

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.

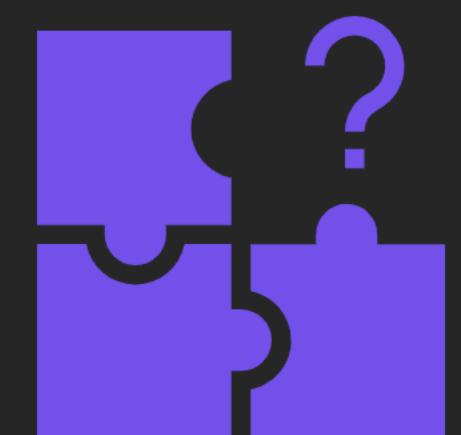
Elles permettent un regroupement facile des questions lors d'une recherche ultérieur par l'utilisateur.

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 ?

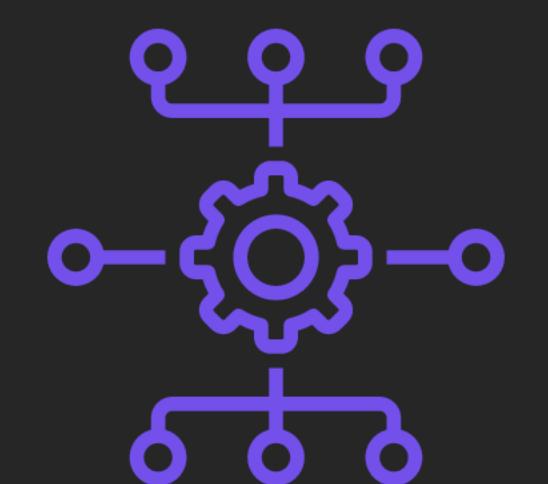


Tu vois, le monde se divise en deux catégories...

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

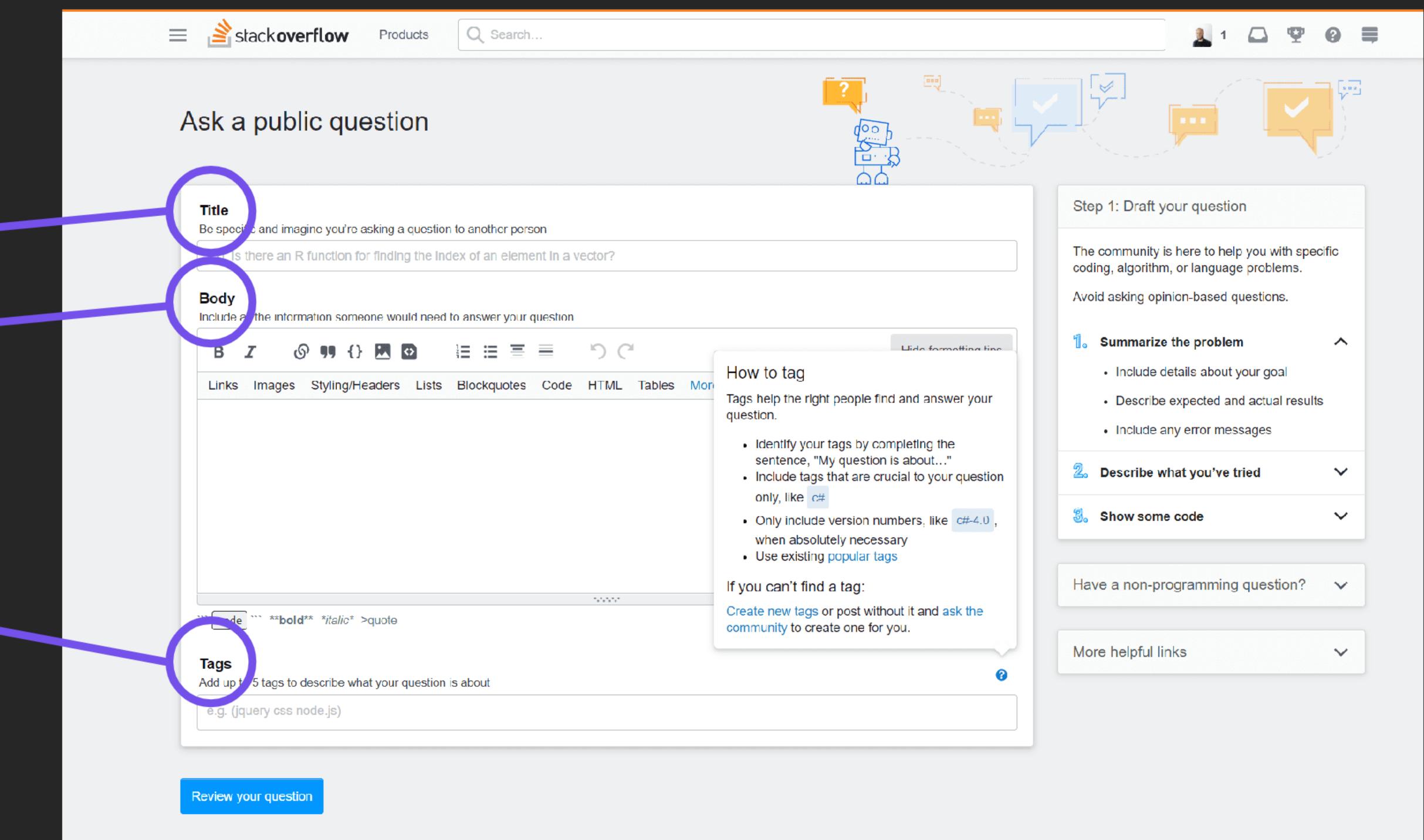
deux classes d'algorithmes d'apprentissage automatique à expérimenter

- 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



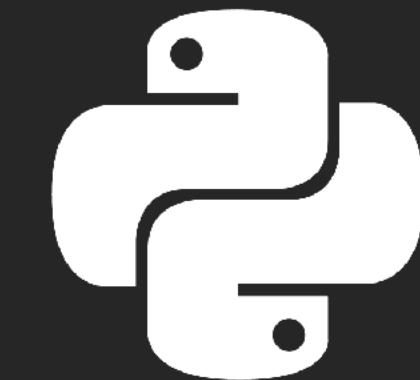
Tu vois, le monde se divise en deux catégories...

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, les librairies
spaCy et **NLTK**

spaCy



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

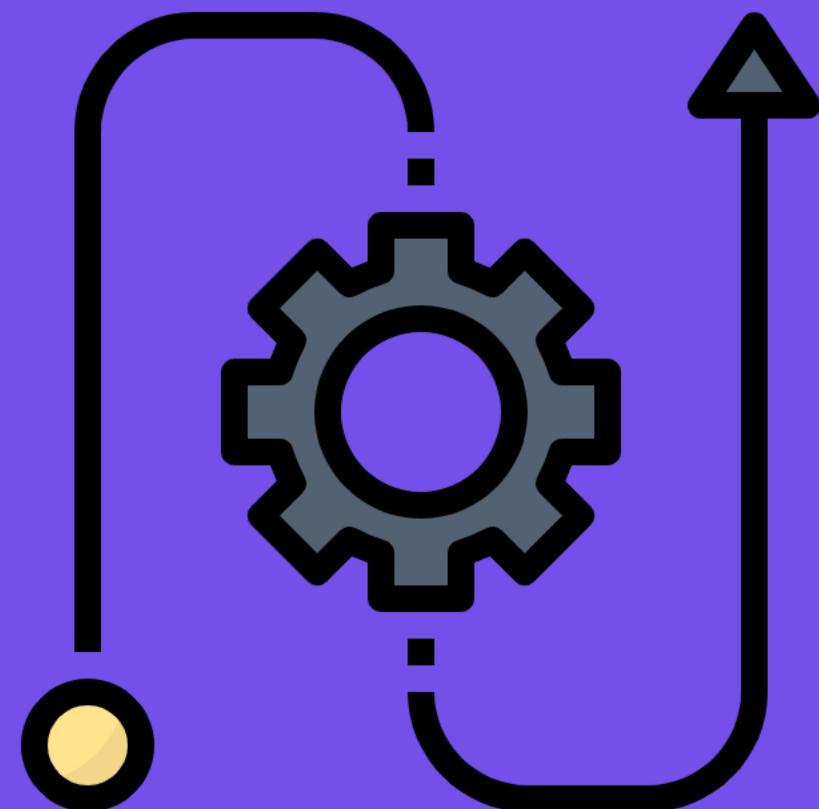


github.com/Loridan/openclassrooms-iml-projects

	main	6 branches	0 tags	Go to file	Add file	Code
 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 the following SQL code:

```
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 of the query editor is a 'Database Schema' pane for the 'Posts' table, which includes columns like Id, PostTypeId, AcceptedAnswerId, ParentId, CreationDate, DeletionDate, Score, ViewCount, Body, and Tags. Below the schema is a list of revisions for a specific post, with revision 1710626 highlighted.

The screenshot shows the results of the executed query. The table has columns for Id, Body, Title, and Tags. A 'Download CSV' button is visible at the top right of the results area.

		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, container
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, core-data, osx-leopard	core-data, osx-leopard
365492	<p>I have read the following properties from ...	I how to read terminalsServices IADs I SUser...	ce, ldap, ca-2.0, terminal-services, iads, suserex	ce, ldap, ca-2.0, terminal-services, iads, suserex

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 ?

non 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

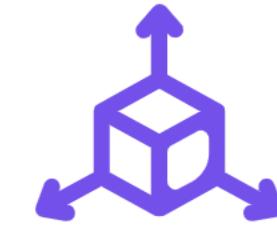
- retirer les chevrons
- séparer sous forme de liste



P05_01_notebook_nettoyage_exploration.ipynb



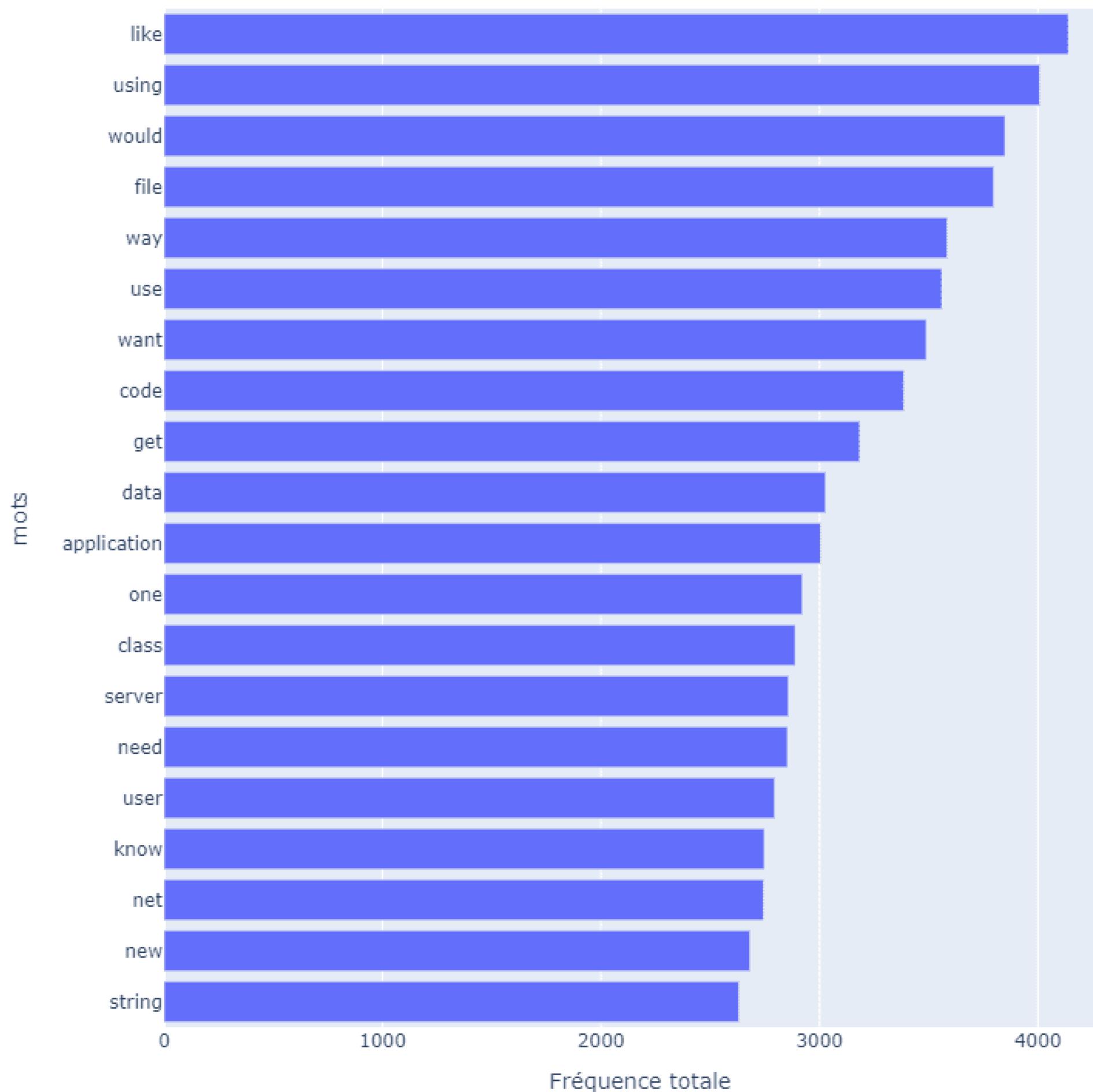
Techniques d'extraction de caractéristiques



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

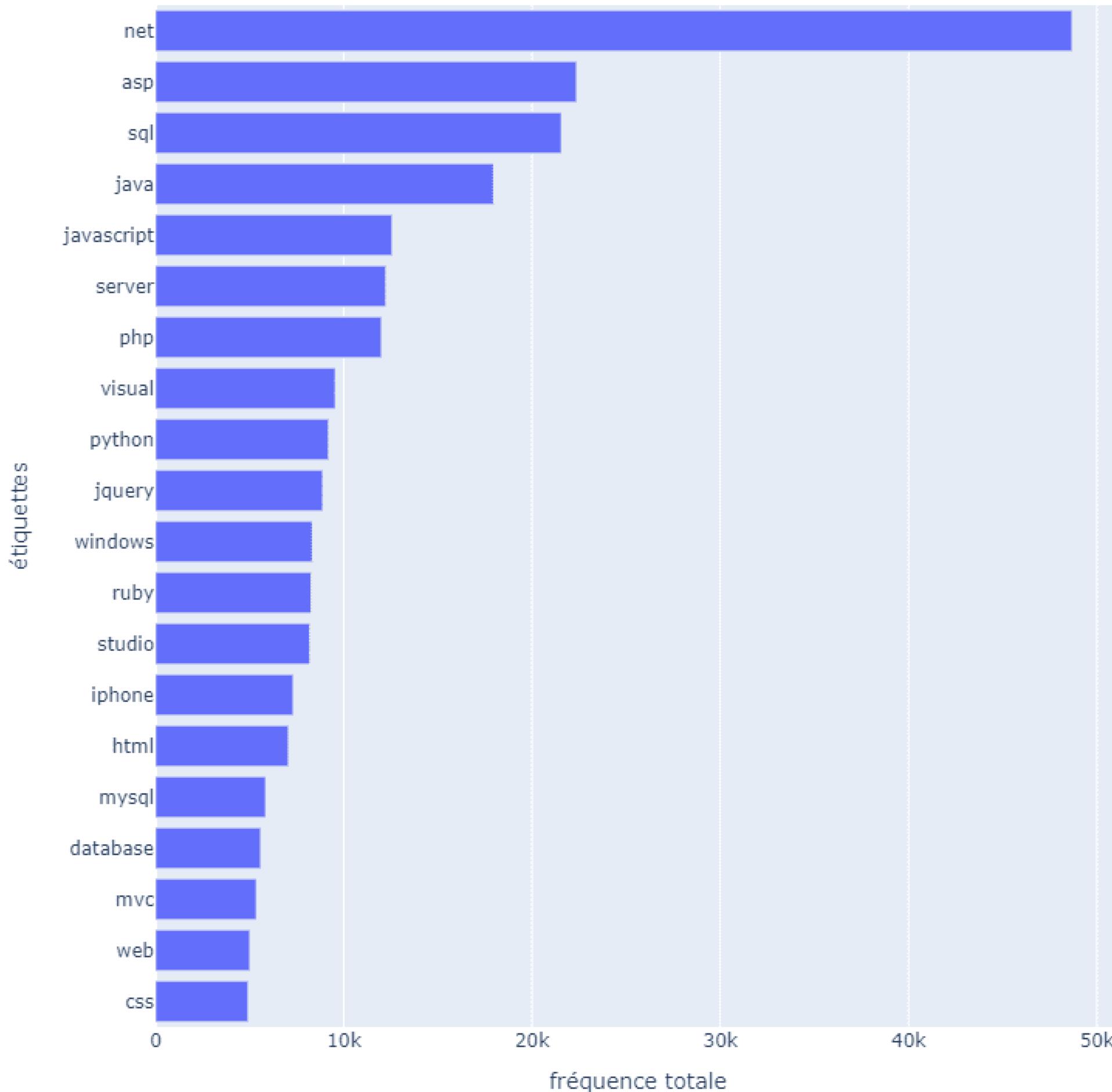


Analyser les $X=20$ mots les plus fréquents





Analyser les X=20 étiquettes les plus fréquentes





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)
```



Réduire la taille du vocabulaire

La Document frequency est le nombre de documents contenant le mot , 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.

CountVectorizer() paramètres :

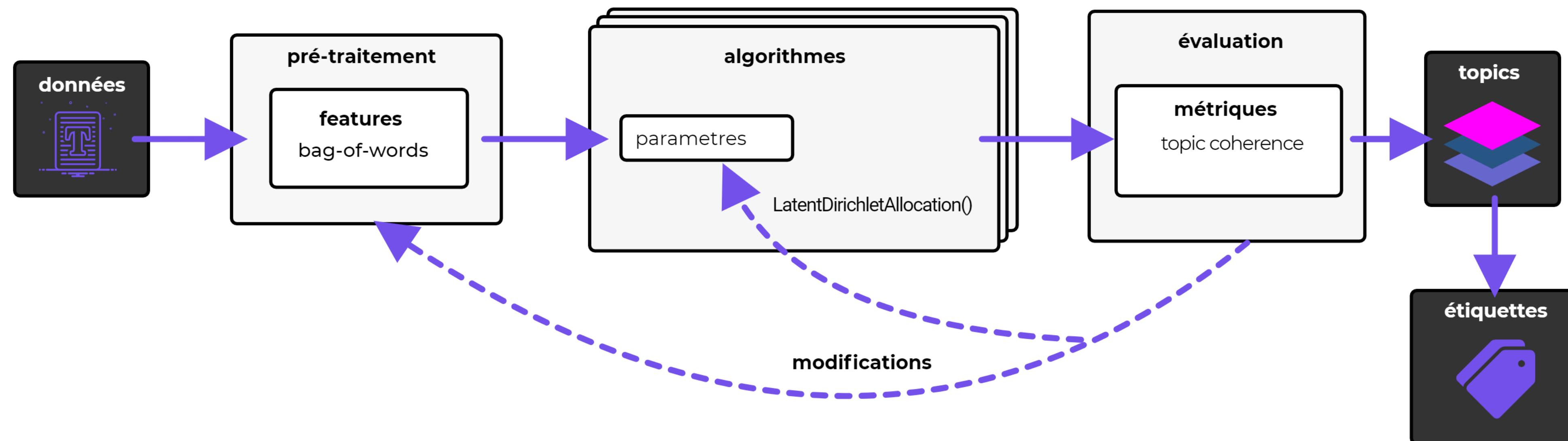
max_df = 0.5

min_df = 10

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



Process de modélisation non supervisée (dév)

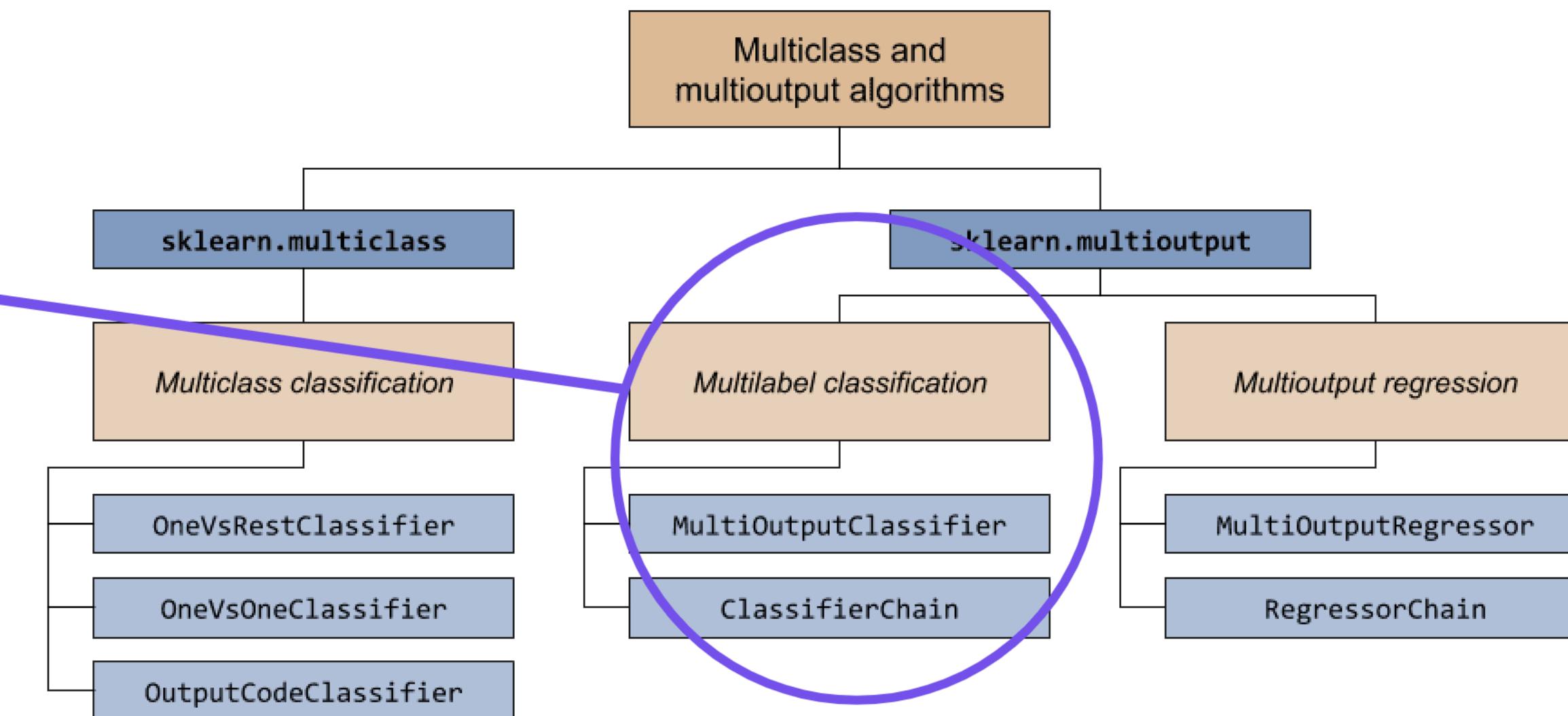




Process de classification supervisée (dév)

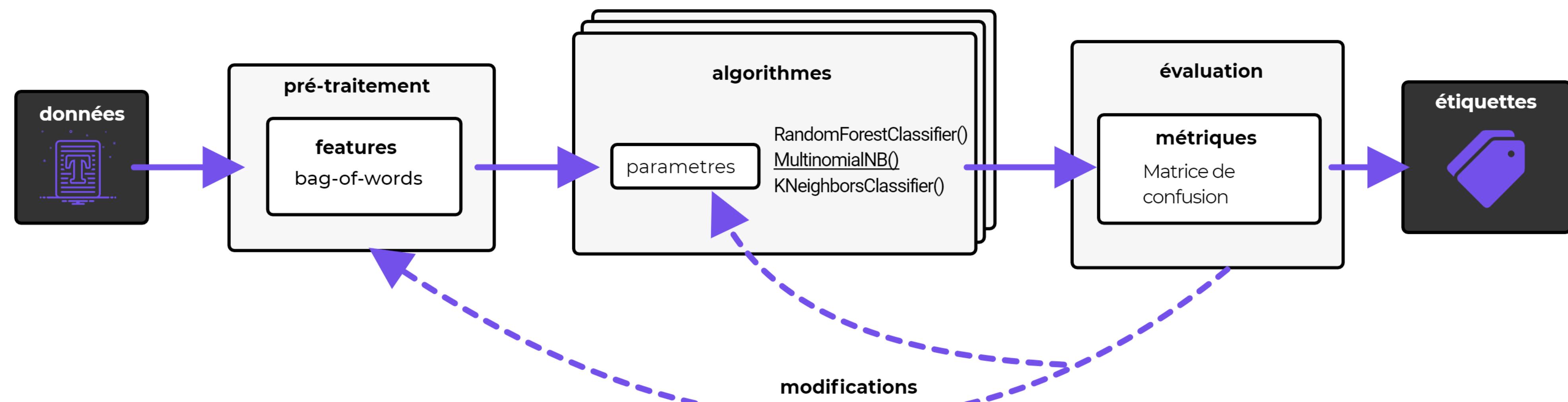
La Multilabel classification est la tâche qui consiste à assigner un ou plusieurs labels ayant pour valeur 0 ou 1. Dans notre cas, cela corresponds à la présence de l'étiquette ou de son absence.

Certains estimateurs sont capables de générer d'office plusieurs étiquettes, d'autres ont besoin de métaclassificateur comme le MultiOutputClassifier ou le ClassifierChain pour la corrélation entre étiquettes.





Process de classification supervisée (dév)



Projet 5

Thank you !

