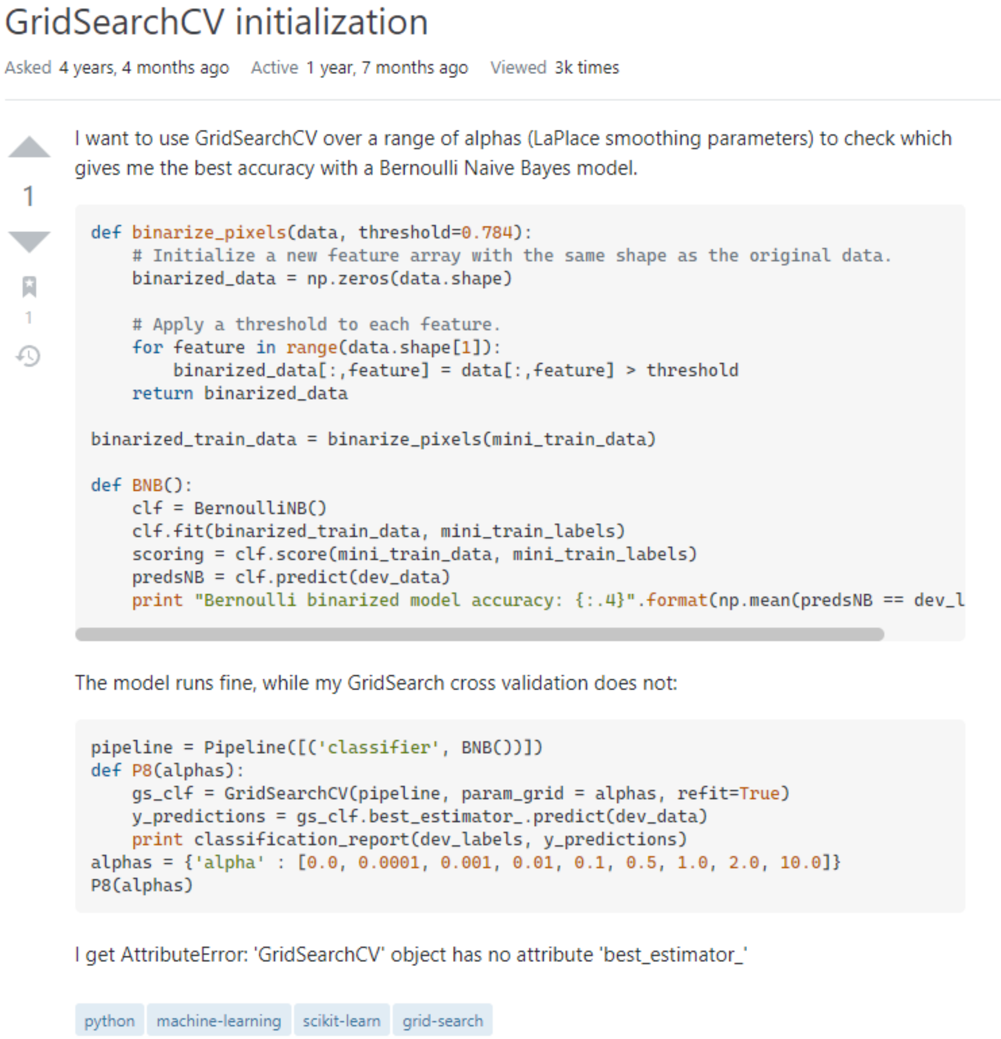
Projet OpenClassrooms – Julien Gremillot

Catégorisation de questions

Le Projet

En utilisant les questions du site Stack Overflow et les tags qui leurs sont associées, le but final de ce projet est de suggérer des tags appropriés à de nouvelles questions.

stackoverflow



GridSearchCV initialization

Asked 4 years, 4 months ago Active 1 year, 7 months ago Viewed 3k times

I want to use GridSearchCV over a range of alphas (LaPlace smoothing parameters) to check which gives me the best accuracy with a Bernoulli Naive Bayes model.

```
def binarize_pixels(data, threshold=0.784):
    # Initialize a new feature array with the same shape as the original data.
    binarized_data = np.zeros(data.shape)

    # Apply a threshold to each feature.
    for feature in range(data.shape[1]):
        binarized_data[:,feature] = data[:,feature] > threshold
    return binarized_data

binarized_train_data = binarize_pixels(mini_train_data)

def BNB():
    clf = BernoulliNB()
    clf.fit(binarized_train_data, mini_train_labels)
    scoring = clf.score(mini_train_data, mini_train_labels)
    predsNB = clf.predict(dev_data)
    print "Bernoulli binarized model accuracy: {:.4}".format(np.mean(predsNB == dev_l
```

The model runs fine, while my GridSearch cross validation does not:

```
pipeline = Pipeline([('classifier', BNB())])
def P8(alphas):
    gs_clf = GridSearchCV(pipeline, param_grid = alphas, refit=True)
    y_predictions = gs_clf.best_estimator_.predict(dev_data)
    print classification_report(dev_labels, y_predictions)
alphas = {'alpha' : [0.0, 0.0001, 0.001, 0.01, 0.1, 0.5, 1.0, 2.0, 10.0]}
P8(alphas)
```

I get AttributeError: 'GridSearchCV' object has no attribute 'best_estimator_'

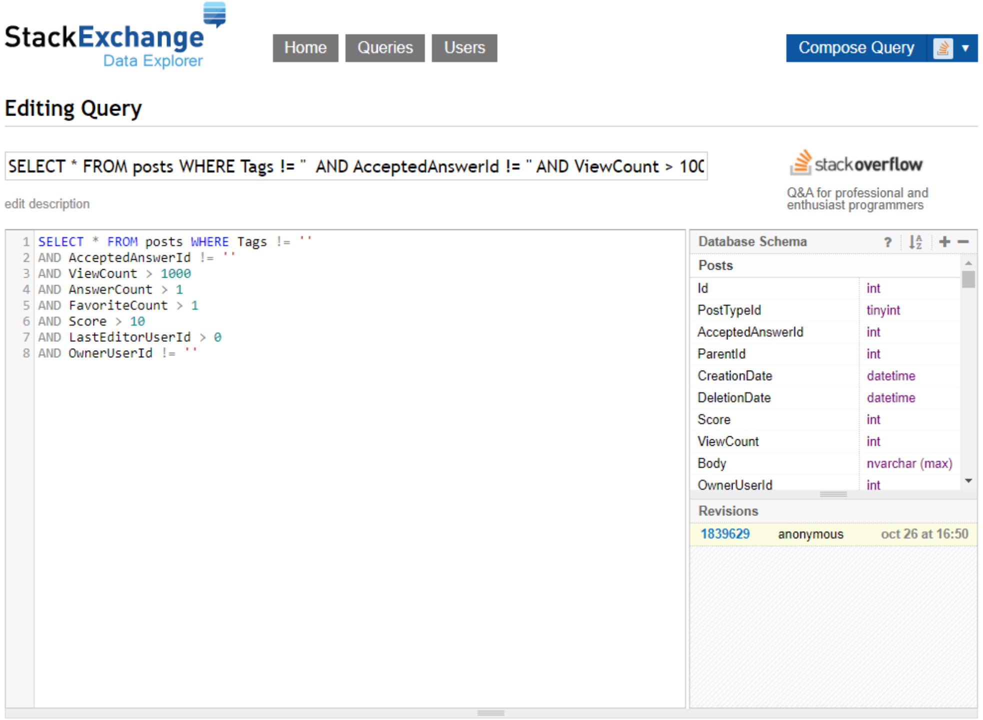
Exploration des données

1. Récupération des questions :

- **Avec tags**
- **Avec réponses**
- **1 réponse acceptée**

- **Mise en favoris**
- **Vues > 1000**
- **Score > 10**

188.065 questions (2008-2014)



Editing Query

SELECT * FROM posts WHERE Tags != '' AND AcceptedAnswerId != '' AND ViewCount > 10

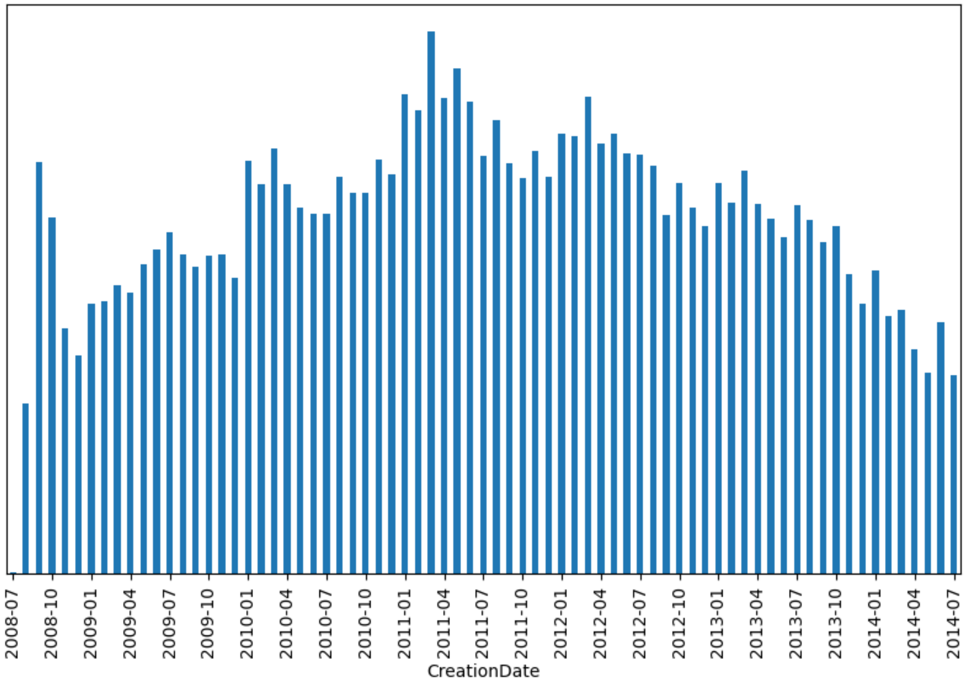
```
SELECT * FROM posts WHERE Tags != ''
AND AcceptedAnswerId != ''
AND ViewCount > 1000
AND AnswerCount > 1
AND FavoriteCount > 1
AND Score > 10
AND LastEditorUserId > 0
AND OwnerUserId != ''
```

Database Schema

Posts	
Id	int
PostTypeId	tinyint
AcceptedAnswerId	int
ParentId	int
CreationDate	datetime
DeletionDate	datetime
Score	int
ViewCount	int
Body	nvarchar (max)
OwnerUserId	int

Revisions

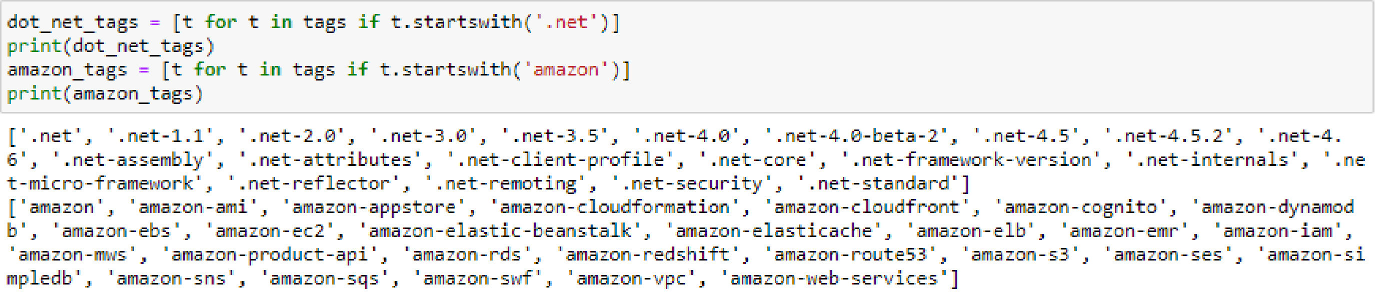
1839629	anonymous	oct 26 at 16:50



Exploration des données

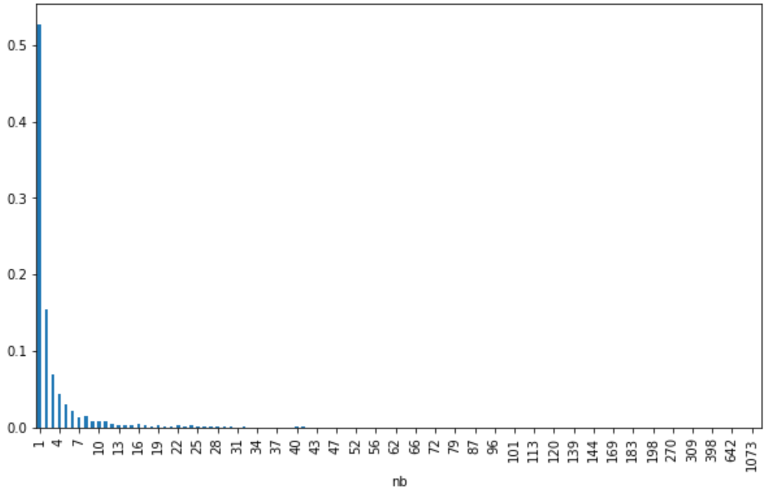
2. Nettoyage des données

- 'Title', 'Body' et 'Tags’
- 4.297 tags
- Nettoyage préfixes (-6%)

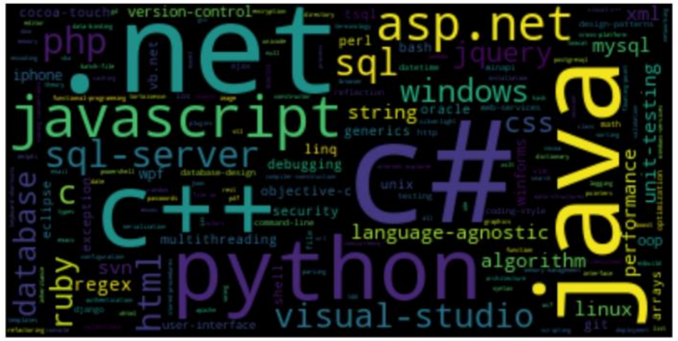


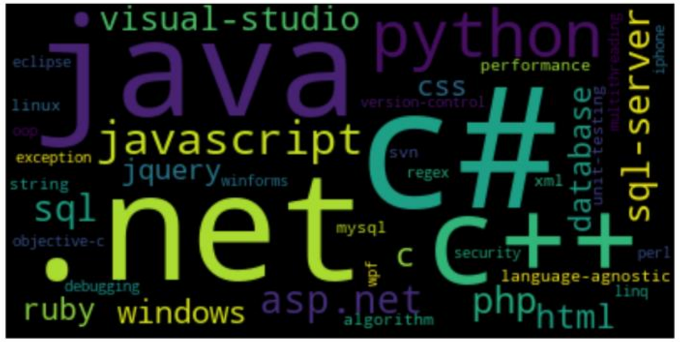
```
dot_net_tags = [t for t in tags if t.startswith('.net')]
print(dot_net_tags)
amazon_tags = [t for t in tags if t.startswith('amazon')]
print(amazon_tags)
```

```
['.net', '.net-1.1', '.net-2.0', '.net-3.0', '.net-3.5', '.net-4.0', '.net-4.0-beta-2', '.net-4.5', '.net-4.5.2', '.net-4.
6', '.net-assembly', '.net-attributes', '.net-client-profile', '.net-core', '.net-framework-version', '.net-internals', '.ne
t-micro-framework', '.net-reflector', '.net-remoting', '.net-security', '.net-standard']
['amazon', 'amazon-ami', 'amazon-appstore', 'amazon-cloudformation', 'amazon-cloudfront', 'amazon-cognito', 'amazon-dynamod
b', 'amazon-ebs', 'amazon-ec2', 'amazon-elastic-beanstalk', 'amazon-elasticache', 'amazon-elb', 'amazon-emr', 'amazon-iam',
'amazon-mws', 'amazon-product-api', 'amazon-rds', 'amazon-redshift', 'amazon-route53', 'amazon-s3', 'amazon-ses', 'amazon-si
mpledb', 'amazon-sns', 'amazon-sqs', 'amazon-swf', 'amazon-vpc', 'amazon-web-services']
```



41 tags présents plus de 100 fois

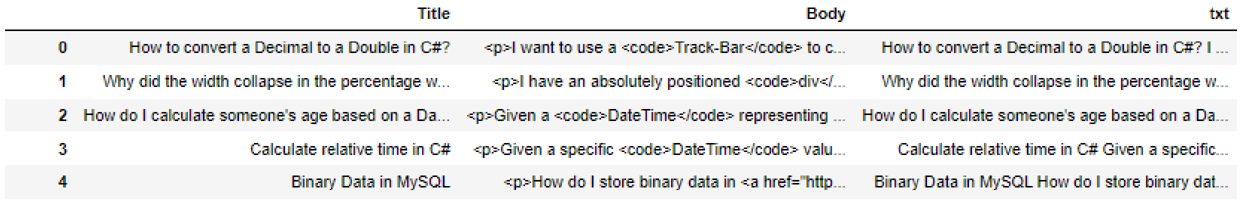




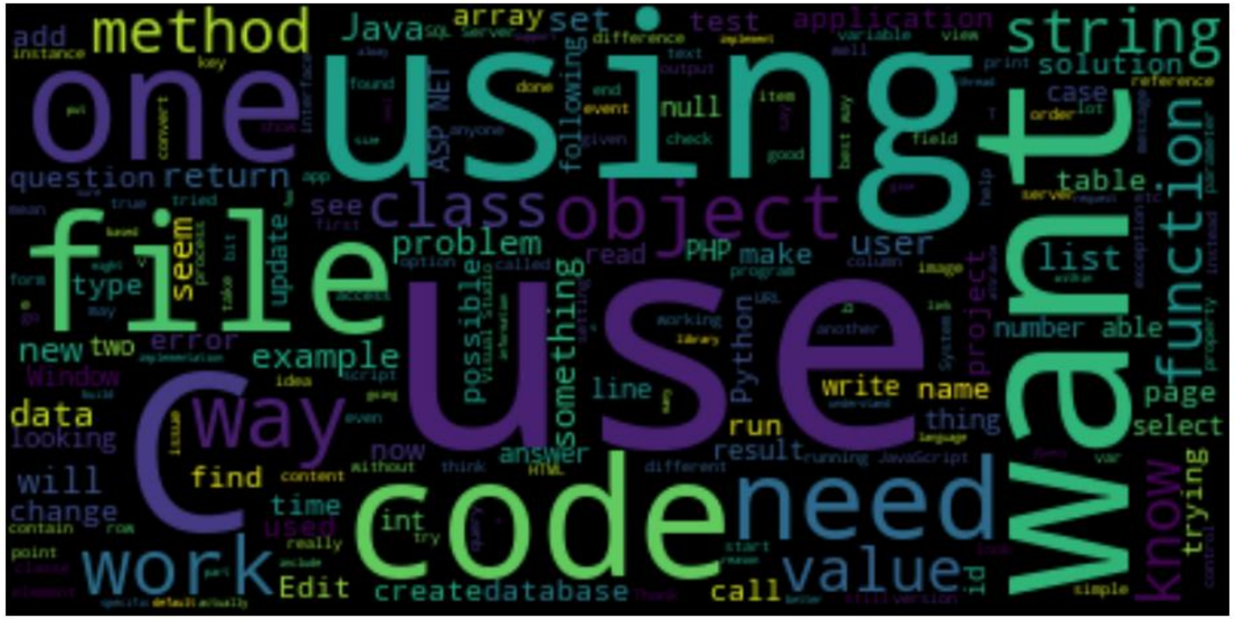
Exploration des données

3. **Suppression du HTML**

4. **Concaténation « Title » & « Body »**

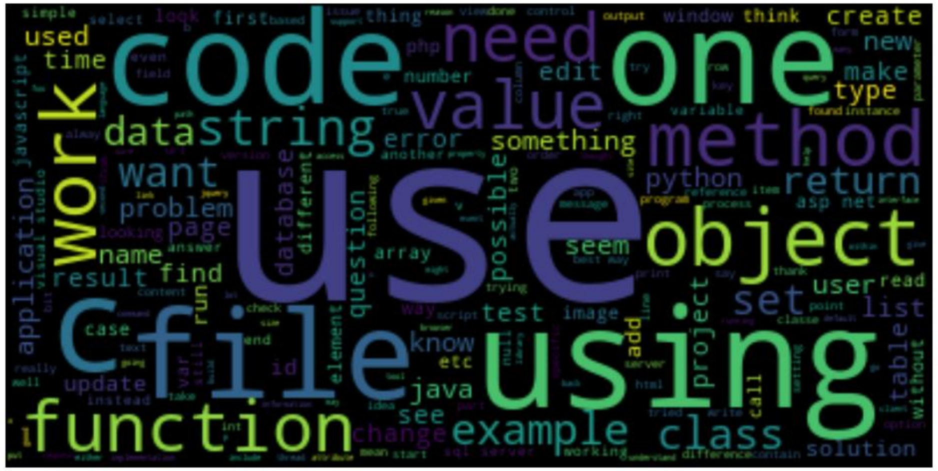


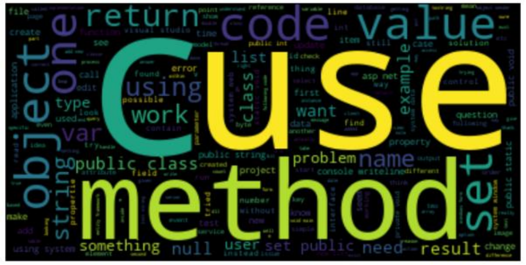
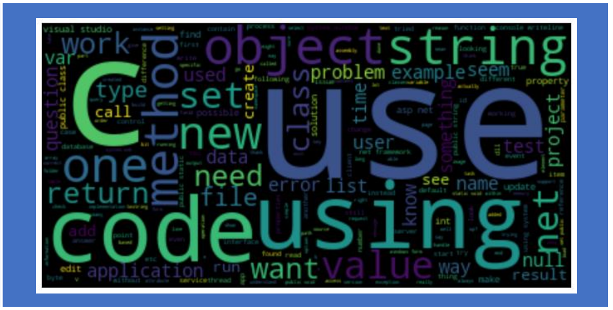
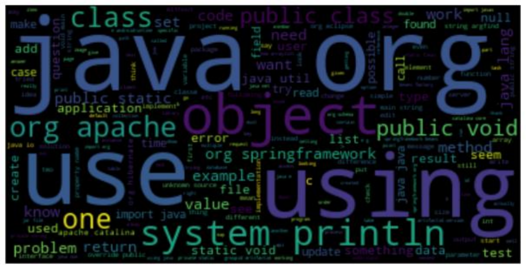
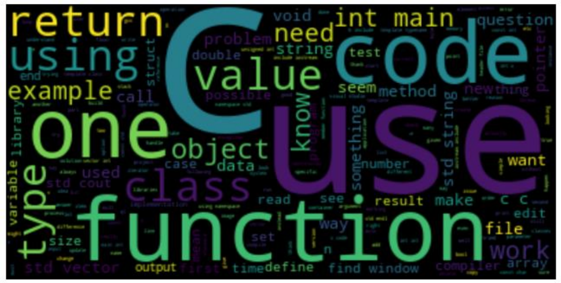
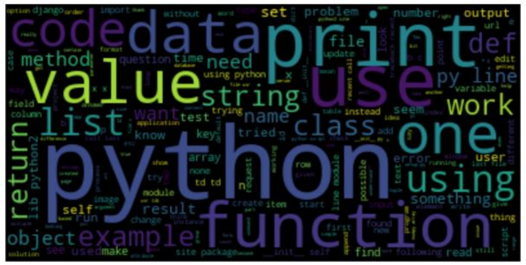
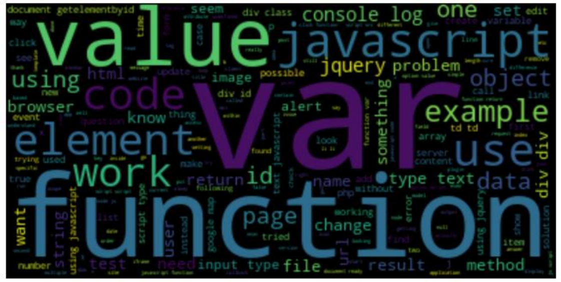
	Title	Body	txt
0	How to convert a Decimal to a Double in C#?	<p>I want to use a <code>Track-Bar</code> to c...	How to convert a Decimal to a Double in C#? I ...
1	Why did the width collapse in the percentage w...	<p>I have an absolutely positioned <code>div</...	Why did the width collapse in the percentage w...
2	How do I calculate someone's age based on a Da...	<p>Given a <code>DateTime</code> representing ...	How do I calculate someone's age based on a Da...
3	Calculate relative time in C#	<p>Given a specific <code>DateTime</code> valu...	Calculate relative time in C# Given a specific...
4	Binary Data in MySQL	<p>How do I store binary data in <a href="http...	Binary Data in MySQL How do I store binary dat...



Exploration des données

5. **Suppression des « stop words »**

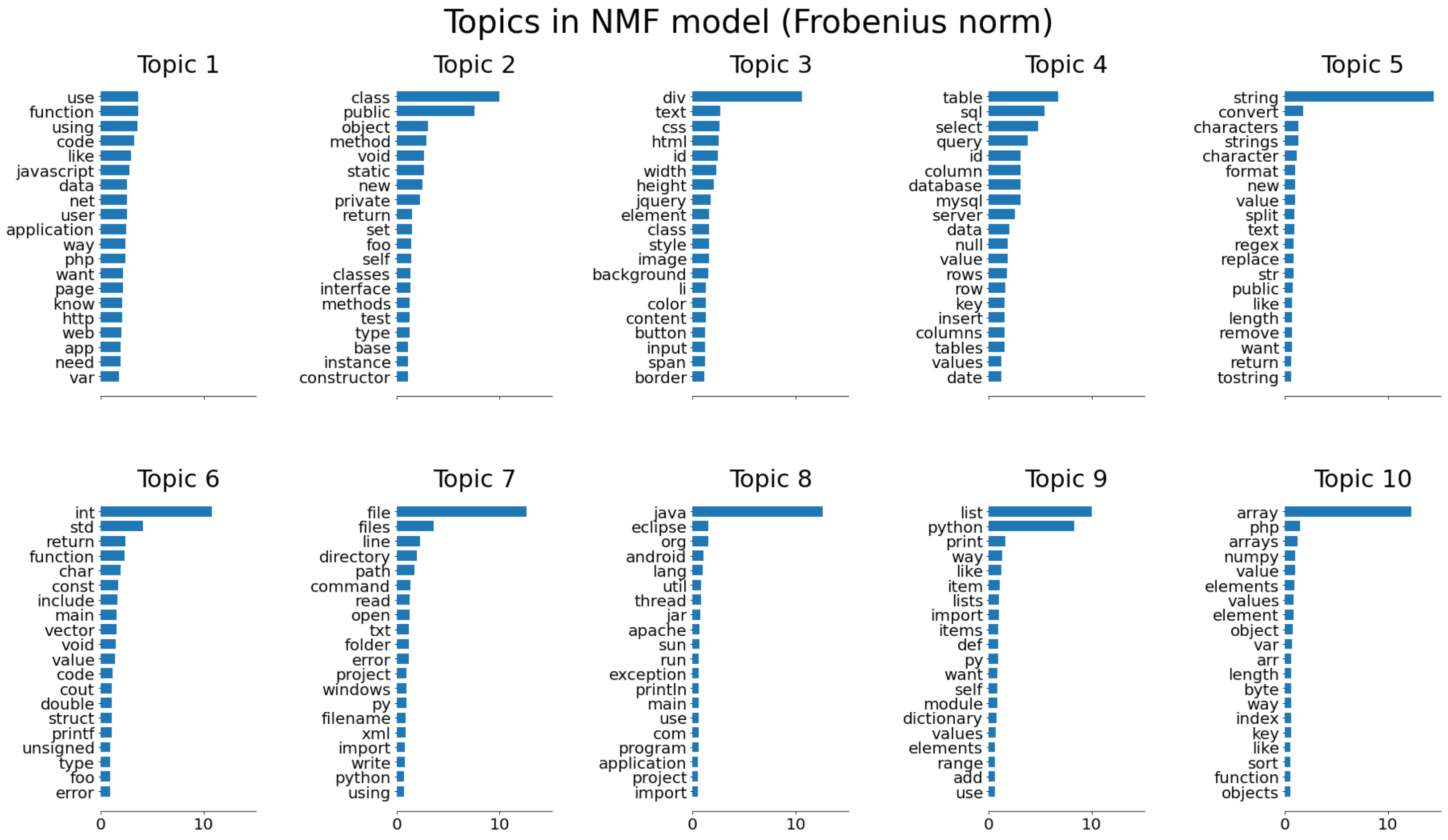


C#		.net	
java		C++	
python		javascript	

Modélisation

Approche non-supervisée

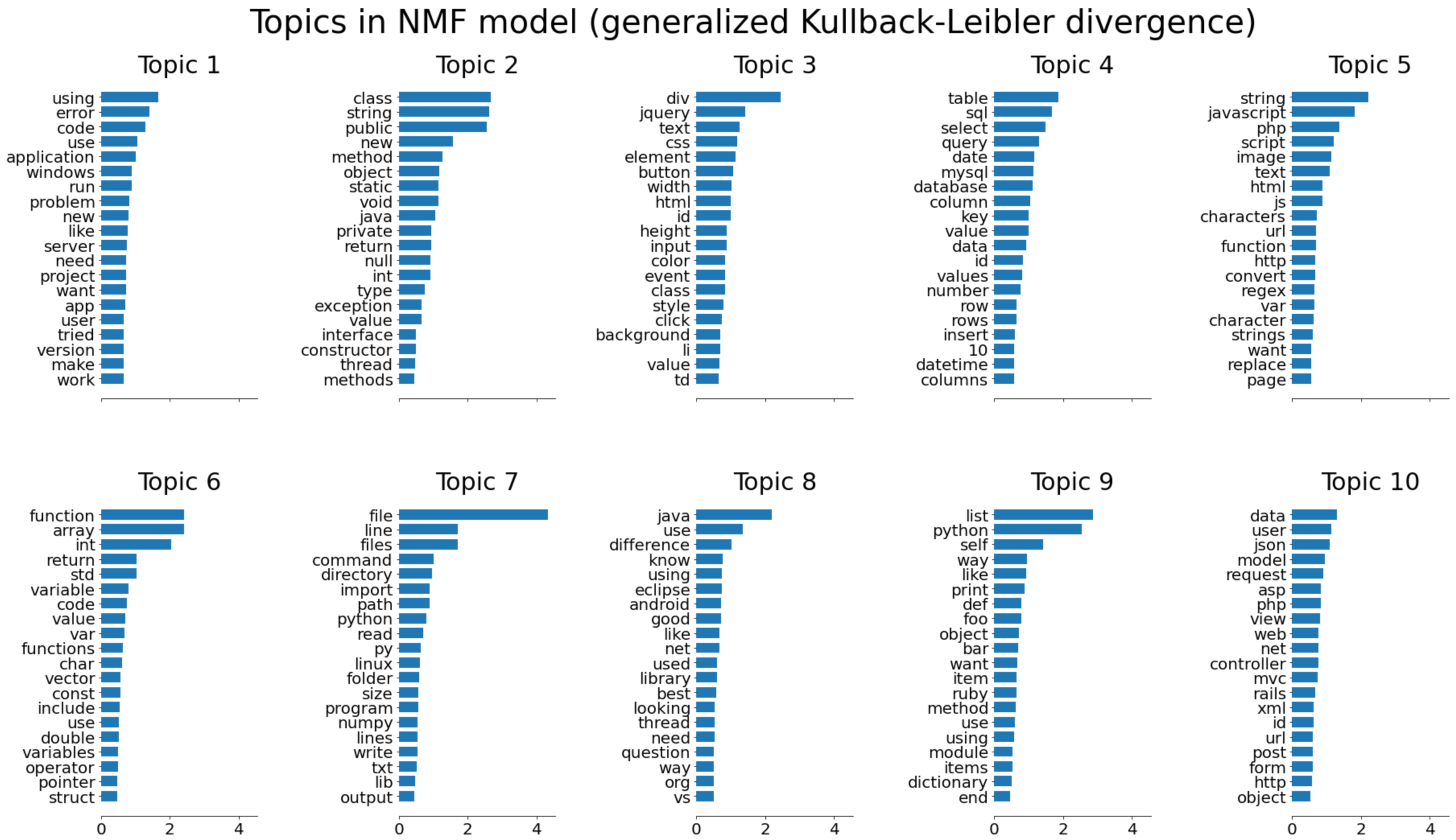
- NMF



Modélisation

Approche non-supervisée

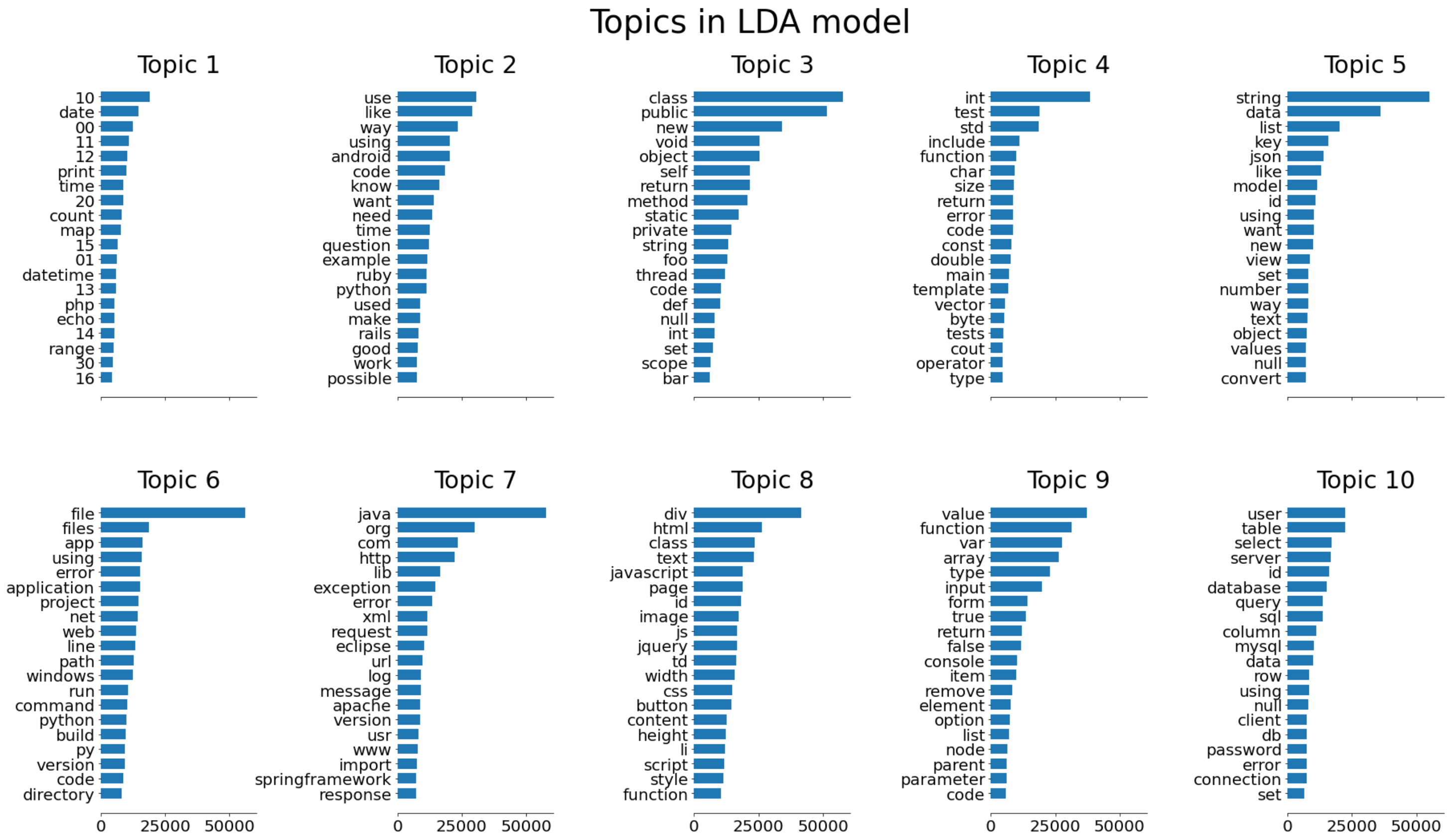
- NMF (KL)



Modélisation

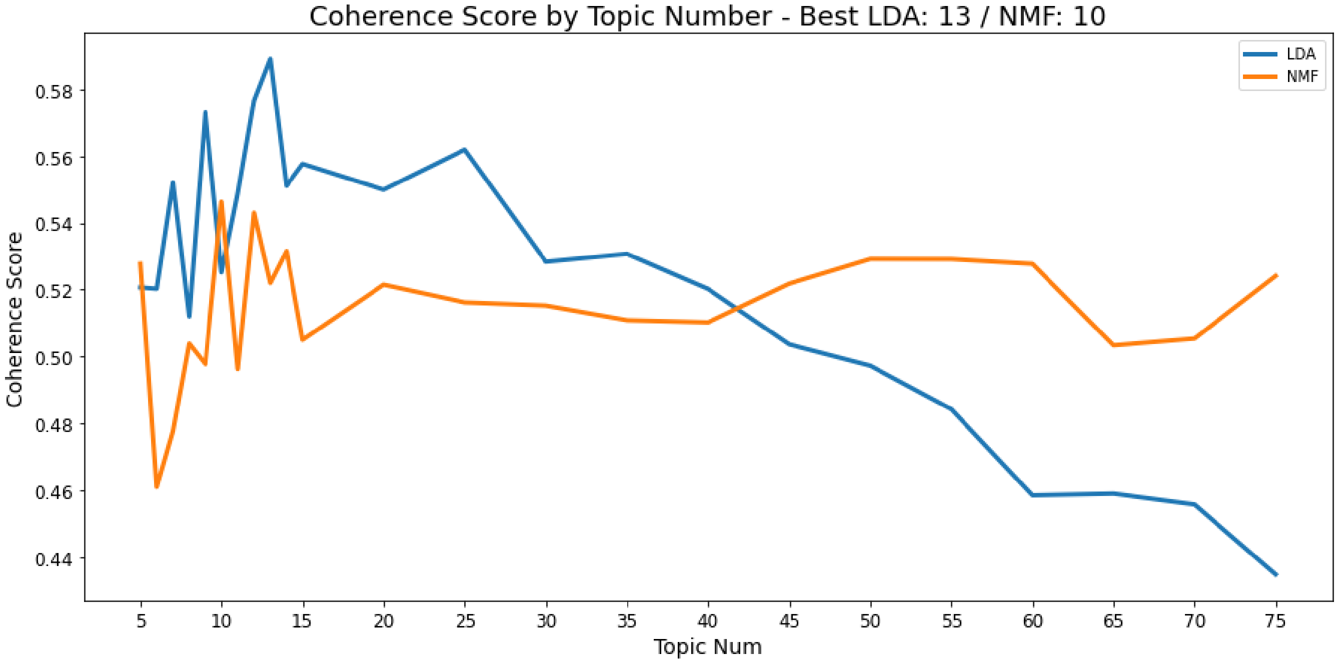
Approche non-supervisée

- LDA



Modélisation

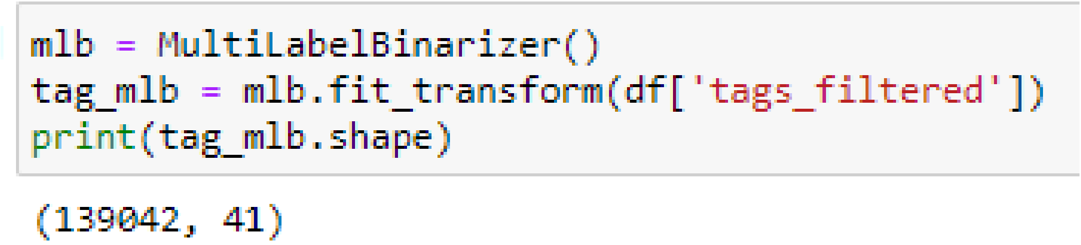
Approche non-supervisée



Modélisation

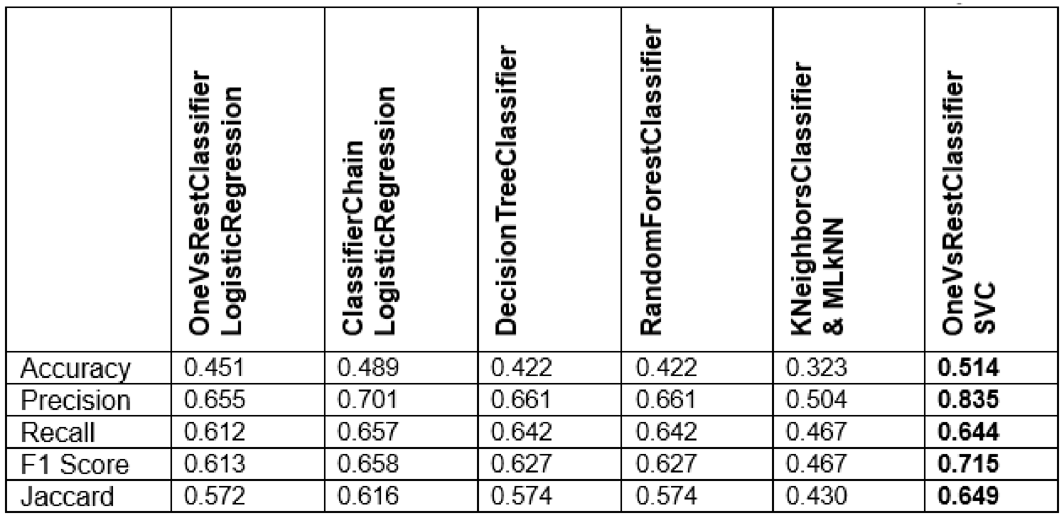
Approche supervisée

- Transformation de la liste de tags en matrice binaire à l'aide d'un **MultiLabelBinazer** de la librairie scikit-learn
- Découpage train / test
- Vectorisation avec le **TfIdfVectorizer**
- **Tests de différents modèles**
- Optimisation avec **GridSearchCV**



```
mlb = MultiLabelBinarizer()
tag_mlb = mlb.fit_transform(df['tags_filtered'])
print(tag_mlb.shape)
```

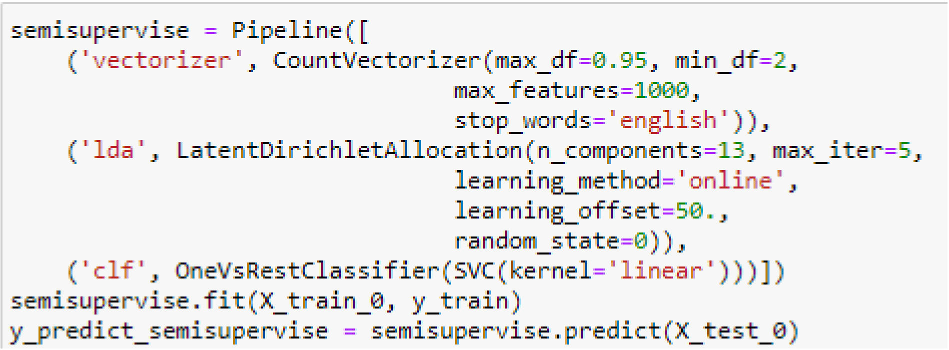
```
(139042, 41)
```



	OneVsRestClassifier LogisticRegression	ClassifierChain LogisticRegression	DecisionTreeClassifier	RandomForestClassifier	KNeighborsClassifier & MLkNN	OneVsRestClassifier SVC
Accuracy	0.451	0.489	0.422	0.422	0.323	**0.514**
Precision	0.655	0.701	0.661	0.661	0.504	**0.835**
Recall	0.612	0.657	0.642	0.642	0.467	**0.644**
F1 Score	0.613	0.658	0.627	0.627	0.467	**0.715**
Jaccard	0.572	0.616	0.574	0.574	0.430	**0.649**

Modélisation – approche semi-supervisée

- CountVectorizer
- LatentDirichletAllocation (13 « topics »)
- OneVsRestClassifier(SVC(kernel='linear'))



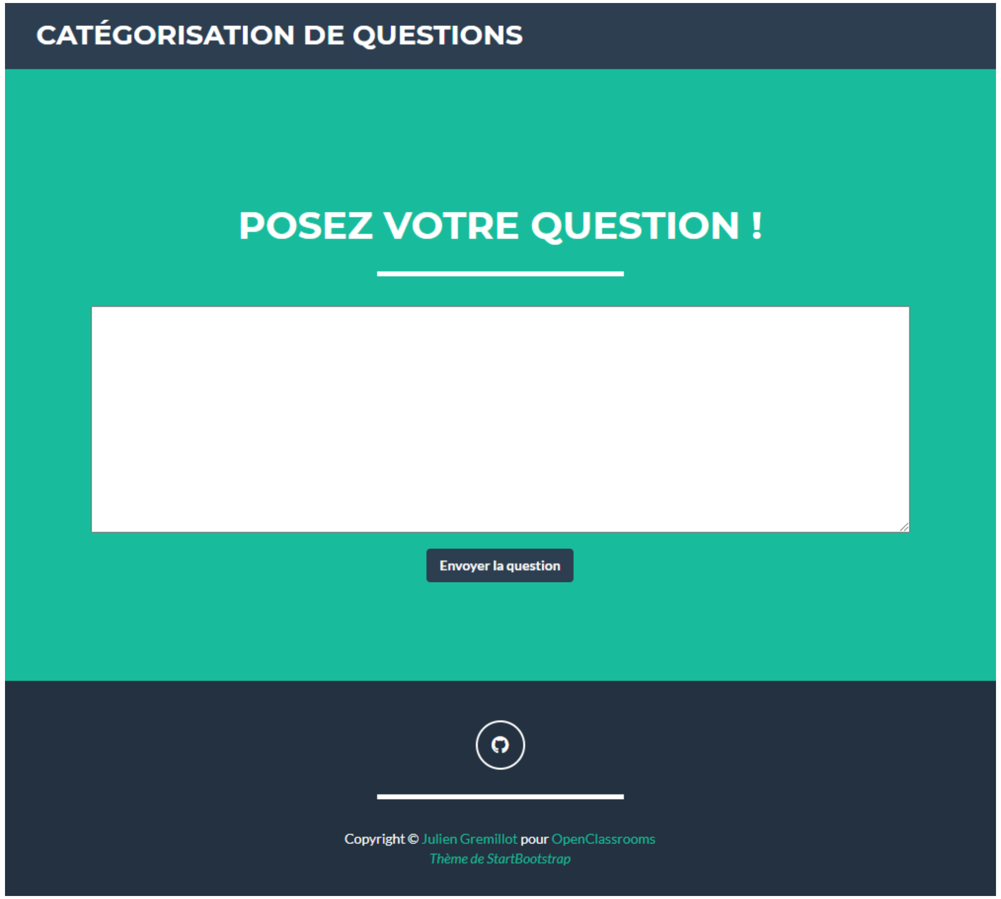
```
semisupervise = Pipeline([
    ('vectorizer', CountVectorizer(max_df=0.95, min_df=2,
                                max_features=1000,
                                stop_words='english')),
    ('lda', LatentDirichletAllocation(n_components=13, max_iter=5,
                                learning_method='online',
                                learning_offset=50.,
                                random_state=0)),
    ('clf', OneVsRestClassifier(SVC(kernel='linear')))])
semisupervise.fit(X_train_0, y_train)
y_predict_semisupervise = semisupervise.predict(X_test_0)
```

Accuracy : 0.042
Precision : 0.115
Recall : 0.049
F1 Score : 0.065
Jaccard : 0.055

Déploiement d’une API

- Développement avec « PyCharm »
- Architecture issue du cours « Concevez un site avec Flask »

https://categorize-questions.herokuapp.com/



CATÉGORISATION DE QUESTIONS

POSEZ VOTRE QUESTION !

Copyright © Julien Gremillot pour OpenClassrooms

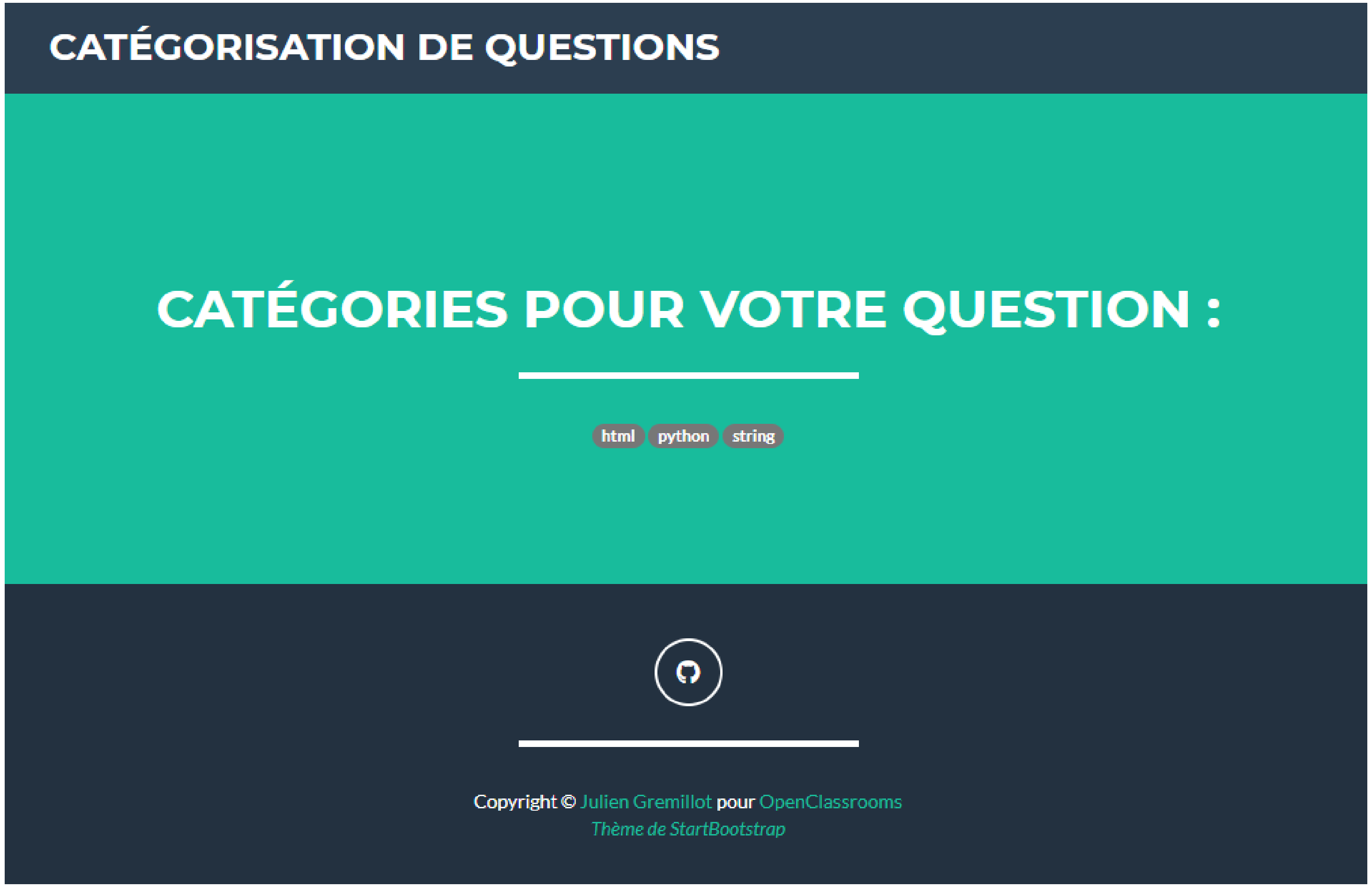
Thème de StartBootstrap

Déploiement d’une API

- Traitement des questions :

- Suppression du HTML (BeautifulSoup)
- Passage en minuscules
- Tokenisation
- Suppression des « stop-words »

- Utilisation du modèle exporté.
- Récupération tags avec MultiLabelBinazer.



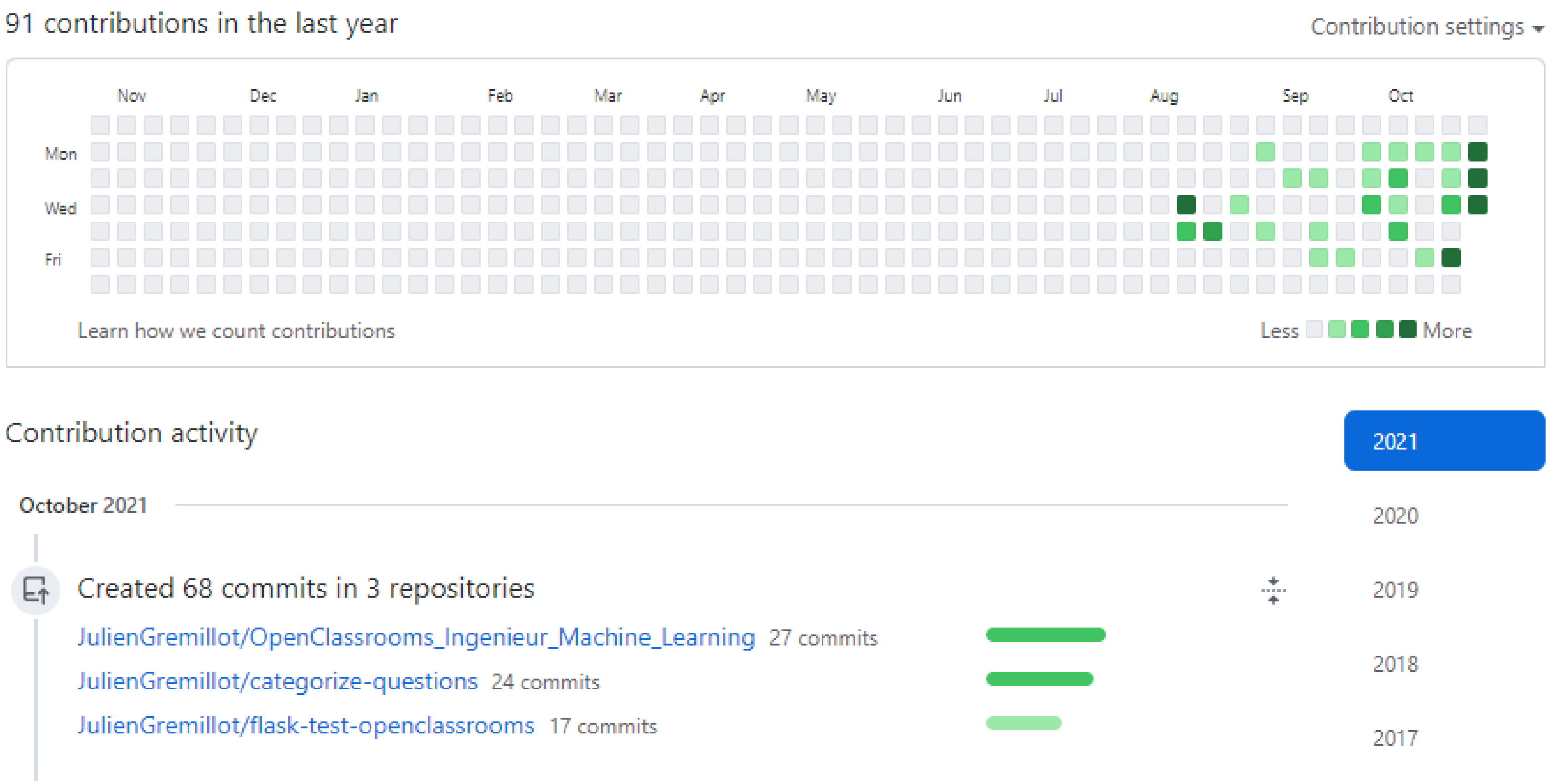
CATÉGORISATION DE QUESTIONS

CATÉGORIES POUR VOTRE QUESTION :

html python string

Copyright © Julien Gremillot pour OpenClassrooms
Thème de StartBootstrap

Gestion des versions



91 contributions in the last year

Learn how we count contributions

Contribution activity

October 2021

Created 68 commits in 3 repositories

JulienGremillot/OpenClassrooms_Ingenieur_Machine_Learning 27 commits

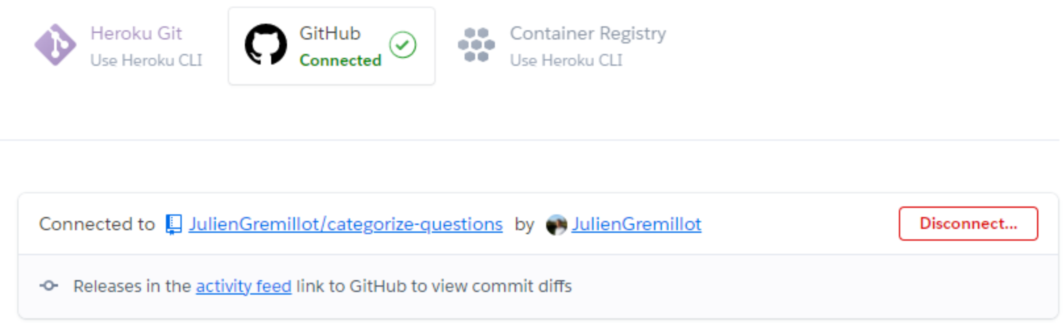
JulienGremillot/categorize-questions 24 commits

JulienGremillot/flask-test-openclassrooms 17 commits

Gestion des versions

Pour déploiement via Heroku :

https://github.com/JulienGremillot/categorize-questions



Connected to JulienGremillot/categorize-questions by JulienGremillot

Releases in the activity feed link to GitHub to view commit diffs