PROJET 6 - Catégoriser automatiquement des questions

FORMENTINI Florian

Sommaire

ÉTUDE DU BESOIN

Tags

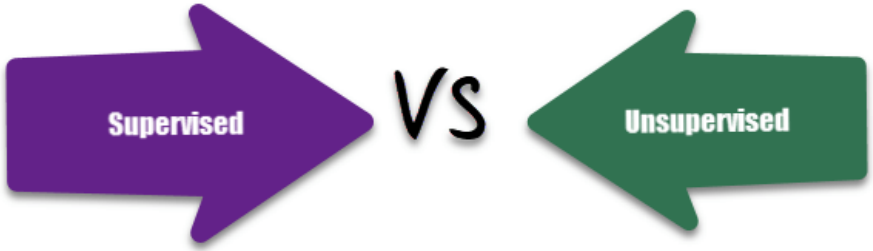
Add up to 5 tags to describe what your question is about

e.g. (angular wordpress reactjs)

Please enter at least one tag; see a list of popular tags.

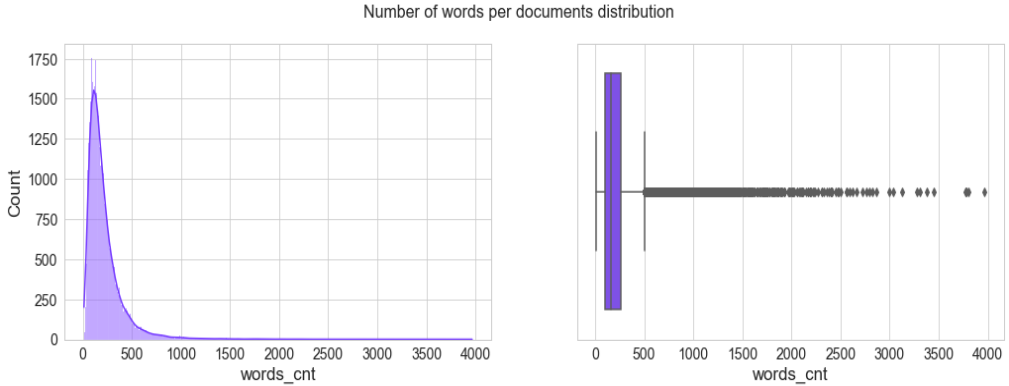
1 – Étude du besoin

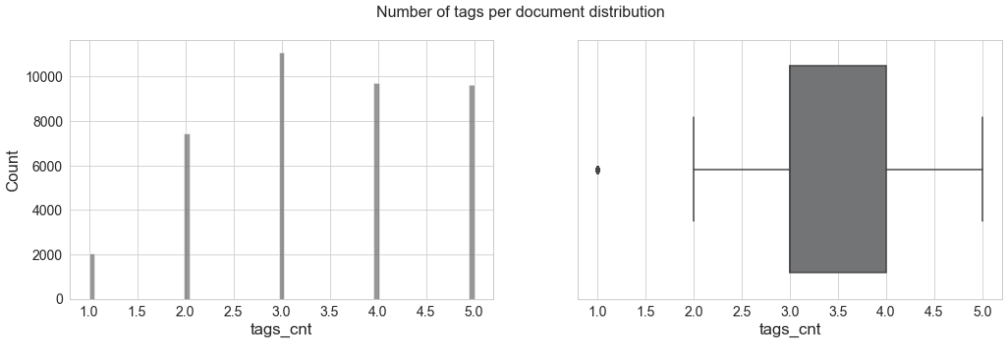
```sql
SELECT
    p.Id as id,
    p.Body as doc,
    p.Title as title,
    p.Tags as tags,
    p.CreationDate as creation_date,
    -- some additionnal metrics
    p.Score as score,
    p.ViewCount as views,
    p.AnswerCount as answers,
    p.CommentCount as comments,
    p.FavoriteCount as favorites,
    p.LastActivityDate as last_activity_date
FROM posts p, PostTypes pt
WHERE p.PostTypeId = pt.Id
AND pt.Name  = 'Question'
AND p.Tags IS NOT NULL
AND p.FavoriteCount > 0
AND p.Score > 0
AND p.CreationDate <= '01-01-2021'  -- to
retrieve recent posts
```

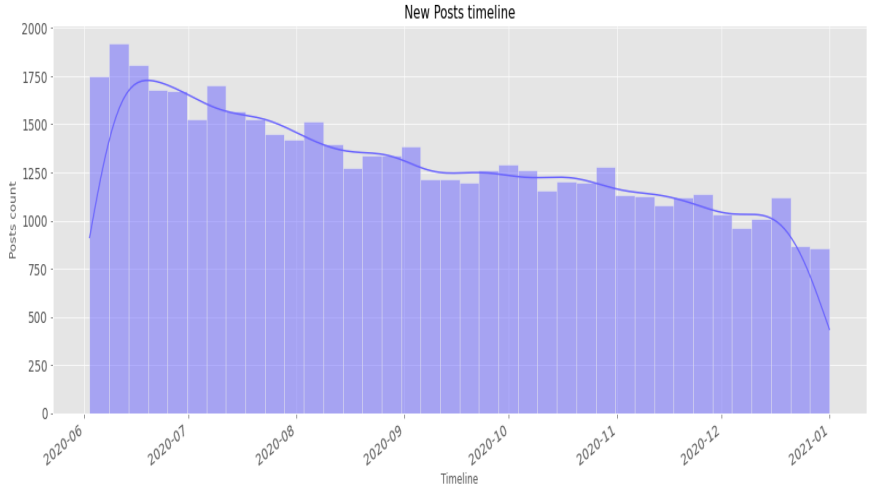


ANALYSE DES DONNÉES

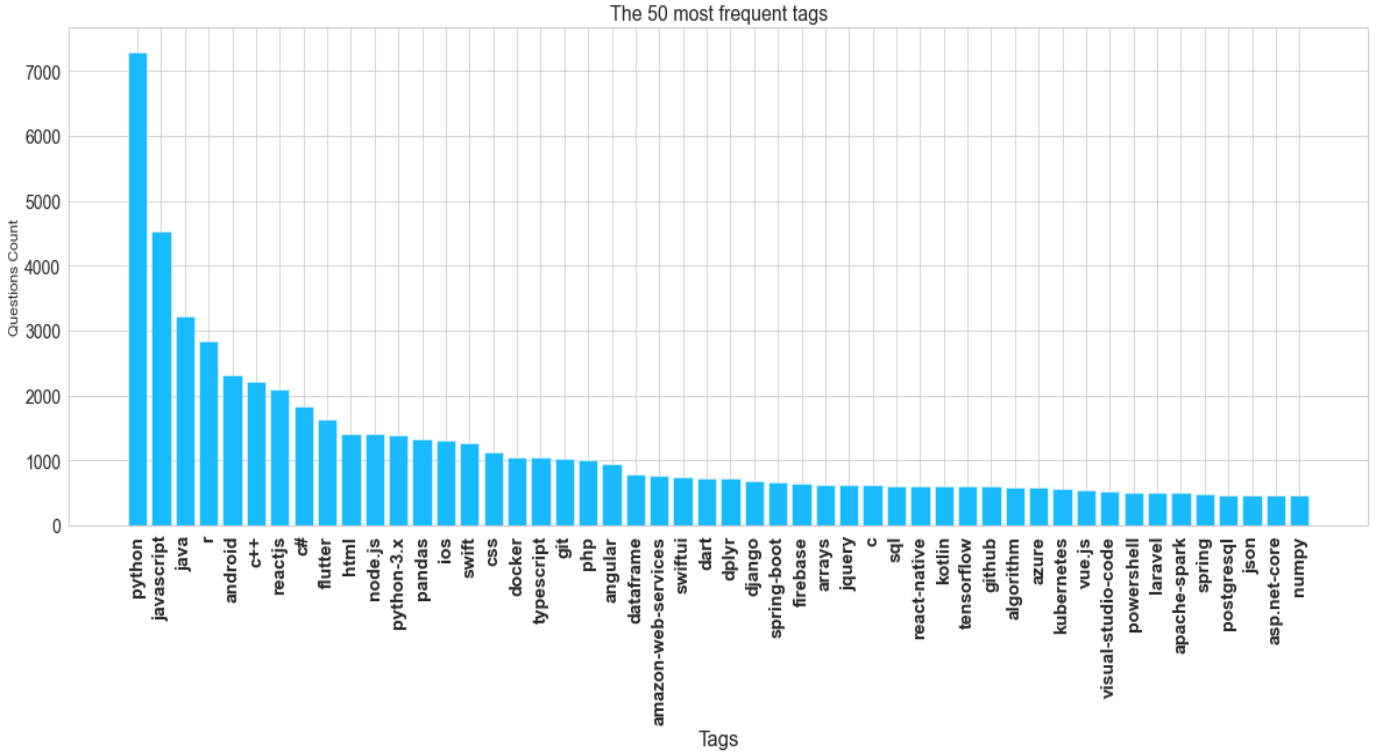
2 - Analyse des données







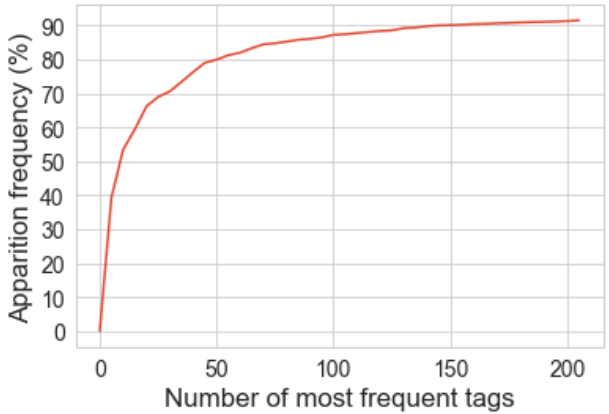
2 - Analyse des données



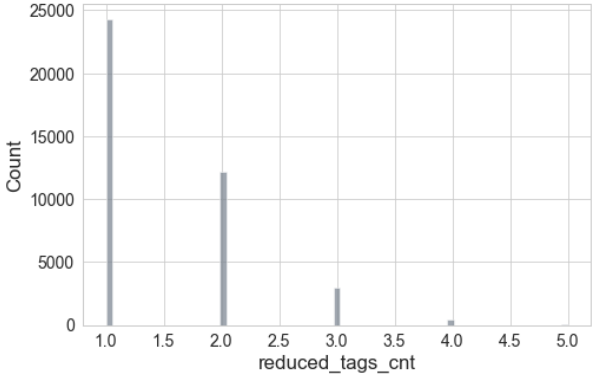
PREPARATION DES DONNÉES

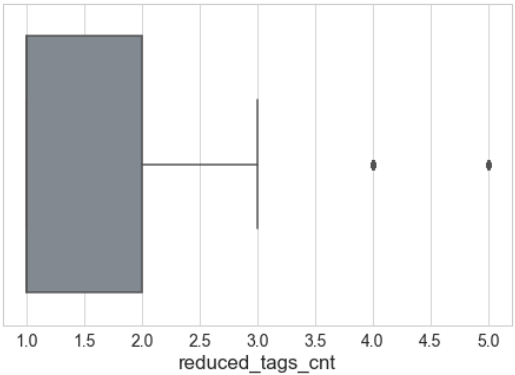
3 – Préparation des données

✓ Reduction des tags



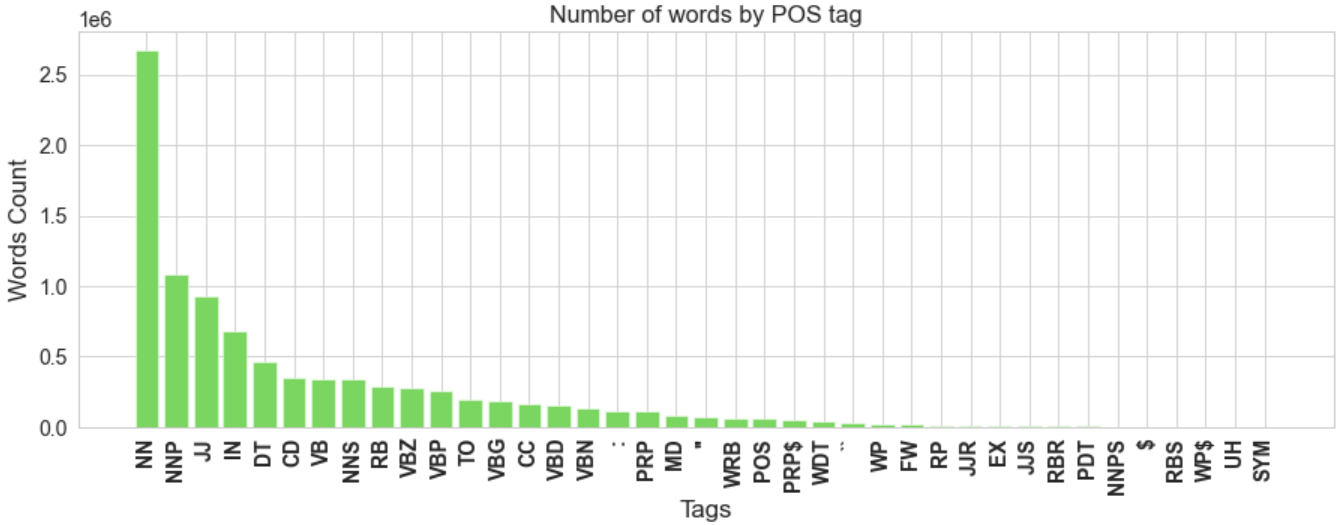
Number of tags per document distribution after reduction





3 – Préparation des données

✓ Tokenisation



- Suppression des Stop Words
- Regex : **(?u)\b\w\w+\b**
- Normalisation des documents : Lemmatisation

3 – Préparation des données

✓ Vectorisation

	the	red	dog	cat	eats	food
1. the red dog →	1	1	1	0	0	0
2. cat eats dog →	0	0	1	1	1	0
3. dog eats food →	0	0	1	0	1	1
4. red cat eats →	0	1	0	1	1	0

$$\text{tfidf}_{ij} = tf_{ij} \times \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

tf_{ij} : fréquence du terme t_i dans le document d_j

$|D|$: nombre total de documents dans le corpus

$|\{d_j : t_i \in d_j\}|$: nombre de documents où le terme t_i apparait.

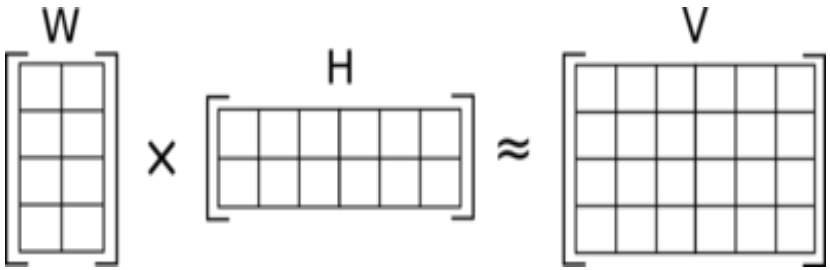
- Utilisation d'uni-grammes
- Fréquence min : 5
- Fréquence max : 75%

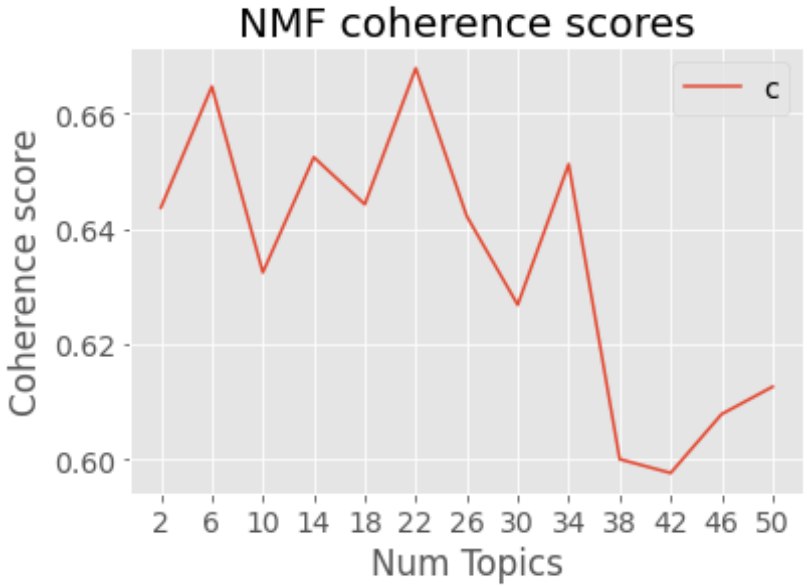
→ **Le dictionnaire obtenu comporte 18104 mots**

APPROCHE NON-SUPERVISÉE

4 – Approche non-supervisée

✓ NMF Topic modeling



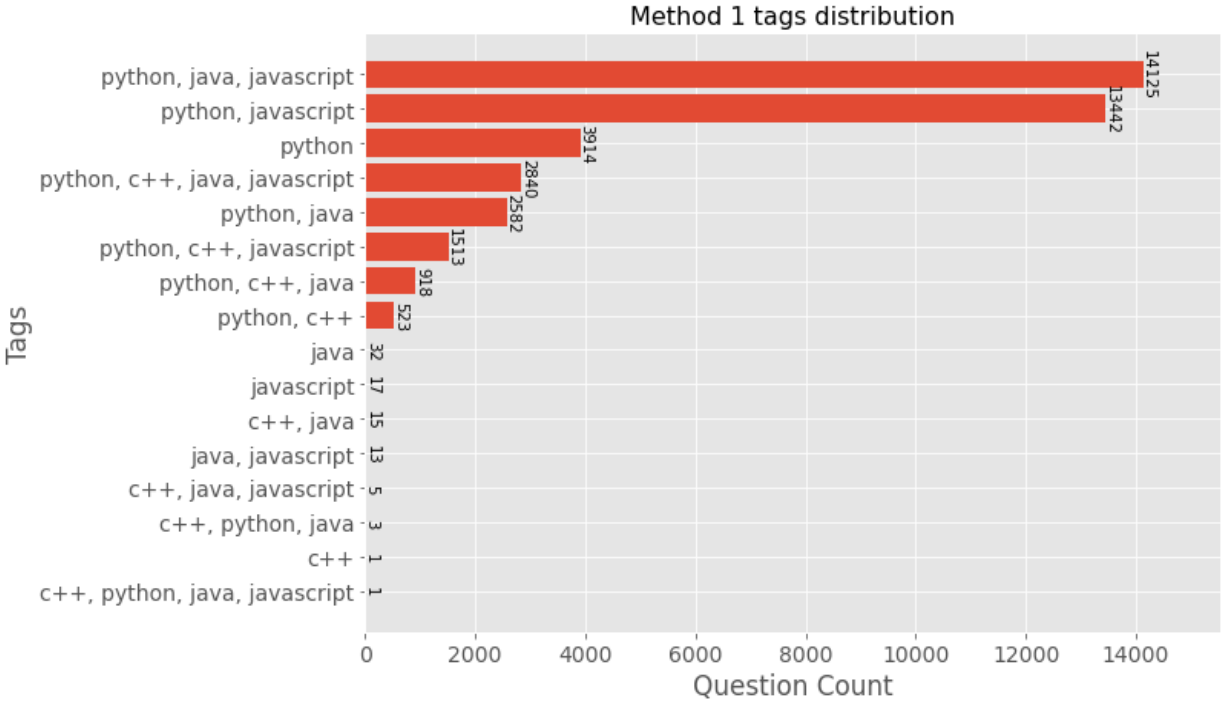


4 – Approche non-supervisée

✓ Nuages de mots

✓ Association tags/thèmes

$$\boldsymbol{C(tags, topics)} = A(documents, tags)^{\mathsf{T}} \times B(documents, topics)$$



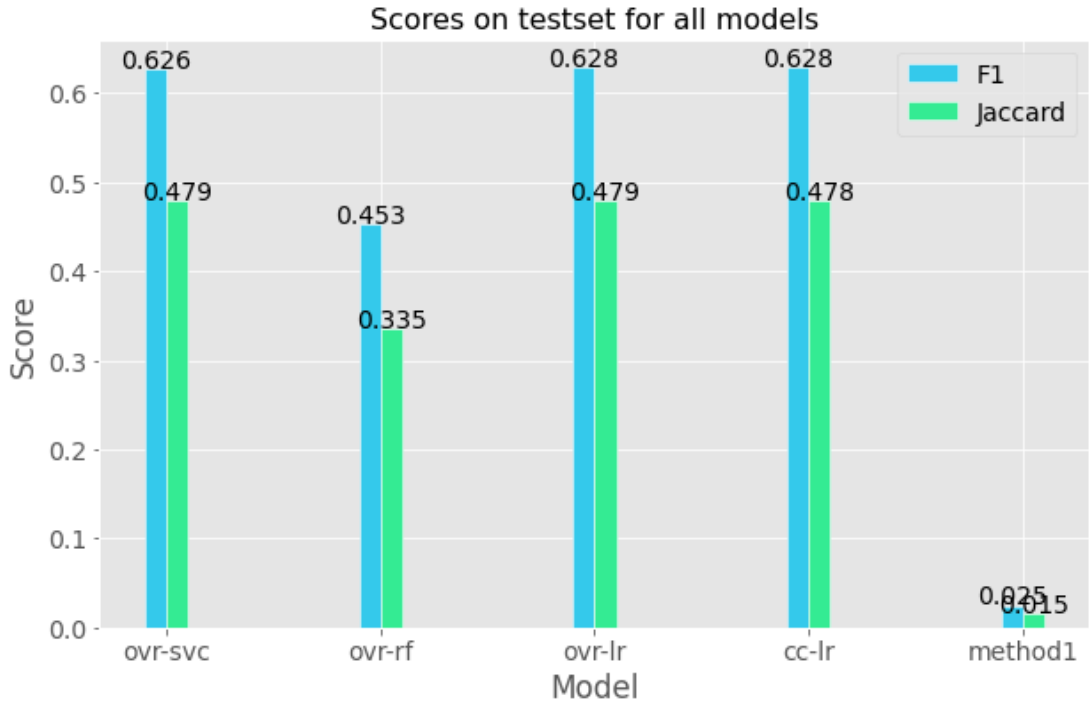
APPROCHE SUPERVISÉE

✓ Classification multi-label

- Méthodes testées : OneVsRest / ClassifierChain
- Classifieurs binaires : LinearSVC, RandomForest, LogisticRegression
- Optimisation des paramètres par GridSearch

COMPARAISON DES MODELES

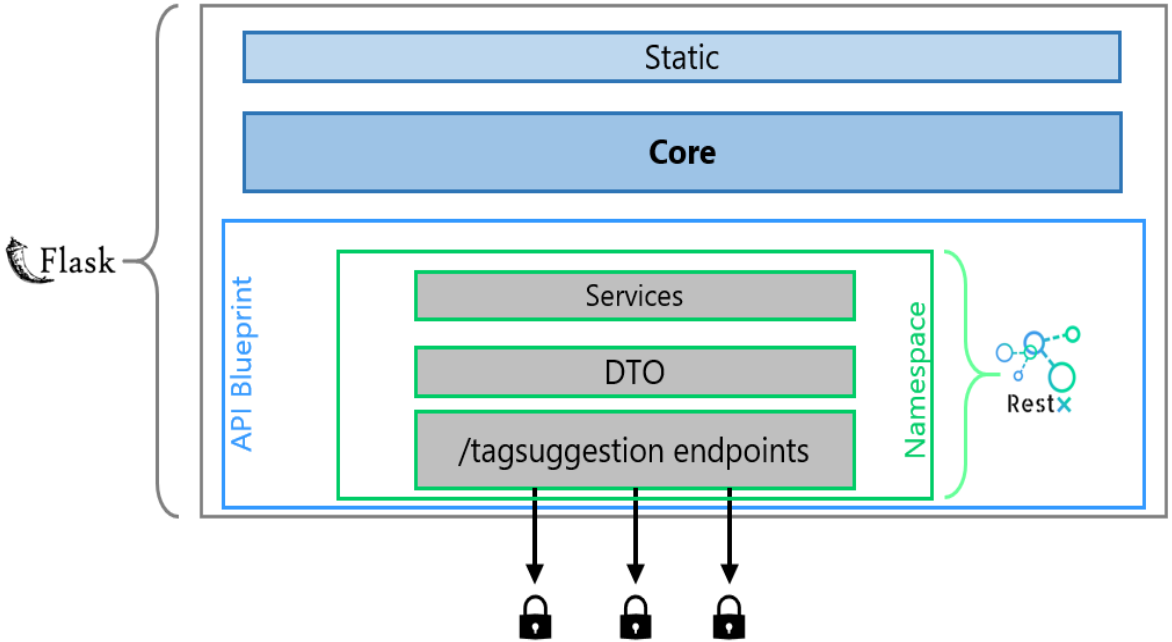
6 – Comparaison des modèles



ovr = OneVsRest
cc = ClassifierChain
svc = LinearSVC
rf = RandomForest
lr = LogisticRegression
method1 = Approche non-supervisée

DÉPLOIEMENT D'UNE API

7 – Déploiement d'une API



HEROKU

GitHub

Démonstration

CONCLUSION

MERCI DE VOTRE ATTENTION

Avez vous des questions ?