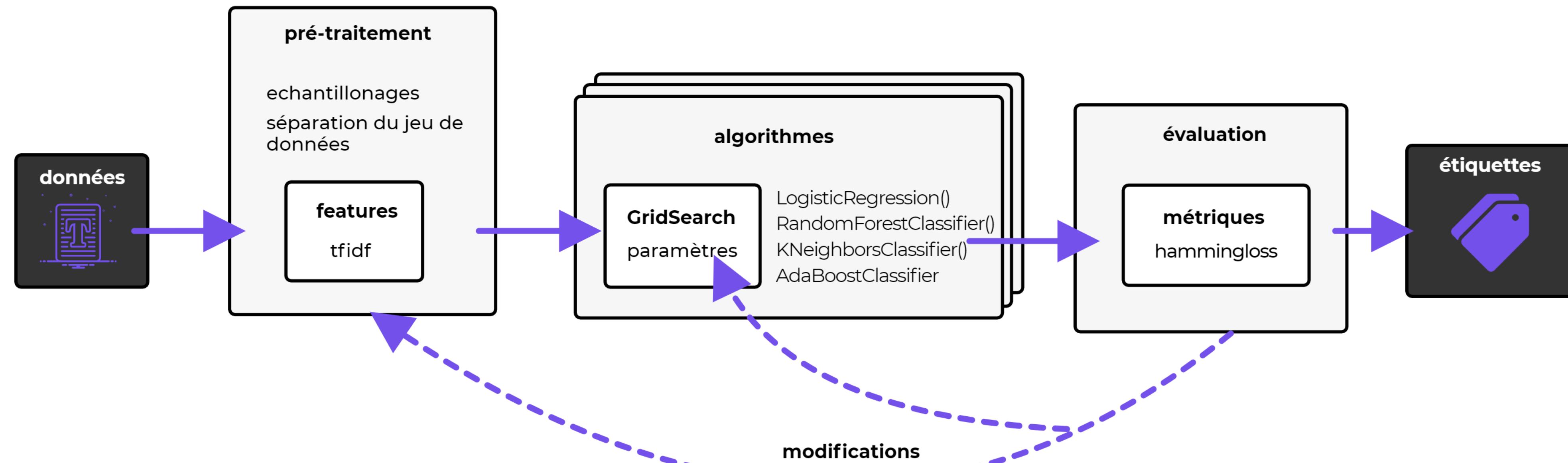
La Multilabel classification est la tâche qui consiste à assigner un ou plusieurs labels ayant pour valeur 0 ou 1. Dans notre cas, cela correspond à la présence de l'étiquette ou de son absence.

Certains estimateurs sont capables de générer d'office plusieurs étiquettes, d'autres ont besoin de méta-estimateur comme le MultiOutputClassifier.





Process de modélisation supervisée (dév)





modifications : réduire la taille du vocabulaire

Nous pouvons réduire la taille du vocabulaire en supprimant les mots étant présents que dans moins de 10 questions et ceux étant présents dans plus de 50% des questions, c'est une réduction de dimensionnalité.

CountVectorizer() paramètres :

max_df = 0.5
min_df = 10

On passe d'une shape (240 399 , 319 814) à (240 399 , 27 119), soit 8% du vocabulaire initial.



métrique : Hamming Loss

Dans la classification multilabel, les prédictions sont des vecteurs d'étiquettes et, par conséquent, la prédiction peut être entièrement correcte, partiellement correcte ou totalement incorrecte. Les métriques par défaut accuracy, precision, recall ne captent pas cette notion.

Nous allons utiliser la métrique "Hamming Loss" qui est la fraction d'étiquettes qui sont incorrectement prédites sur l'ensemble de nos vecteurs. Elle prendra donc en compte les faux positifs et les faux négatifs. Elle est comprise entre 0 et 1, une HL = 0 indique l'absence totale d'erreur, cad plus sa valeur est petite plus la classification est performante.

```
Entrée [69]: y_pred = [ [0,1,1,0] , [0,0,1,1] , [0,0,0,0] ]
y_true = [ [1,1,0,0] , [0,1,1,1] , [0,0,0,1] ]
hamming_loss(y_pred,y_true)

Out[69]: 0.3333333333333333
```

exemple:

$(3,4) = 3*4 = 12$ labels prédits
on compte 4 erreurs de prédictions sur nos 12 labels

$4/12 = 1/3 = 0.33333333$



base line pour notre comparatif

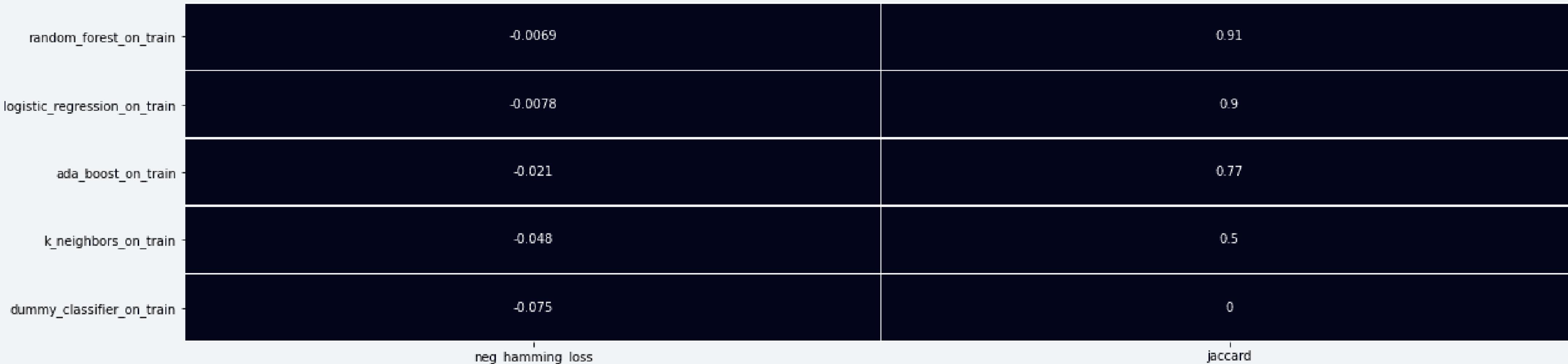
On va utiliser également un `DummyClassifier()` qui effectue des prédictions à l'aide de règles simples. Ce classificateur est utile comme base de référence simple pour comparer avec d'autres classificateurs.

Strategy = “most_frequent”: prédit toujours l'étiquette la plus fréquente dans l'ensemble d'apprentissage.

Dans notre cas, il va toujours prédire zéro 0 étant donné que l'on a une matrice creuse.



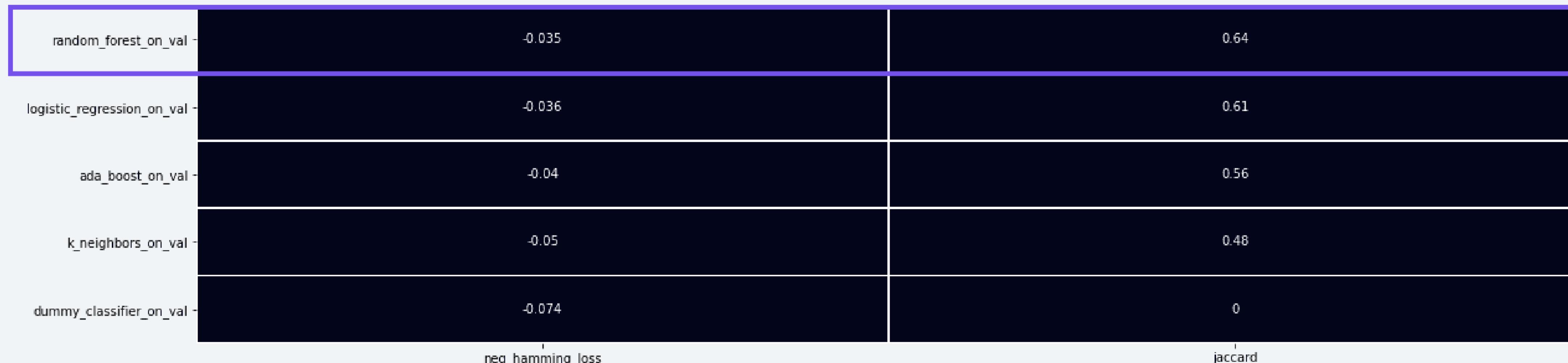
Comparaison de nos modèles : train set





Comparaison de nos modèles : validation set

Le modèle ensembliste Random forest est le plus performant

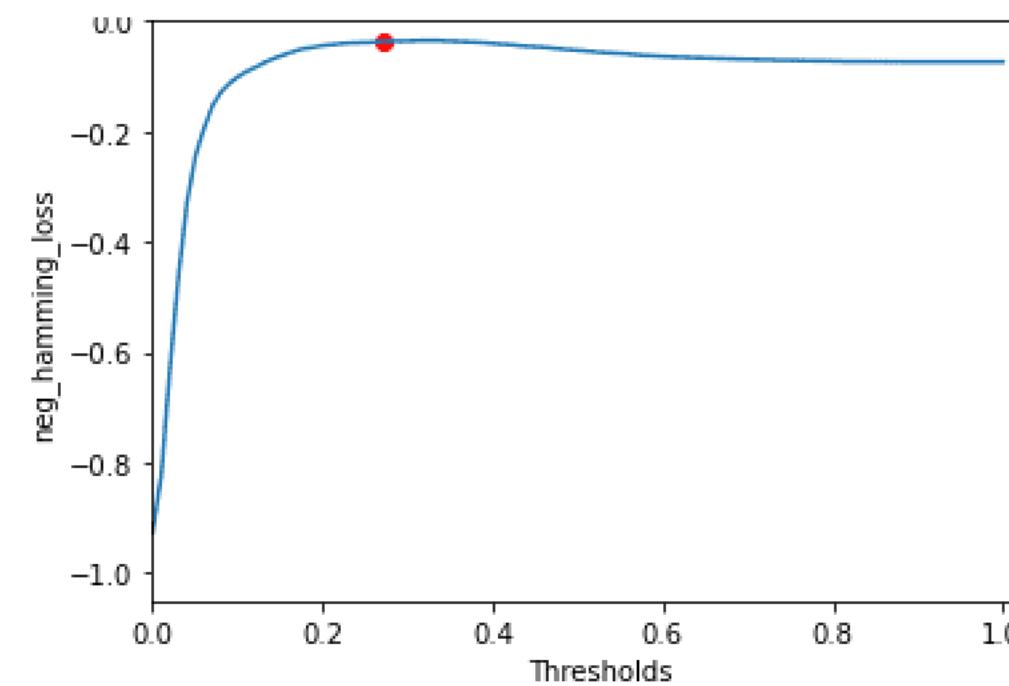




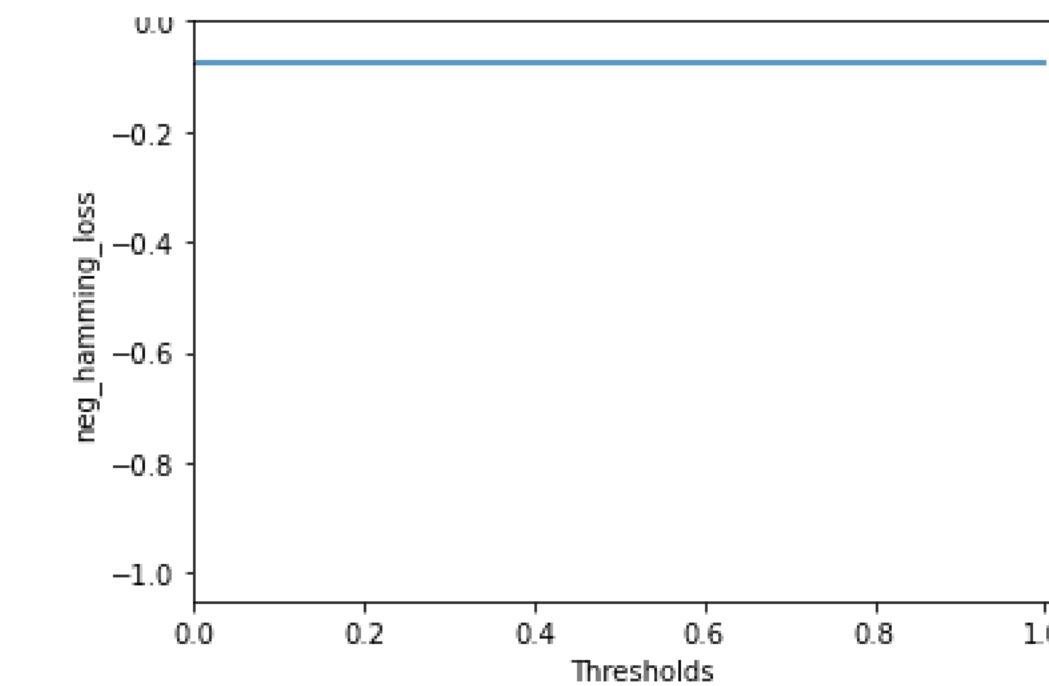
Hamming Loss

Le threshold est déterminé par la meilleure performance possible sur notre métrique Hamming Loss

randomforest



baseline

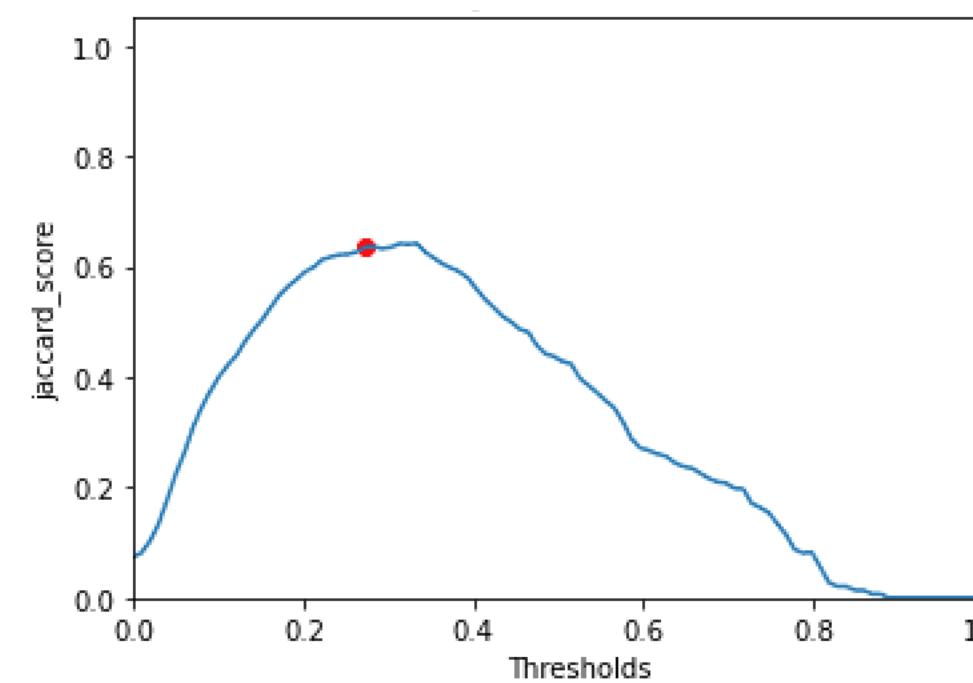




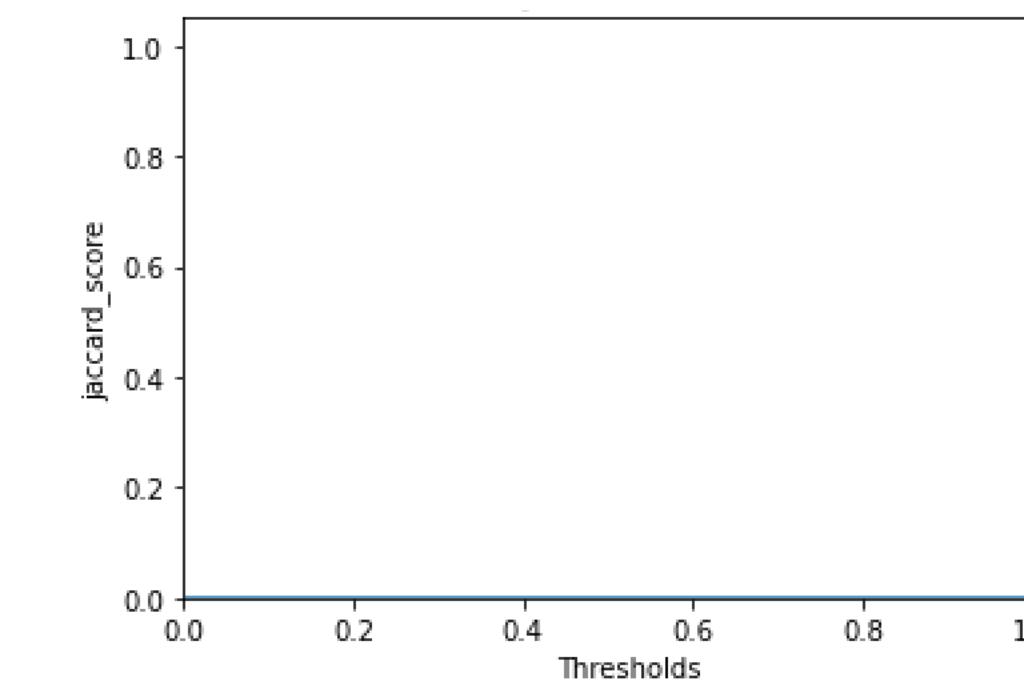
Jaccard

Une fois notre threshold déterminé, on peut observer et comprendre sur d'autres métriques

randomforest



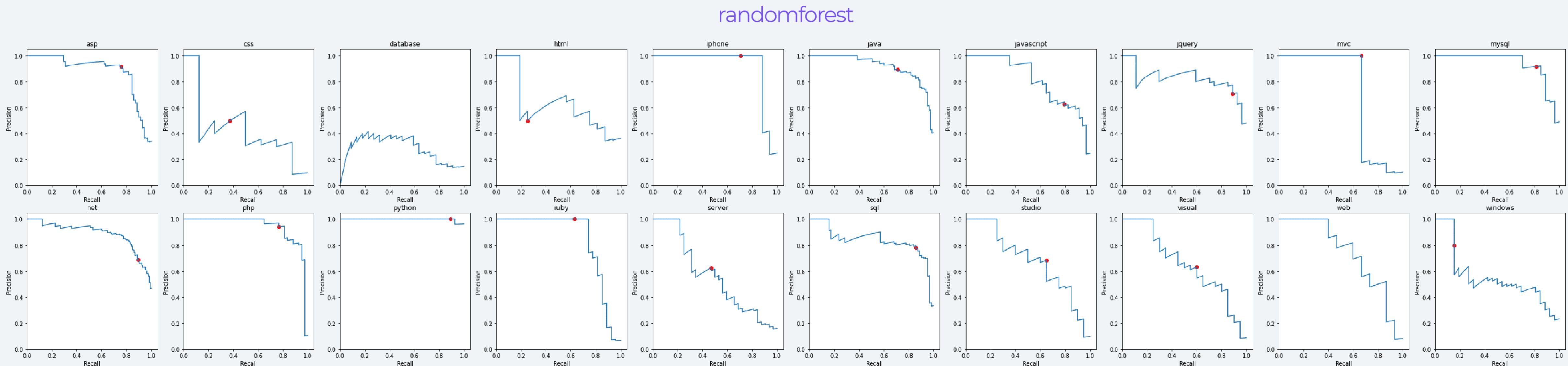
baseline





Courbe precision-recall

On peut voir la performance de prédictions pour nos 20 étiquettes séparément.
Plus le point rouge est proche du coin haut droit, plus la prédiction est performante.
Sa position indique le trade-off entre faux positifs et faux négatifs.

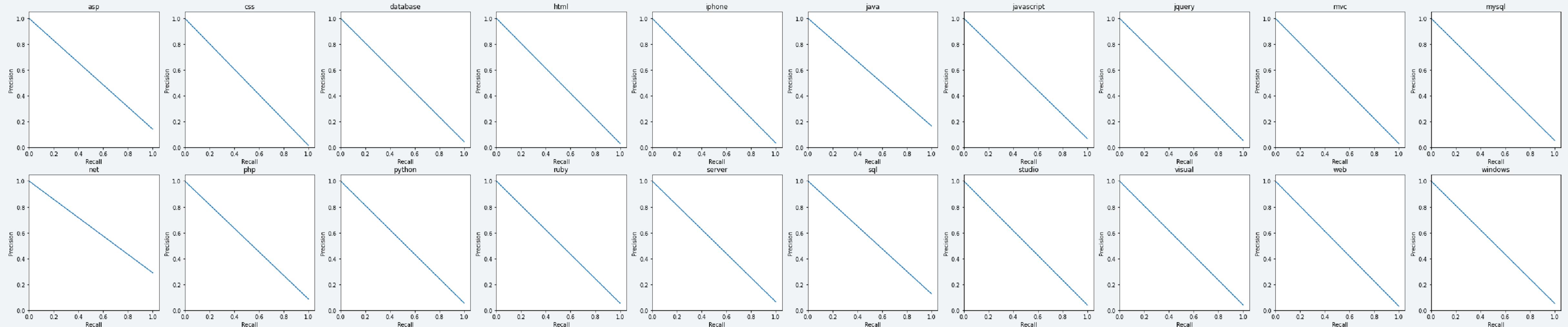




Courbe precision-recall

On peut voir la performance de predictions pour nos 20 étiquettes séparemment.
Plus le point rouge est proche du coin haut droit, plus la prédiction est performante.
Sa position indique le trade-off entre faux positifs et faux négatifs.

baseline



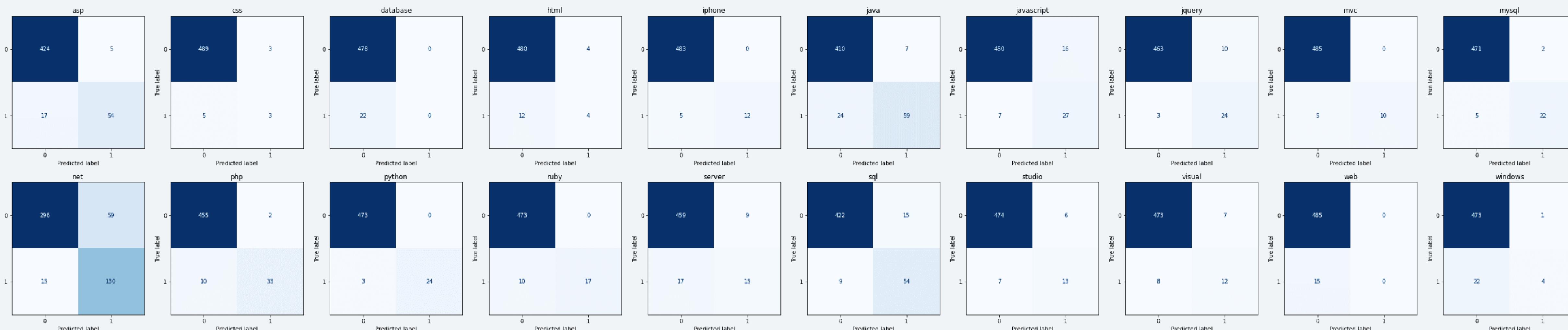


matrice de confusion

On voit bien pour chaque étiquette, le jeu n'est pas équilibré, la classe positive étant peu présente.

La matrice de confusion de baseline nous montre que la strategy most frequent est respectée.

randomforest



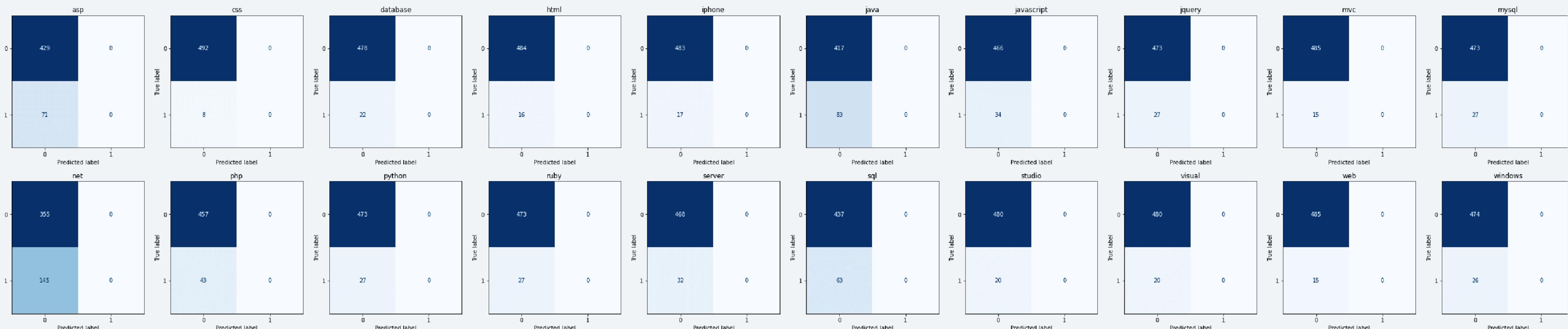


matrice de confusion

On voit bien pour chaque étiquette, le jeu n'est pas équilibré, la classe positive étant peu présente.

La matrice de confusion de baseline nous montre que la strategy most frequent est respectée.

baseline





Conclusion

Nous avons réussi à réaliser un modèle de prédictions assez performant pour être utilisé en production. Notre choix s'est porté sur un algorithme de classification supervisée, le random forest. On peut afficher une liste d'étiquettes de suggestions pour aider l'utilisateur à remplir le formulaire. Des améliorations sont évidemment possibles, on peut continuer à expérimenter, choisir un vocabulaire plus restreint, augmenter la taille des échantillons, proposer plus d'étiquettes, revoir l'implémentation du multioutput, appliquer un threshold différent par label, etc.

Tu vois, le monde se divise en deux catégories...

ceux qui ont la classe et les autres

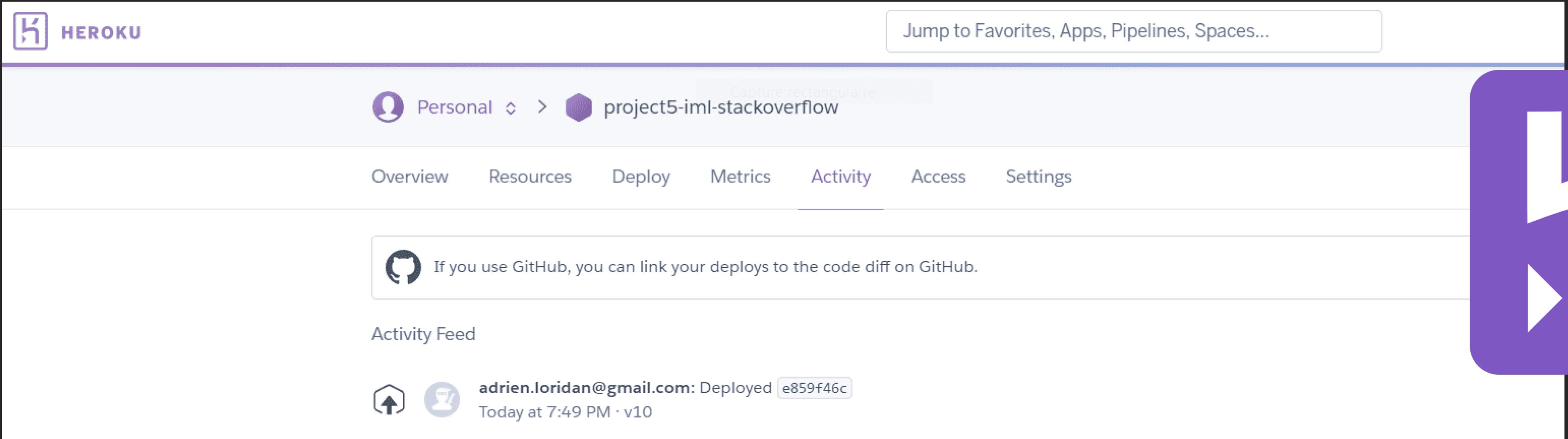
code final à déployer

Le code final est un serveur flask qui a été déployé sur [Heroku](#),
Le random forest trop gourmand en mémoire n'a pas pu fonctionné.
Le multioutput logistic regression a donc été choisi pour sa compacité.

```
2021-05-01T17:34:25.892170+00:00 heroku[web.1]: Process running mem=1279M(250.0%)
2021-05-01T17:34:25.939108+00:00 heroku[web.1]: Error R15 (Memory quota vastly exceeded)
2021-05-01T17:34:25.941747+00:00 heroku[web.1]: Stopping process with SIGKILL
2021-05-01T17:34:25.965790+00:00 heroku[router]: at=error code=H13 desc="Connection closed without response" host=precise-ml.stack...
```



Tu vois, le monde se divise en deux catégories...
ceux qui ont la classe et les autres



The screenshot shows the Heroku web interface for a project named "project5-iml-stackoverflow". The "Activity" tab is selected in the navigation bar. A callout bubble from GitHub suggests linking deployments to code diff on GitHub. The activity feed shows a deployment by "adrien.loridan@gmail.com" at 7:49 PM today, with commit hash "e859f46c". A large purple Heroku logo is overlaid on the right side of the screenshot.

Tu vois, le monde se divise en deux catégories...

ceux qui ont la classe et les autres

api de test url

[https://project5-iml-stackoverflow.herokuapp.com/
tags?question=As everyone else suggested you, I suggest asp.net learning pages](https://project5-iml-stackoverflow.herokuapp.com/tags?question=As%20everyone%20else%20suggested%20you,%20I%20suggest%20asp.net%20learning%20pages)

```
{  
  "nombre_tags": 2,  
  "tags": "asp net"  
}
```



Projet 5

Thank you !

