|  |
| --- |
| Parcours Machine Learning Engineer OpenClassrooms |
| Catégorisez automatiquement des questions |
| L'objectif de ce projet est de développer un système de suggestion de tags pour une question posée sur le site Stack Overflow. Le but est d'aider les membres du site à avoir un système de recommandations de tags pour leur questions (5 au maximum) afin de leur faire gagner du temps et leur faciliter l’accessibilité de leur question dans la bonne catégorie et donc une réponse pertinente.  Dans un premier temps, je vais récupérer les données à partir d'une API du site Stack Overflow, puis je vais les analyser et les traiter en utilisant des méthodes propres au traitement du langage naturel afin d’utiliser des techniques issues des dernières recherches dans le domaine.  Dans un second temps, je vais mettre en œuvre 2 approches différentes de recommandation de tags. La première, non supervisée, visera à trouver le sujet principal d’une question et à proposer des mots relatifs au sujet détecté. La seconde, supervisée, visera à généraliser, à des questions non classifiées, les tags des questions déjà classifiées fournis par l’API Stack Overflow.  Le système de recommandation de tags mettant en œuvre les 2 approches sera intégré au travers d’une simple application web. Pour finir, des ouvertures à l'amélioration seront proposées. |

|  |
| --- |
| Mehdi Doubiani  18/03/2020 |

Table des matières

[Récupération des données 1](#_Toc35524605)

[Analyse de la variable *Tags* 2](#_Toc35524606)

[Analyse de la variable *Body* 4](#_Toc35524607)

[Natural Language Processing de la variable *Body* 5](#_Toc35524608)

[Préprocessing des données avant modélisation 6](#_Toc35524609)

[Transformation de la variable Tags 6](#_Toc35524610)

[Séparation des jeux de données en jeux d’entraînement et validation 7](#_Toc35524611)

[Transformation de la variable *Body* en « bag of words » 8](#_Toc35524612)

[Analyse non supervisée 8](#_Toc35524613)

[Méthode Latent Dirichlet Allocation 8](#_Toc35524614)

[Analyse supervisée 12](#_Toc35524615)

[Indice de Jaccard 12](#_Toc35524616)

[Binary Relevance 12](#_Toc35524617)

[Transformation de la variable *Body* avec TermFrequency-InverseDocumentFrequency 12](#_Toc35524618)

[Dummy Classifier 13](#_Toc35524619)

[Recherche par grille du meilleur modèle 13](#_Toc35524620)

[LogisticRegression 13](#_Toc35524621)

[Visualisations 14](#_Toc35524622)

[API 17](#_Toc35524623)

[Pistes d’amélioration 17](#_Toc35524624)

# Récupération des données

L'API de Stack Overflow permet de requêter en SQL les diverses données publiques nécessaires au projet, notamment le contenu de la question ainsi que les tags associés.

Je ne récupère que les documents de type « Question » et leur note qui va me permettre de m’assurer de la qualité de leur contenu. Je ne conserve que les enregistrements ayant un favori count (popularité du post) supérieur à 50.

Exemple de la dernière requête effectuée :

SELECT body, title, tags FROM posts WHERE title is not null and body is not null and tags is not null and FavoriteCount>50

Je charge le tout dans un CSV puis dans un dataframe pandas. J’obtiens alors un jeu de données qui comporte **20 028 questions**.

J’utilise le module Pandas Profiling afin de m’aider à explorer les données. Il en ressort notamment que le jeu de données ne possède **aucune valeur manquante et aucune question en doublon**.

Exemple d’un document :

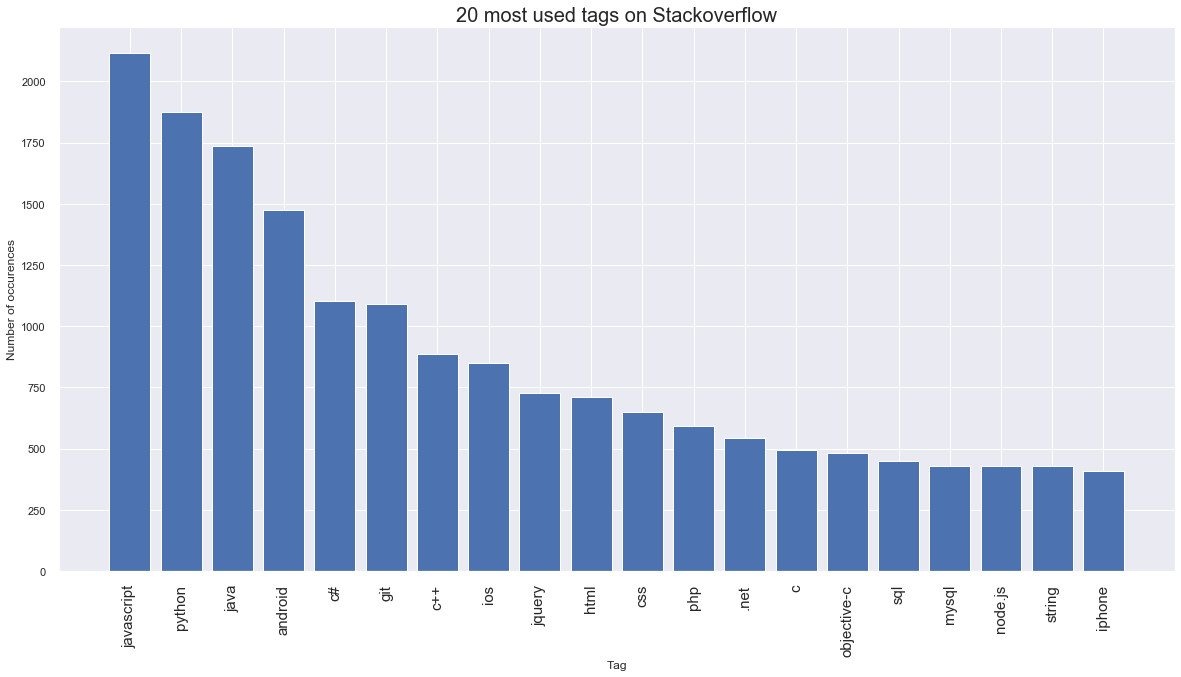
|  |  |
| --- | --- |
| Id | 19142 |
| Quality | 502 |
| Body | <p>How can I request a random row (or as close... |
| Title | How to request a random row in SQL? |
| Tags | <sql><random> |

La colonne *Body* est parsemée de sauts de ligne ou balises HTML et devra faire l’objet d’un nettoyage spécifique. Je vais la fusionner avec la colonne *Title*. Et la variable issue de cette fusion s’appelle Text

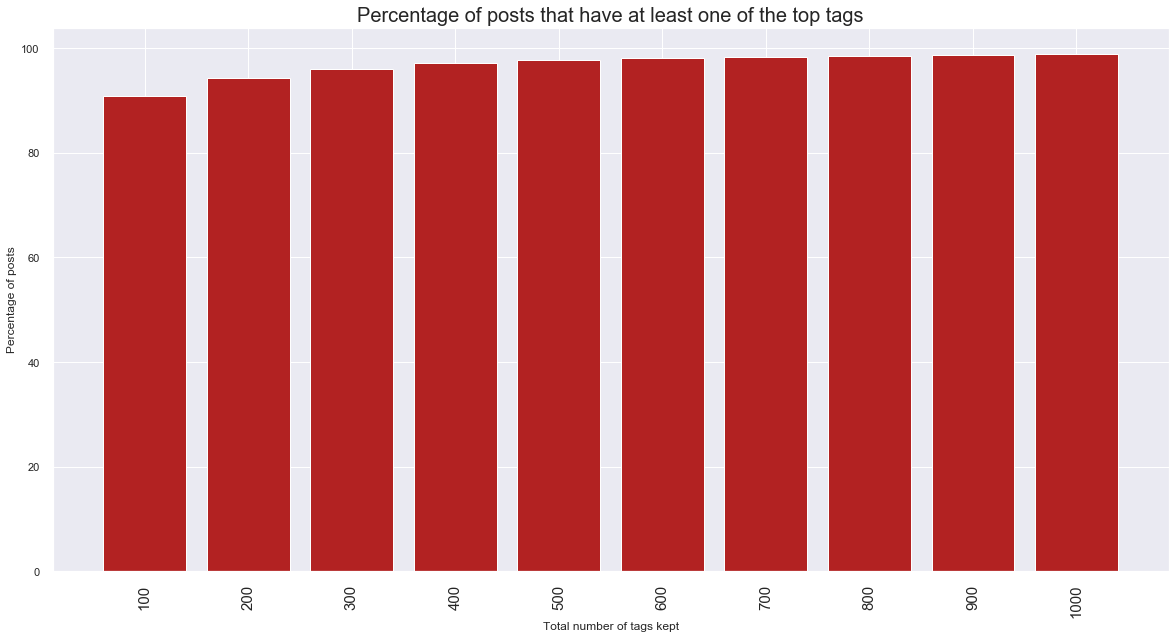
|  |  |
| --- | --- |
|  | La colonne Quality a été crée en additionnant la variable Score et la variable CommentCount (nombre de commentaires). Cela nous donne une indication sur la qualité du post. |

# Analyse de la variable *Tags*

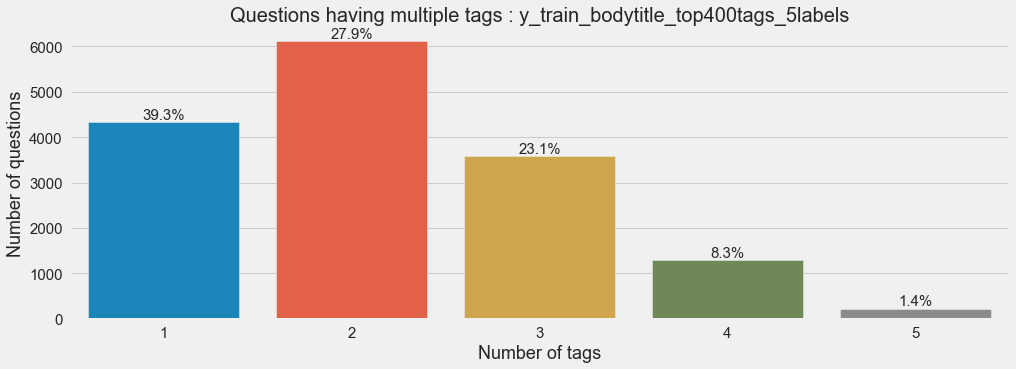
La variable *Tags* comporte **6997 tags distincts**. Je calcule le nombre d’occurrences de chaque tag dans le dataset et produit un graphique du top 20. Les tags les plus fréquents sont javascript, java, python et android. Ce qui est cohérent avec la popularité des développements actuels dans le monde informatique.

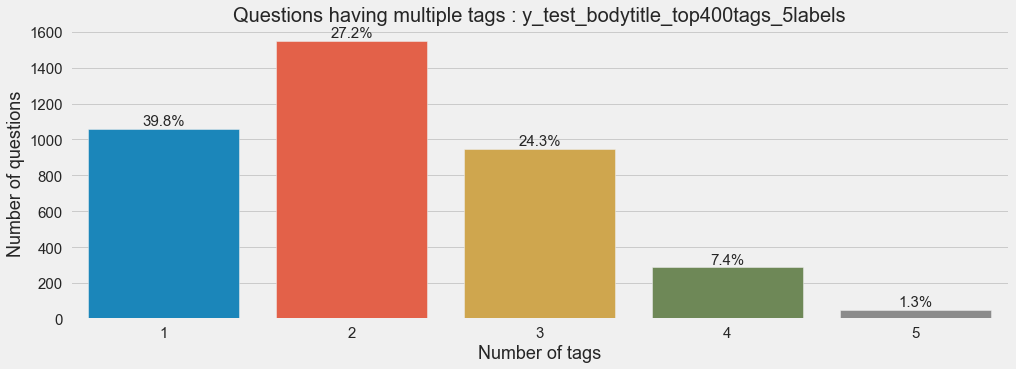


Etant donné le grand nombre de tags, je vais ne conserver qu'un nombre réduit de tags. D'une part, je m’affranchirais en partie du fléau de la dimension, cela me sera plus facile de prédire le bon tag à partir d'un nombre réduit de mots ; d'autre part, dans les presque 7000 tags, certains représentent des technologies peu usitées.

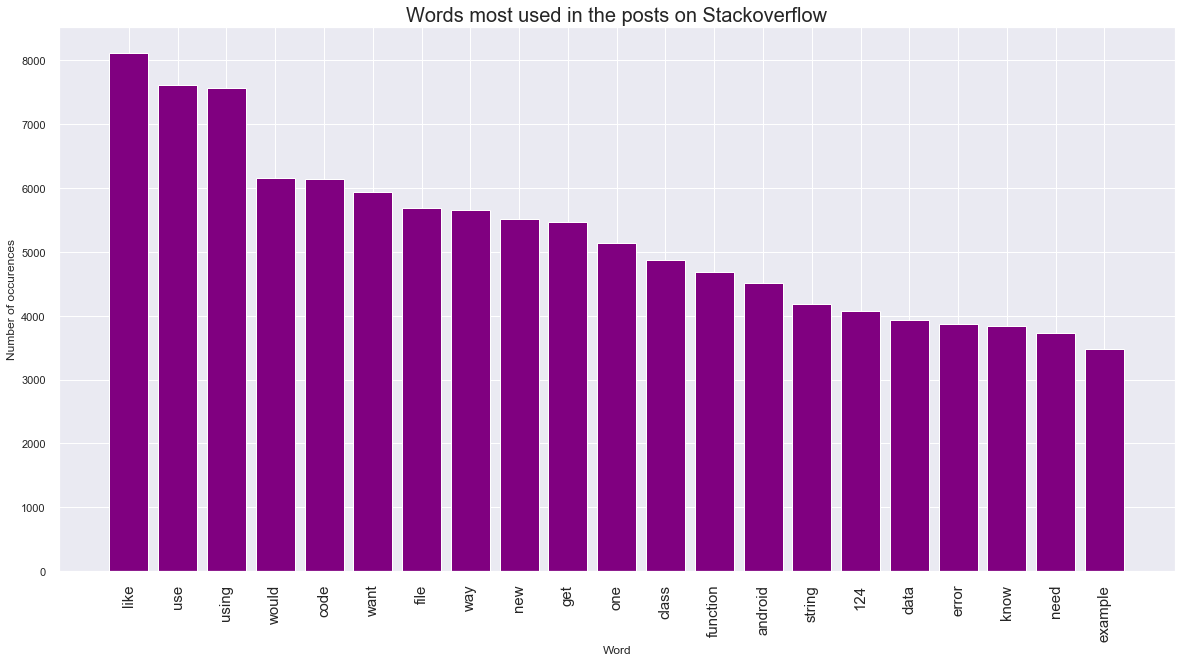


**Ne conserver que les 400 tags les plus fréquents me semble être un bon compromis** dans la mesure où ils apparaissent au moins une fois dans 97% des tags des documents. Je vais donc supprimer les tags exclus de la colonne *tags* et supprimer les documents qui ne possèdent plus de tags. Le dataset possède désormais **19 446** **questions**.

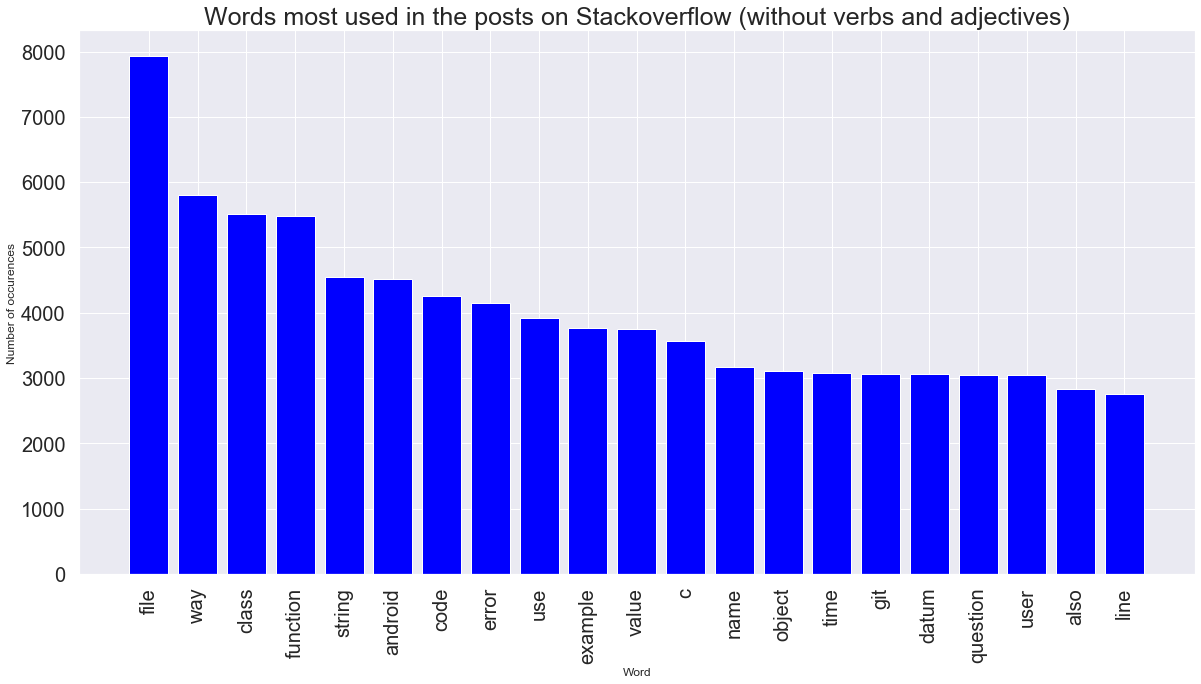


Un document peut posséder jusqu’à 5 tags. L’on constate que la grande majorité des questions contiennent au maximum 3 tags. Mais je décide de toute conserver. L’on constate également une répartition similaire des tags entre le jeu de test et d’entrainement.

# Analyse de la variable *Body*



Les mots les plus utilisés dans les questions avant traitement, c’est surtout des verbes ou des pronoms et/ou adjectifs.



L’on constate une différence une fois les verbes et adjectifs/pronoms retirés.

# Natural Language Processing de la variable *Body*

La variable *Body* nécessite d’être nettoyée et transformée pour être exploitable. Je vais donc lui appliquer des méthodes de processing propre au traitement du langage naturel :

40240 <p>I have the following string and I would like to remove <code>&lt;bpt \*&gt;\*&lt;/bpt&gt;</code> and <code>&lt;ept \*&gt;\*&lt;/ept&gt;</code> (notice the additional tag content inside them that also needs to be removed) without using a XML parser (overhead too large for tiny strings).</p>\r\n\r\n<pre><code>The big &lt;bpt i="1" x="1" type="bold"&gt;&lt;b&gt;&lt;/bpt&gt;black&lt;ept i="1"&gt;&lt;/b&gt;&lt;/ept&gt; &lt;bpt i="2" x="2" type="ulined"&gt;&lt;u&gt;&lt;/bpt&gt;cat&lt;ept i="2"&gt;&lt;/u&gt;&lt;/ept&gt; sleeps.\r\n</code></pre>\r\n\r\n<p>Any regex in VB.NET or C# will do.</p>\r\n Regular expression to remove XML tags and their content

* Mise en minuscules du texte, suppression des caractères « whitespace » et du code au motif que le vocabulaire contenu entre des balises code car il n’apporte rien au but recherché.

40240 <p>i have the following string and i would like to remove and (notice the additional tag content inside them that also needs to be removed) without using a xml parser (overhead too large for tiny strings).</p> <pre></pre> <p>any regex in vb.net or c# will do.</p> regular expression to remove xml tags and their content

* Suppression du format HTML avec le package Beautiful Soup.

40240 i have the following string and i would like to remove and (notice the additional tag content inside them that also needs to be removed) without using a xml parser (overhead too large for tiny strings). any regex in vb.net or c# will do. regular expression to remove xml tags and their content

* Suppression de la ponctuation.

40240 i have the following string and i would like to remove and notice the additional tag content inside them that also needs to be removed without using a xml parser overhead too large for tiny strings any regex in xyzspecialtags16zyx or xyzspecialtags26zyx will do regular expression to remove xml tags and their content

* Suppression des stopwords proposés par les modules NLP (NLTK et Spacy).
* Lemmatisation pour réduire les mots à leur forme neutre canonique.
* Suppression des mots qui ne sont pas des noms (POS tagging) considérant que les verbes ou adverbes n’apportent pas de valeur ajoutée pour la problématique.

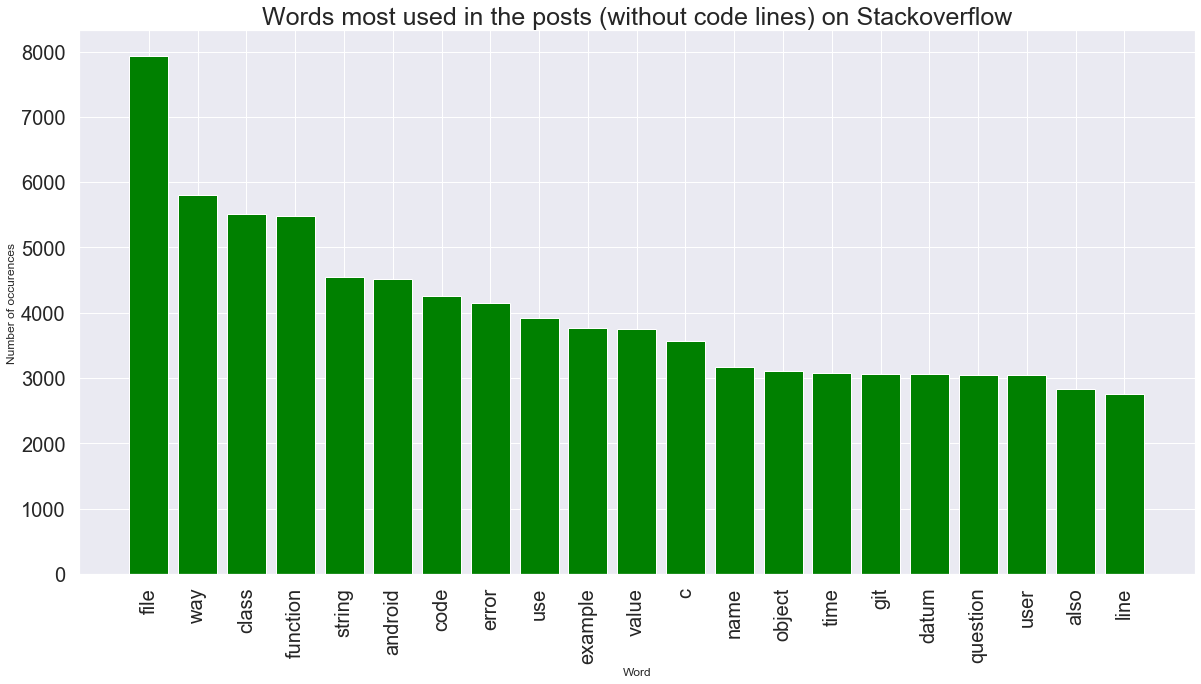
40240 string notice tag content xml parser string regex expression xml tag

* Suppression de stopwords. A ce stade, un rapide coup d’œil sur les 200 premiers mots par occurrence met en évidence que certains noms sont très génériques et n’apportent pas de valeur ajoutée, ils sont supprimés.

'file', 'way', 'user', 'use', 'problem', 'work', 'example', 'method', 'question', 'value', 'thank', 'solution', 'thing', 'number', 'change', 'idea', 'answer', 'issue', 'update', 'lot', 'message', 'information', 'people', 'reason', 'help', 'want', 'run', 'need', 'end', 'default', 'difference', 'suggestion', 'approach', 'task', 'implementation', 'check', 'e', 'custom', 'place', 'practice', 'support', 'experience', 'product', 'stuff', 'comment', 'note', 'argument', 'year'

Conclusion après NLP :

* Toutes les questions possèdent au moins un tag.
* Le top 20 des mots les plus fréquents ci-dessous met surtout en avant des mots propres aux langages informatiques :



# Préprocessing des données avant modélisation

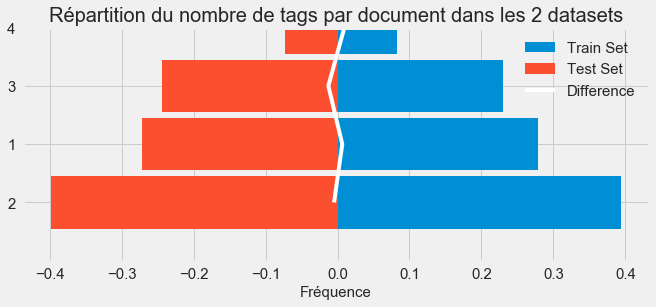
## Transformation de la variable Tags

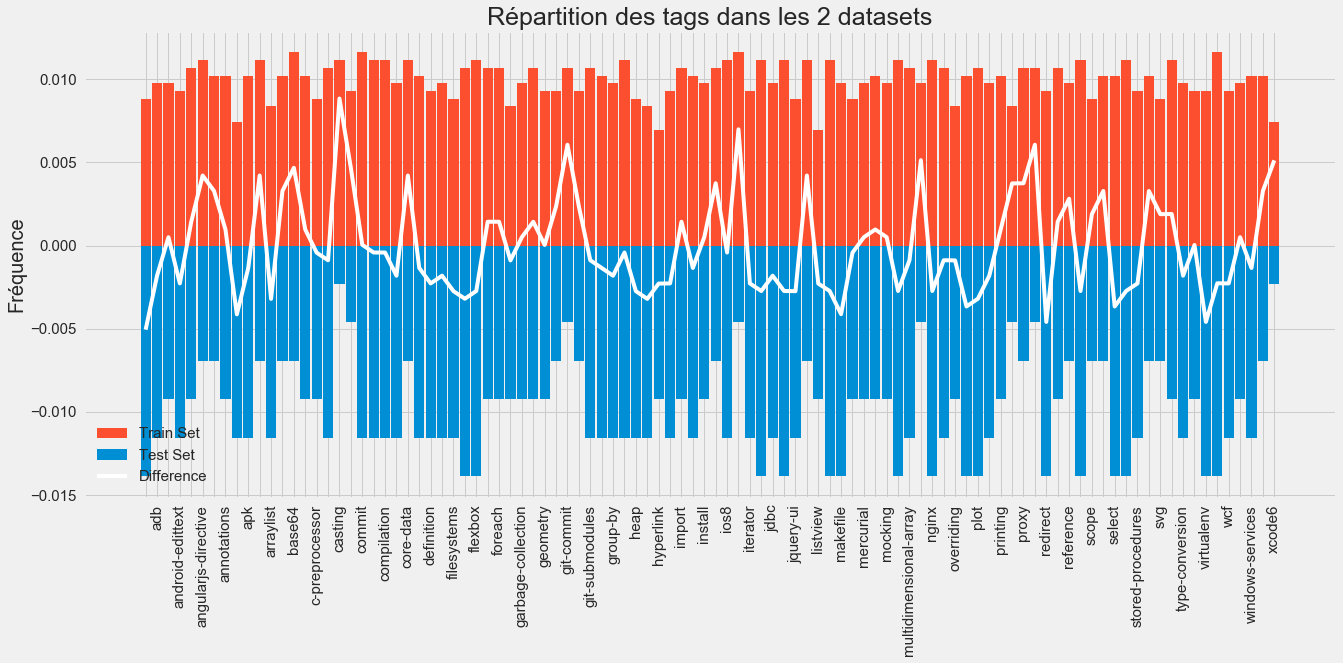
Je commence par transformer la variable *Tags* à l’aide d’un MultiLabelBinarizer pour la modélisation supervisée. J’obtiens une matrice de taille **(posts x tags) soit (19 446 x 400)** **de valeurs binaires** indiquant la présence ou non d’un ou plusieurs tags pour chaque document.



## Séparation des jeux de données en jeux d’entraînement et validation

Je sépare les jeux de données *Tags* et *Body* en jeux d’entraînement (80% soit **15 556 posts**) et validation (20% soit **3 890 posts**). Afin de m’assurer de l’équilibre de la répartition des tags dans les jeux d’entraînement et validation, je produis 2 pyramides, la première, de la répartition du nombre de tags par document dans les 2 datasets, et la seconde, de la fréquence d’apparition des tags dans les 2 datasets.





NB : je n’affiche qu’un tag sur 2

Je constate que la courbe représentant la différence entre les fréquences observées des tags dans le jeu d’entraînement et dans le jeu de validation oscille dans un intervalle de fréquence entre 0.5 et 1%. Les jeux de données sont donc à peu de choses près, bien équilibrés.

## Transformation de la variable *Body* en « bag of words »

Un bag of words est une matrice de dénombrement des mots dans le corpus qui va alimenter le LDA.

Etant donné la grande dimension de mon vocabulaire, je vais chercher à le réduire en fixant une valeur pour le paramètre *min\_df* de la méthode *TfIdfVectorizer* (**term frequency–inverse document frequency)** afin de ne pas intégrer des mots trop rares. J’itère sur plusieurs valeurs entières jusqu’à obtenir des résultats qui me paraissent satisfaisants pour l’analyse non supervisée. Après plusieurs itérations, je retiens ***min\_df = 0.001*** ce qui permet de réduire mon vocabulaire à **1816 mots**,soit une matrice d’entraînement de dimension **15 556 x 1816**.

Par ailleurs, je ne prends en compte que les *unigrams* et modifie le *token\_pattern* par défaut pour tenir compte des mots d’un seul caractère (par exemple, le langage C).

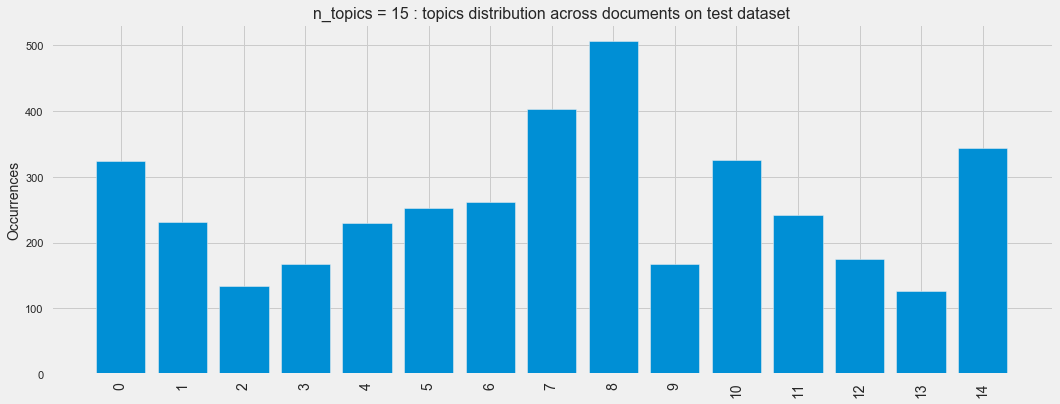
# Analyse non supervisée

## Méthode Latent Dirichlet Allocation

LDA suppose que chaque posts d'un corpus contient un mélange de topics que l'on retrouve dans l'ensemble du corpus. La structure des topics est masquée, nous ne pouvons observer que les mots et posts, pas les topics eux-mêmes. Parce que la structure est cachée (ou latente), cette méthode cherche à déduire la structure des topics du corpus en fonction des mots et des posts connus.

J'ai donc essayé dans un premier temps d'entraîner plusieurs LDA en faisant varier le nombre de topics pour optimiser la perplexité, mais sur sklearn, il semblerait que le calcul de la perplexité souffre d'un bug qui fait qu'elle croît avec le nombre de topics, alors qu'elle devrait décroître.

Je choisis le nombre de topics de façon à ce qu’il discrimine le mieux possible les documents. A mon sens, le meilleur compromis entre l’homogénéité de la répartition des topics dans les documents, la variété des topics et leur cohérence s’obtient avec un **nombre de topics égal à 15**.



Les top 5 des mots décrivant les topics sont globalement plutôt cohérents.

n\_topics = 15 : Top 5 words for each topic on test dataset

|  | **Word 0** | **Word 1** | **Word 2** | **Word 3** | **Word 4** |
| --- | --- | --- | --- | --- | --- |
| **Topic 0** | text | input | html | jquery | page |
| **Topic 1** | git | repository | mysql | commit | github |
| **Topic 2** | image | docker | install | container | path |
| **Topic 3** | branch | git | directory | master | scope |
| **Topic 4** | array | exception | line | element | xml |
| **Topic 5** | java | php | vim | random | number |
| **Topic 6** | json | request | xcode | http | api |
| **Topic 7** | android | table | column | sql | database |
| **Topic 8** | project | error | import | build | application |
| **Topic 9** | python | number | algorithm | plot | pip |
| **Topic 10** | difference | object | list | js | node |
| **Topic 11** | javascript | css | jquery | browser | element |
| **Topic 12** | bar | foo | interface | value | const |
| **Topic 13** | date | commit | git | net | asp |
| **Topic 14** | static | memory | java | language | programming |

|  |  |
| --- | --- |
| **Topic 0**    Vocabulaire orienté web | **Topic 2**    Vocabulaire virtualisation, cloud et système |
| **Topic 3**    Vocabulaire orienté développement et gestion | **Topic 11**    Vocabulaire développement web |
| **Topic 9**    Vocabulaire orienté data science | **Topic 7**    Vocabulaire orienté BDD |
| **Topic 14**    Vocabulaire orienté langage objet | **Topic 12**    Vocabulaire orienté interfaces UI/UX |

Le modèle LDA appliqué à une matrice TF permet de produire 2 matrices :

* Une de dimension (d x t) contenant les probabilités des topics sachant le document
* Une de dimension (t x w) contenant les probabilités des mots sachant le topic

Je peux donc opérer un produit matriciel entre les 2 matrices pour obtenir une matrice de dimension (d x w) qui contiendra les probabilités des mots sachant le document. Il ne me restera plus qu'à choisir les N mots les plus probables pour proposer pour un document donné, les N mots les plus liés au topic latent du document.

Je réalise une fonction permettant de recommander 5 mots pour des documents passés en paramètre et observe sur quelques échantillons que si les propositions manquent de spécificité, certains termes demeurent assez adaptés.

|  |  |  |
| --- | --- | --- |
|  | **Mots proposés** | **Cleaned Body** |
| **44492** | python, c++, c, language, function | c c++ loop statement |
| **7894** | control, form, button, string, server | parentusercontrol host load parentusercontrol access property parentusercontrol childusercontrol time property parent control child control |
| **36398** | time, memory, class, program, process | class asset class class definition asset getdefinition class definition asset getdefinitionbyname |
| **17197** | java, project, c++, eclipse, source | java effect point operation java |
| **64422** | web, asp.net, application, site, object | application predispatch uri authentication store session zend request object thought authentication zend |

# Analyse supervisée

## Indice de Jaccard

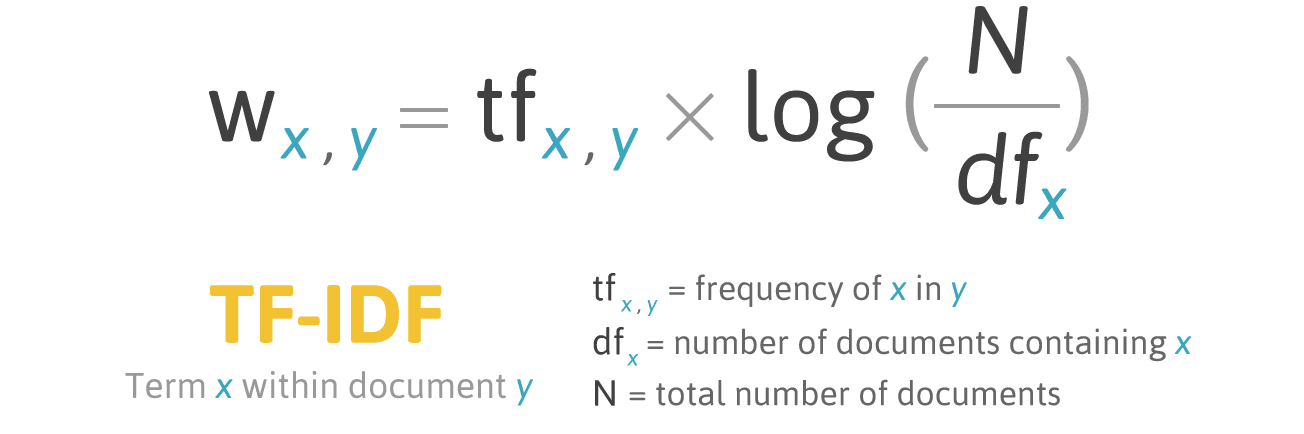
|  |  |
| --- | --- |
| L’indice de Jaccard, défini comme étant la taille de l’intersection divisée par la taille de l’union de 2 ensembles de labels, est utilisé pour comparer un ensemble de labels prédits pour un échantillon aux labels réels correspondant.  Il s’étend aux problématiques « multilabel » en calculant une moyenne du score obtenu pour chaque label. J’utilise l’indice de Jaccard pondéré qui détermine la moyenne des métriques calculée pour chaque label, pondérée par leur distribution réelle observée. |  |

## Binary Relevance

|  |  |
| --- | --- |
| Résultat de recherche d'images pour "binary relevance multilabel" | Dans le cas d’un problème « multilabel », il est possible d’avoir plusieurs labels pour une même instance.  Je vais utiliser l’approche Binary Relevance pour décomposer la tâche d'apprentissage multilabel en un certain nombre de tâches d'apprentissage binaires indépendantes (une par label), tout en ayant bien conscience que je préjuge de l’indépendance des labels entre eux.  Sur Sklearn, la méthode Binary Relevance est implémentée dans *OneVsRestClassifier*. |

## Transformation de la variable *Body* avec TermFrequency-InverseDocumentFrequency

Je transforme désormais le vocabulaire de mon corpus en une matrice de valeurs numériques en utilisant TF-IDF. Cette méthode de pondération vise à accorder une pertinence lexicale à un terme au sein d’un document. Un terme aura plus de chances d’être pertinent pour un document, si celui-ci en possède une occurrence plus élevée en son sein que les autres documents où le terme apparaît. La pertinence lexicale se mesure donc avec TF-IDF grâce à une relation entre la rareté d’un terme au sein d’un ensemble de documents et son occurrence dans un seul document.



Etant donné la grande dimension de mon vocabulaire, je vais chercher à le réduire en optimisant les différents paramètres *min\_df*, *max\_df* et *max\_features* de *TfidfVectorizer* à l’aide d’une recherche par grille visant à maximiser l’indice *jaccard\_weighted en %* qui me sert de métrique d’évaluation pour l’analyse supervisée.

Par ailleurs, je ne prends en compte que les *unigrams* et modifie le *token\_pattern* par défaut pour tenir compte des mots d’un seul caractère (par exemple, le langage C).

Etant donné la grande dimension de mon vocabulaire et le nombre important de paramètres à tuner, je vais entraîner un classifieur *MultinomialNB* (multinomial Naive Bayes) peu gourmand en ressources.

Le meilleur résultat est donné pour ***min\_df=0.001***, ***max\_df=0.1*** et ***max\_features=2000***, ce qui porte mon vocabulaire à **1 816 mots**,soit une matrice d’entraînement de dimension **15 556 x 1816**.

## Dummy Classifier

Pour évaluer l‘intérêt des modèles que j’entraîne, je vais comparer leur résultat à un classifieur naïf que j’ai entraîné à prédire en respectant la distribution des labels dans l’échantillon d’entraînement.

|  |  |
| --- | --- |
| Jaccard weighted Train | 1.6172225375310207 |

## Recherche par grille du meilleur modèle

J’ai procédé à une comparaison exhaustive des modèles suivants avec leur score de jaccard weighted en % respectifs :

* DummyClassifier : 1.62%
* SGDClassifier : 36.45 %
* LogisticRegression : 28.45%
* MultinomialNB : 13.06%
* LinearSVC : 39.78%
* Perceptron : 35.41%
* PassiveAggressiveClassifier : 39.95%
* MLPClassifier : 38.71%
* XGBClassifier : 40.45%
* KNeighborsClassifier : 20.33%
* RandomForestClassifier : 32.45%

On peut déjà constater que le XGBClassifier arrive déjà en tête en matière de score de jaccard , et ceci sans aucune optimisation d’hyperparamètres.

Après mise en place d’une optimisation d’hyperparamètres exhaustives pour tous les modèles cités, il s’avère que c’est la régression logistique qui offre le score de jaccard le plus élevé.

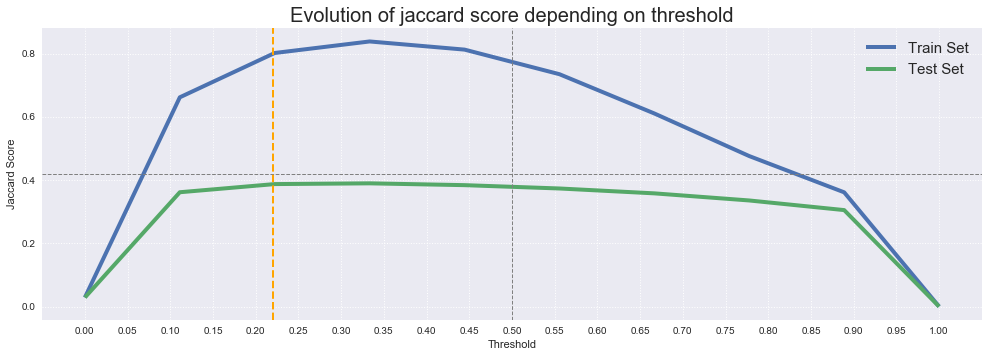
### LogisticRegression

|  |  |
| --- | --- |
| Paramètres testés | penalty : ['l1', 'l2']  C : [0.1, 1, 10, 100, 1000] |
| Meilleurs paramètres | penalty : l1  C : 10 |
| Jaccard weighted Test % | 41.04% |

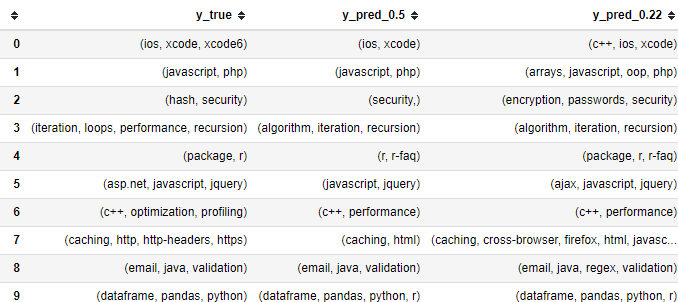
Avec la régression logistique, la performance s’améliore mais le sur-apprentissage existe malgré la pénalisation. C’est le modèle que je vais retenir pour l’API.

**Optimisation du threshold**

Je teste différentes valeurs pour le seuil de probabilité visant à déterminer la classe binaire finale pour chaque label mais constate que modifier le seuil par défaut (= 0.5) n’aurait pas vraiment d’impact sur la capacité du modèle à mieux généraliser. Toutefois, je vais réduire le **seuil à 0.22**, ce qui va avoir pour effet de favoriser la capacité de mon modèle à fournir une prédiction sans en dégrader la performance.



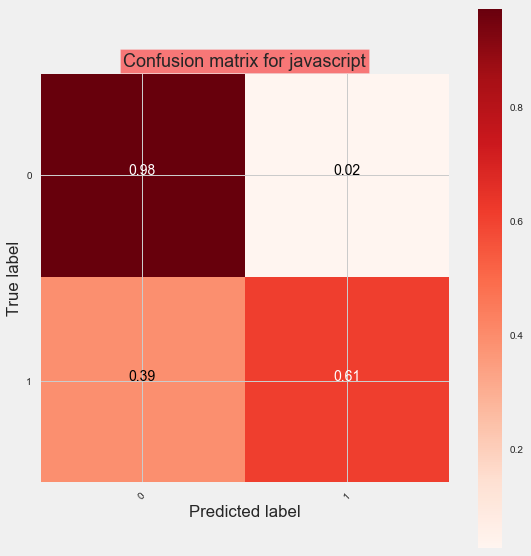
On peut constater l’impact de la modification du seuil sur un échantillon tiré aléatoirement, le modèle est plus enclin à proposer un tag pour un seuil à 0.22 contrairement au seuil à 0.5, sans pour autant que le ou les tags proposés ne soient pas pertinents.



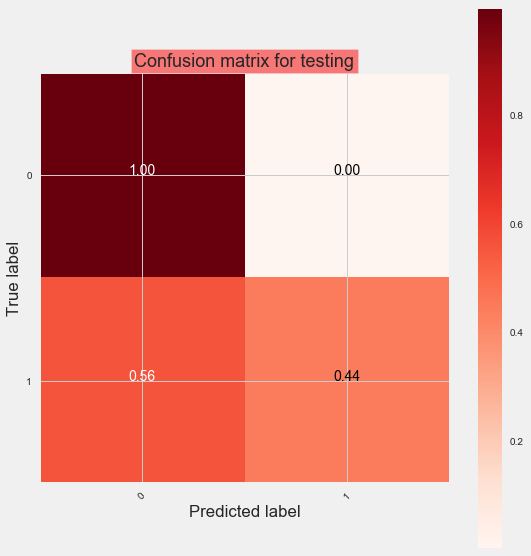
# Visualisations

En visualisant la matrice de confusions pour des tags dans le top, middle et bottom des 400 premiers tags , il s’avère que le modèle donne à peu près la même performance de prédiction.

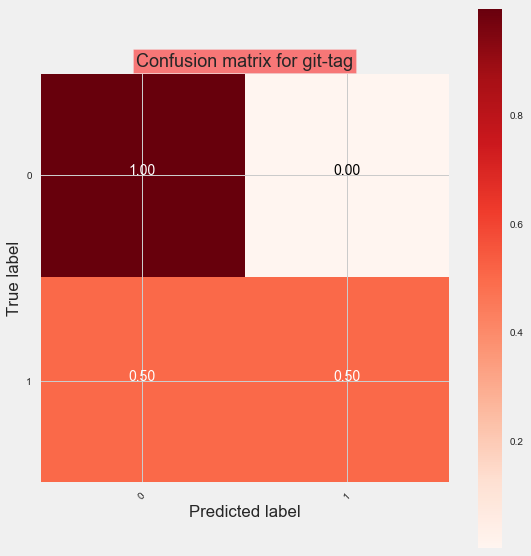
TOP LABEL



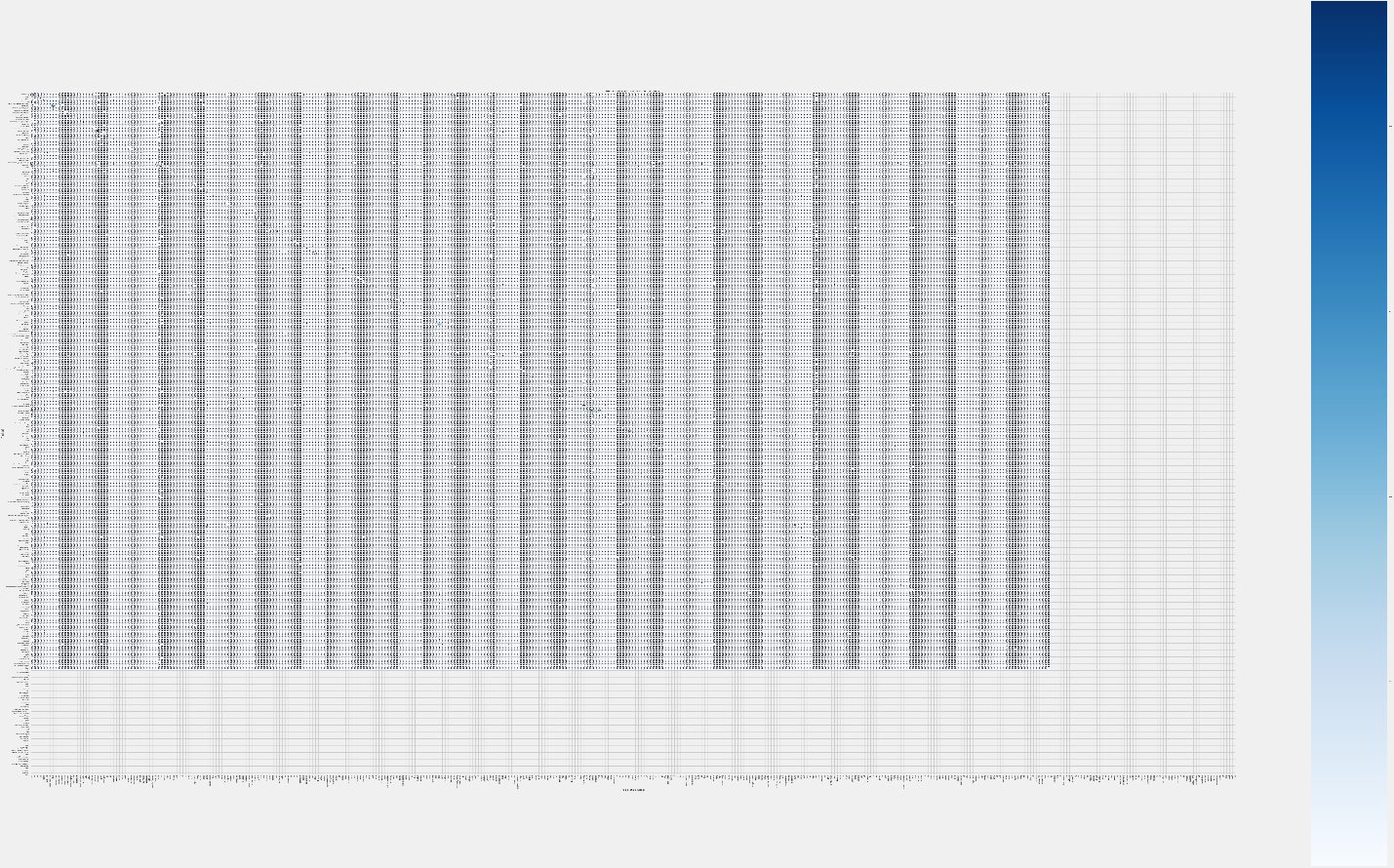
MIDDLE LABEL



BOTTOM LABEL



La visualisation d’une matrice de 400x400 est très difficile à interpréter mais cela donne une idée de la performance de notre modèle



# API

L’application est réalisée en utilisant Dash. C’est une application basique qui propose une liste de tags StackOverflow (jusqu’à 5 prédits) relatifs à une question saisie traitant de sujets informatiques et en Anglais.

Pickles (.pkl) nécessaires :

* multilabel\_binarizer (préprocessing) : multilabelbinarizer pour transformer les prédictions supervisées en libellé
* vectorizer\_dfText (préprocessing) : transformer TFIDF
* lda\_model (recommandation) : modèle non supervisé
* top\_tags (recommandation) : les 400 top tags

L’ensemble des fonctions réalisées dans le cadre de ce projet ont été stockées dans un module *utils.py* :

* clean\_text
* clean\_punct
* stopWordsRemove
* lemmatization
* pred\_nwords\_unsupervised
* recommend\_tags
* avg\_jaccard

Avant prédiction, le texte saisi passe par toutes les étapes de préprocessing NLP évoqués plus haut :

text = text.apply(lambda s: clean\_text(s))

text = text.apply(lambda s: BeautifulSoup(s,features="lxml").get\_text())

text = text.apply(lambda s: clean\_punct(s,top\_tags))

text = text.apply(lambda s: stopWordsRemove(s))

text = text.apply(lambda s: lemmatization(s,['NOUN', 'ADV'],top\_tags))

Puis il est transformé en matrice TF-IDF avant application des modèles supervisés et non supervisés.



# Pistes d’amélioration

* Pour le modèle non supervisé, supprimer les mots les plus fréquents par topic pourrait permettre d’amener un peu plus de spécificité. Je pourrais peut-être aussi gagner en spécificité en intégrant des n-grams ou en utilisant des techniques de plongements de mots.
* Il faudrait arriver à gérer l’effet de bord induit par le faible taille des documents qui n’aide pas à bien discriminer les mots qui importent dans chaque document au sein d’une matrice TF-IDF.
* Je pourrais éviter de prédire les mêmes tags entre supervisé et non supervisé et donc intégrer les tags comme stopwords pour la matrice TF-IDF.
* Au lieu d’avoir un unique modèle supervisé pour prédire l’ensemble des tags, il ne serait pas inintéressant d’isoler les tags identifiant une technologie des autres tags. Je spécialiserais alors un classifieur à prédire la technologie ou langage concernée par la question tandis qu’un autre classifieur se concentrerait plutôt à décrire la nature du problème.
* Utiliser les dernières techniques issues du deep learning comme BERT ou ELMo