

# Rapport

## StackOverflow – Modélisation de prédiction des tags

<b>Auteur</b>	Alain ROUILLON		
<b>Projet</b>	Parcours OC Ingénieur Machine Learning		
<b>Sujet</b>	Catégorisation automatique des questions du site StackOverflow		
<b>Révision</b>	<b>Date</b>	<b>Objet</b>	<b>Description</b>
	02/05/2021	Initiale	

## **Table des matières**

1. Description du projet.....	3
1.1 Contexte.....	3
1.4 Livrables.....	4
2. Préparation des données.....	5
2.1 Cleaning.....	5
2.2. Exploration des tags.....	5
2.3 Préparation du texte.....	6
2.4 Exploration du texte.....	7
2.5 Feature Engineering.....	7
4. Modélisations.....	8
4.1 Dataset.....	8
4.1 Modélisation non supervisée.....	8
4.2 Modélisation supervisée.....	10
5. Fonctionnement retenu.....	11
6. API.....	11
7. Axes d'amélioration.....	12

# 1. Description du projet

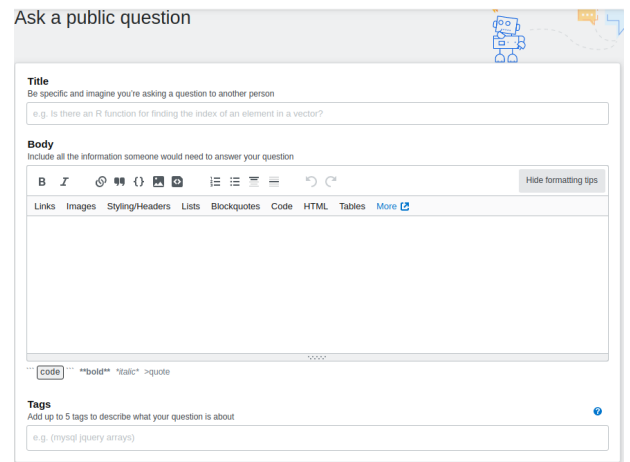
## 1.1 Contexte

Le site StackOverflow (<https://stackoverflow.com/>) offre la possibilité aux informaticiens de poster des questions à la communauté sur des problématiques rencontrées.

L'interface du site permet de saisir :

- un titre
- une description de la problématique rencontrée
- De coller du code pour préciser le problème rencontré
- D'attribuer des tags pour décrire le sujet

La saisie des tags, limitée à 5 choix parmi les tags paramétrés du site, permettra de répertorier la question et de driver le moteur de recherche en vue de recherche par les internautes.



## 1.2 Problématique

Le périmètre des problématiques rencontrées est extrêmement vaste, puisque le métier d'informaticien couvre des missions très diverses, allant du développement à l'ingénierie système, à l'administration de base de données, à la robotique, etc, ...

L'idée est donc de guider l'utilisateur en soumettant une proposition de tags pouvant correspondre à son post à partir des mots clés de celui-ci.

## 1.3 Démarche suivie

Pour ce projet nous avons retenu deux approches possibles :

- Une approche non supervisée
  - on cherche à dégager des topics des posts existants
  - la finalité est alors d'associer à ces topics les tags des posts pour lesquels ils sont prédominants
- Une approche supervisée
  - à partir des posts existants déjà taggués on cherche à modéliser le lien qui existe entre eux pour effectuer par la suite une probabilité de choix possibles
  - il s'agit d'une modélisation de classification multi-label

## **1.4 Livrables**

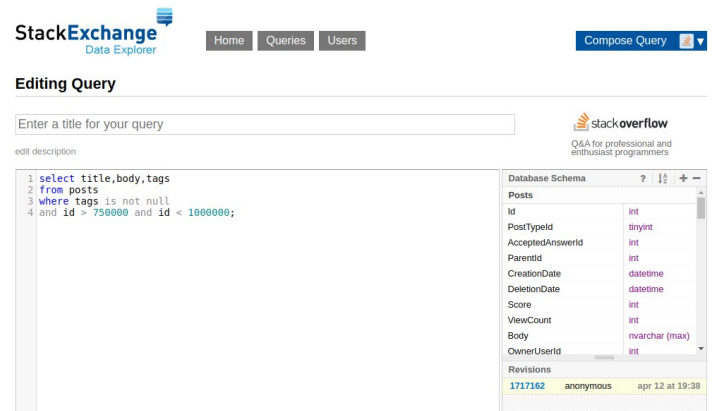
L'objectif est de communiquer à l'utilisateur la suggestion des tags possibles au moment de la saisie par le portail du site. Pour cela on met en place une API qui peut être intégrée en appel REST par le site.

Pour ce projet l'API est interfacée pour permettre l'interaction avec les fonctions de prédictions soutenues par le(s) modèle(s) mis en place.

## 2. Préparation des données

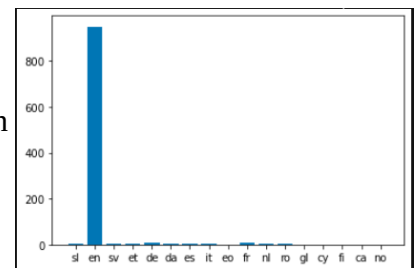
Les données sont récupérées depuis le site <https://data.stackexchange.com/stackoverflow>

Les données sont à récupérer par paquets de 50 000 enregistrements, Pour obtenir un dataset consistant l'idée retenue est de ne récupérer que les posts qui sont taggués.



### 2.1 Cleaning

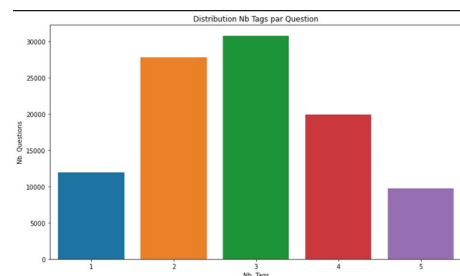
- Après analyse de la distribution des langues représentées on ne conserve que les posts rédigés en anglais
- Suppression des enregistrements dont le body du post est NaN
- Suppression des enregistrements dont les tags sont NaN



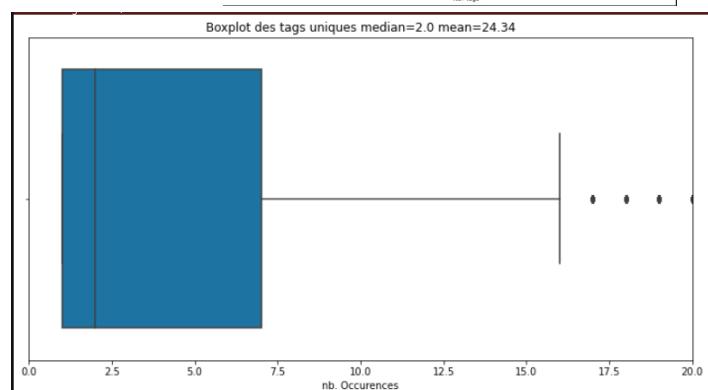
### 2.2. Exploration des tags

Sur le dataset étudié on recense 11812 tags distincts.

La distribution du nombre de tags par question présente un maximum de 3.



50 % des tags ne sont utilisés que 2 fois.



## StackOverflow - Catégorisation automatique de questions

La distribution de la fréquence d'utilisation des tags montre que peu de tags sont souvent représentés, il sera donc très difficile d'entraîner un modèle performant pour les prédire.

Ce point est important à souligner car il va conditionner notre démarche de modélisation supervisée.



### 2.3 Préparation du texte

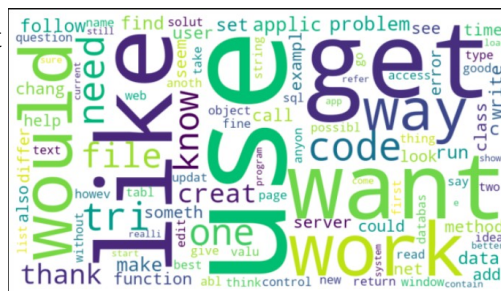
Afin d'effectuer notre modélisation nous effectuons un pré-processing des données textuelles portées par le corps des questions.



## 2.4 Exploration du texte

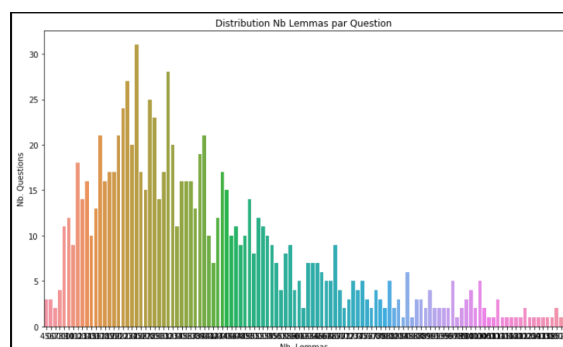
Après pré-traitement (voir point précédent) on fait un petit zoom sur le contenu des termes qui vont constituer le corpus.

On constate visuellement que les termes servant à introduire les questions restent très présents, ce qui ne sert pas directement notre objectif. Cela pourra constituer une piste d'amélioration (voir point 7)



La distribution de termes racine par question est presque normale avec un léger décalage à gauche.

Globalement les bodies des posts qui vont servir au training sont bien distribués en termes d'apport de mots racine.



## 2.5 Feature Engineering

En vue de la modélisation on procède préalablement à une étape de feature engineering portant sur la colonne du corps du post (nommé document dans la suite). L'idée est de vectoriser le document pour apporter de l'information à l'algorithme d'apprentissage :

- indicateur de fréquence de chaque terme dans le document
- notion de séquençement du terme (ordre)

Plusieurs techniques ont été testées :

- Bag of Words
- Count Vectorizer
- TF-IDF

TF-IDF est la méthode qui permet de mettre en place le plus d'informations, en donnant un indicateur calculé comme étant le produit de la fréquence du terme dans un document (TF) et de la fréquence de document dans lequel le terme apparaît (IDF) :

$$\text{TF-IDF} = \text{TF}(t,d) * \text{IDF}(t)$$

Cela permet de coupler à un terme donné un facteur de fréquence à une dimension tenant compte de 2 dimensions.

## 4. Modélisations

### 4.1 Dataset

Par la suite, pour des problématiques de ressources machine, le dataset utilisé en entrée de training est limité à 9 342 enregistrements.

### 4.1 Modélisation non supervisée

L'objectif est de construire un moyen de matérialiser le dataset dont le contenu est inconnu a priori.

L'algorithme Latent Dirichlet Allocation (LDA) permet de construire des topics représentant les thèmes abordés dans les documents en entrée.

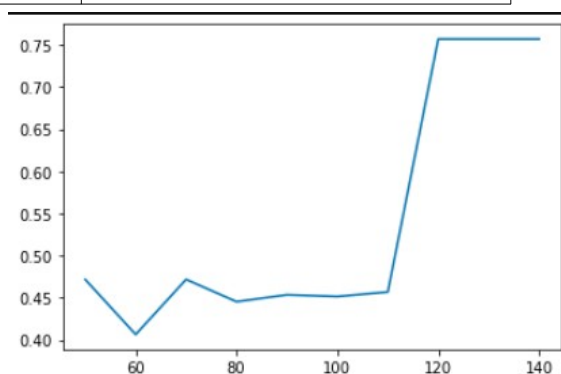
L'algorithme repose sur l'identification d'un nombre fini de topics donné en entrée, et cherche pour chaque terme de chaque document à maximiser la probabilité que le document soit rattaché au thème  $t$  construit à partir de l'ajout du nouveau terme, et à maximiser la probabilité que le thème  $t$  soit assigné au terme.

Nous avons choisi l'implémentation LDAMulticore du package gensim.

L'entraînement du modèle est passé par une phase de détermination des meilleurs paramètres par itération sur des plages de valeurs candidates pour les paramètres principaux :

Paramètre	Description	Valeur optimale
Topics	Nombre de thèmes à construire	50
alpha	Contrainte de l'algorithme à considérer qu'un document est composé de plus ou moins de topics. Plus alpha est bas, plus l'algorithme tendra à associer peu de topics dominants.	0,91
beta	Contrainte de l'algorithme à considérer qu'un topic est composé de plus ou moins de termes. Plus beta est bas, plus l'algorithme va considérer que les topics sont composés d'un nombre de termes limité.	symmetric

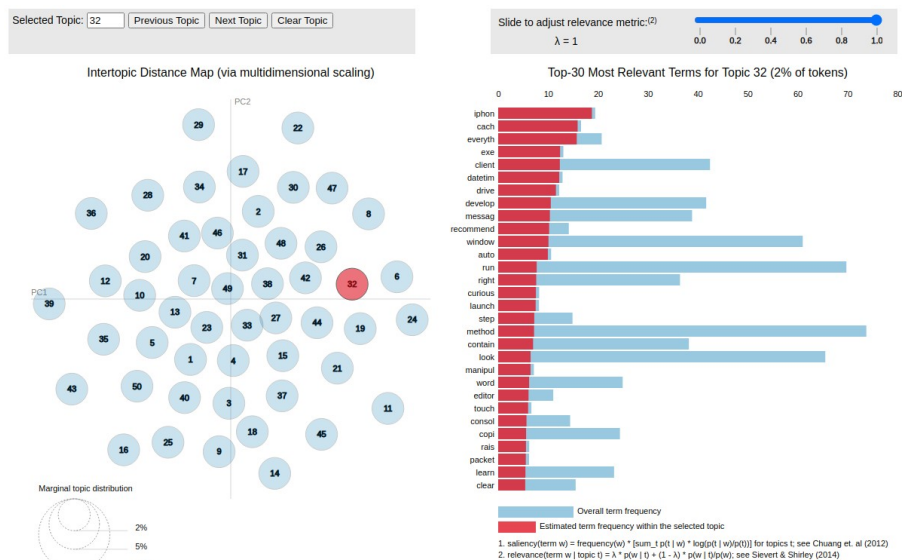
	Validation_Set	Topics	Alpha	Beta	Coherence
319	100% Corpus	50	0.9099999999999999	symmetric	0.275046





## StackOverflow - Catégorisation automatique de questions

La visualisation des topics par pyLDAvis nous permet de constater assez rapidement qu'en l'état notre modèle résultant n'est pas réellement exploitable. Chaque topic a le même poids (environ 2%)

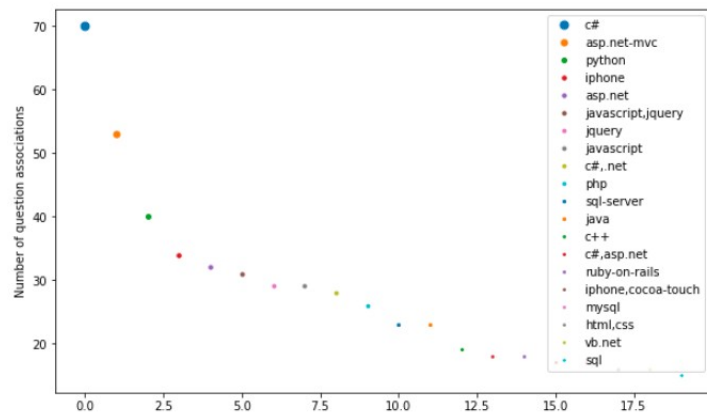


Cela est confirmé par le calcul. La prédiction pour un nouveau document est de 2 % pour chaque topic. Par conséquent il ne serait pas possible en tout état de cause de dégager un topic dominant et d'en ressortir une suggestion de tags.

Ce constat s'explique probablement du fait du nombre limité de données du dataset.

## 4.2 Modélisation supervisée

Sur le périmètre limité aux 9 342 enregistrements la distribution de tags est la suivante :



Nous partons sur une modélisation pour les 20 tags les plus fréquents.

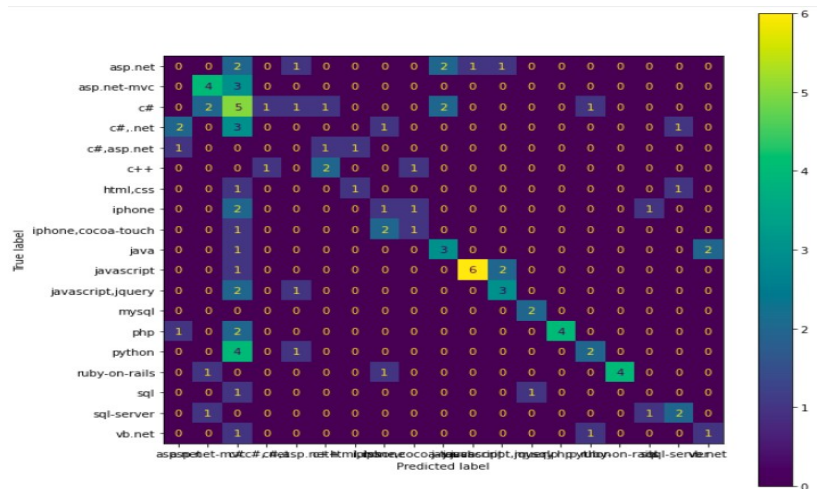
Le modèle choisi est XGBoostClassifier.

Le réglage des hyperparamètres est fait par pipeline à partir de GridSearchCV.

L'estimateur des performances du modèle est l'accuracy.

L'accuracy obtenue sur le jeu de test est de 0,39.

La matrice de confusion montre les cas d'erreurs rencontrés.



Ce résultat n'est pas très bon, il ne permet pas d'envisager une mise en production du modèle en l'état.

L'amélioration des résultats est possible en intégrant un nombre de cas plus important pour couvrir plus de combinaisons. Ce point nécessite un plus gros dataset, avec donc plus de ressources machine pour processor.


## 5. Fonctionnement retenu

En l'état le modèle non supervisé n'est pas exploitable.

Nous aurions pu coupler les deux modèles en privilégiant le modèle supervisé, et en cas de défaut de réponse donner le relai au modèle non supervisé.

A défaut nous fonctionnons uniquement avec le modèle supervisé.

## 6. API

Etape	Ecran
Splash screen	 <p>The splash screen displays the StackOverflow logo, which consists of a stylized 'S' made of horizontal bars of increasing height, followed by the text 'stackoverflow'. Below the logo, it says 'Welcome to StackOverflow tagging API !' and 'Please have a look !'.</p>
Invite de saisie de la question et bouton de soumission	 <p>The screen shows a text input area with the placeholder text 'Question Stackoverflow:'. The input contains the following text: 'Hello, I am developing a new application and I need to create a new table in sql-server database. My application is an HTML form and I submit it through a javascript function. How can I do that ? Thanks !'. Below the input is a blue button labeled 'SEND'.</p>
Résultat de la prédiction	 <p>The screen shows the same text input area as the previous screen, with the same text. Below the input, there is a section titled 'Tags suggérés' (Suggested tags) with a list of tags: 'c#', 'sql', 'sql-server', and 'vb.net'. At the bottom of the screen is a blue circular button with a white left-pointing arrow.</p>

## **7. Axes d'amélioration**

Général :

- Enrichir les stopwords pour éliminer les termes inutiles pour l'objectif : mots question, termes d'interaction de dialogue, etc, ...
- Augmenter taille du dataset
- GPU/RAM

Non supervisé

- Dataset plus large : en augmentant le nombre de données les calculs sont plus longs, mais surtout les combinatoires plus importantes, et donc le calcul de probabilité a de meilleures chances de donner des résultats pertinents.

Supervisé

- CNN : un réseau de neurone pourrait améliorer l'accuracy en mettant en place plusieurs couches cachées qui serviraient à affiner la classification, et donc réduire les cas de faux positifs.