

# Catégoriser automatiquement des questions

Elise Andro – 06/08/2021

[Table des matières](#)

Introduction.....	2
Qui sont les clients ?.....	2
Quelle est leur problématique ?.....	3
Quelles données avons-nous à disposition ? .....	3
Traitement des données .....	3
Nettoyage .....	3
Analyse Descriptive .....	4
Analyse des tags .....	5
Modèles non-supervisés .....	6
Modèles supervisés .....	9
Conclusion .....	10

## Introduction

### Qui sont les clients ?



Stack Overflow est un site web proposant des questions et réponses sur un large choix de thèmes concernant la programmation informatique. Il fait partie du réseau de sites Stack Exchange.

### Quelle est leur problématique ?

L'objectif de ce projet est de simplifier l'utilisation du site StackOverFlow et sa recherche par tag. Les utilisateurs entrent des tags pour rechercher une question et ses réponses.

Développer un système de suggestion de tag pour le site. Le système assignera automatiquement plusieurs tags pertinents à une question.

### Quelles données sont à disposition ?

Données : requête SQL depuis le site <https://data.stackexchange.com/stackoverflow/query/new>

```
1  select top(50000) Id, ViewCount, Body, Title, Tags, Score
2  from Posts
3  where ViewCount > 10000 and Score > 100
```

## Traitement des données

### Nettoyage

Les étapes de nettoyage effectués sont les suivantes :

- ✓ Nettoyage des caractères spéciaux (html)

- ✓ Fusion du titre et du contenu
- ✓ Tokenization
- ✓ Stop words
- ✓ Bigram et trigram
- ✓ Lemmatization

Résultats du nettoyage :

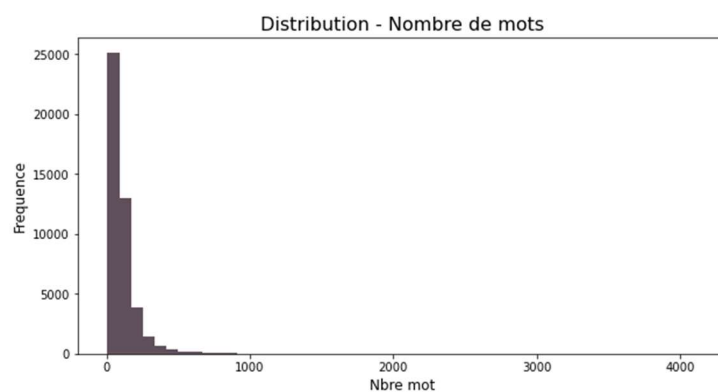
	content	Tags	preprocess_content
0	is there any way to tinker with the iphone sdk...	ios,iphone,windows	[way, machine, plan, version, way, think, run,...
1	i have this gigantic ugly string j transaction...	regex	[gigantic, ugly, string, transaction, start, p...
2	i am working with autolayout and constraints a...	ios,xcode,storyboard,autolayout,xcode6	[work, autolayout, constraint, find, constrain...
3	i m learning objective c and keep bumping into...	objective-c	[learn, objective, keep, bump, symbol, use, di...
4	is there a simple way in a pretty standard uni...	bash,unix,scripting	[simple, way, pretty, standard, unix, environm...

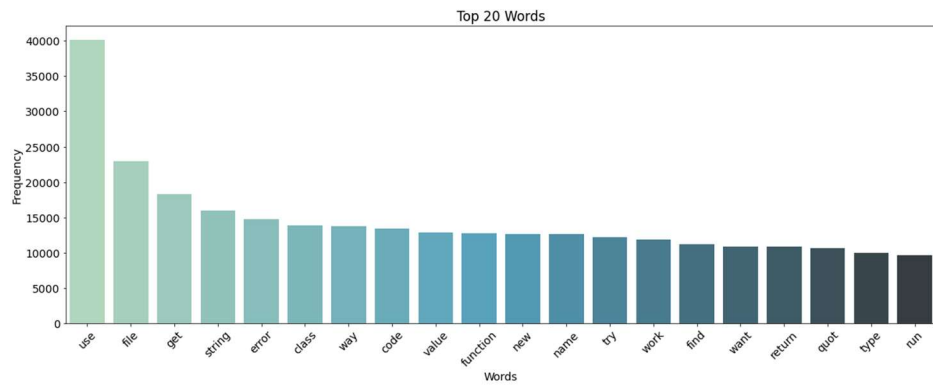


	Body	Title	Tags
0	<p>Is there any way to tinker with the iPhone ...	How can I develop for iPhone using a Windows d...	<ios><iphone><windows>
1	<p>I have this gigantic ugly string:<p>\n<pre>...	My regex is matching too much. How do I make i...	<regex>
2	<p>I am Working with autolayout and constraint...	What is "Constrain to margin" in Storyboard in...	<ios><xcode><storyboard><autolayout><xcode6>
3	<p>I'm learning objective-c and keep bumping i...	What does the @ symbol represent in objective-c?	<objective-c>
4	<p>Is there a simple way, in a pretty standard...	Delete all but the most recent X files in bash	<bash><unix><scripting>

## Analyse Descriptive

Analyse et exploration du contenu (titre+texte) pour prendre en main le corpus de document et avoir une idée de la diversité des données.



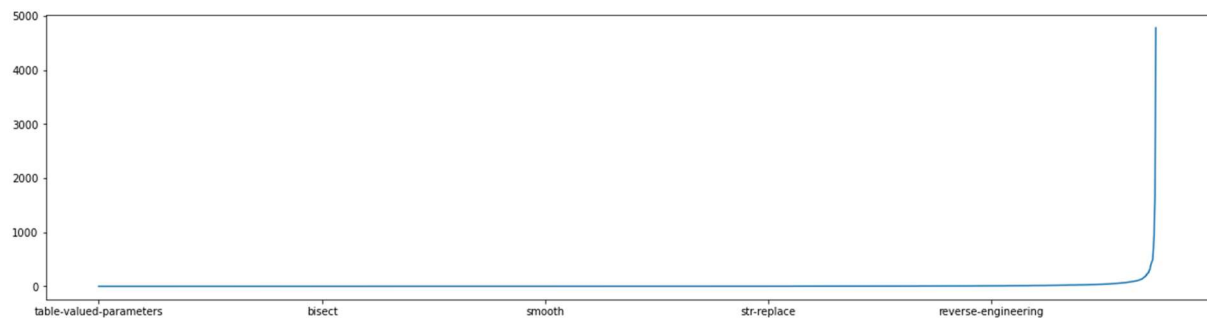


## Analyse des tags

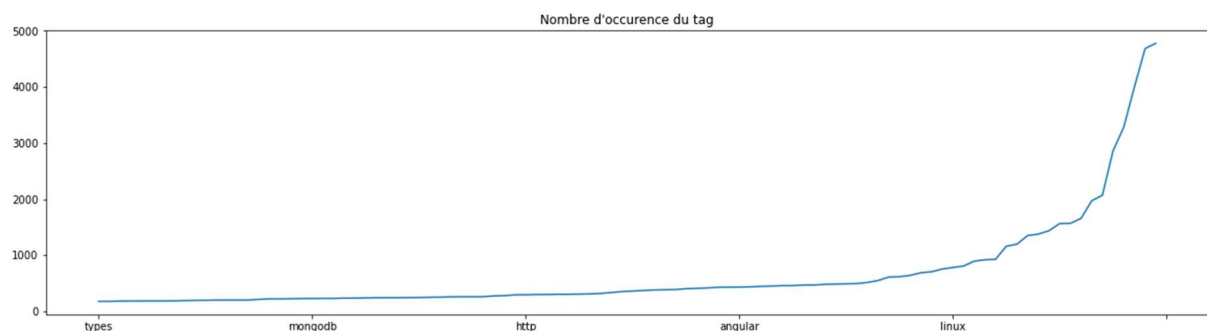
Pour analyser les tags, on crée un dataframe contenant une colonne par tags :

	content	Tags	preprocess_content	word_count	between	visual-c++-2012	cllocationmanager	cpp-core-guidelines	android-background	no-www	...	final	trailing	tic-tac-toe	am ec;
0	is there any way to tinker with the iphone sdk...	[ios, iphone, windows]	[way, machine, plan, version, way, think, run,...]	70	0	0	0	0	0	0	...	0	0	0	0

Ce dataframe permet d'effectuer l'analyse du nombre de tags et fréquence des tags. On peut constater que beaucoup de tags sont très peu utilisés.

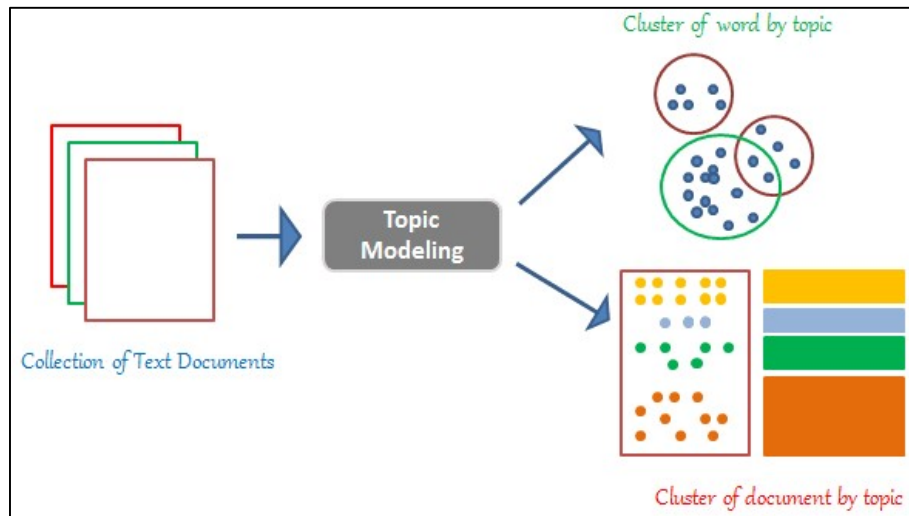


Pour faciliter l'analyse et les modèles, on ne conserve dans un premier temps que les 100 tags les plus utilisés :



## Modèles non-supervisés

Topic Modeling, le principe :



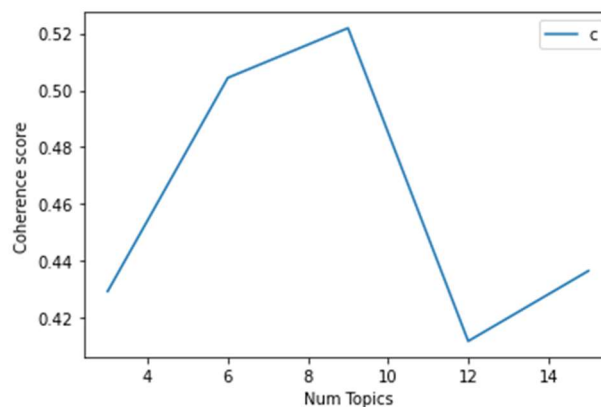
### Latent Dirichlet Allocation

Il s'agit d'un modèle génératif probabiliste permettant d'expliquer des ensembles d'observations, par le moyen de groupes non observés, eux-mêmes définis par des similarités de données. Le modèle LDA suppose que chaque document est un mélange d'un petit nombre de sujets ou thèmes, et que la génération de chaque occurrence d'un mot est attribuable (probabilité) à l'un des thèmes du document.

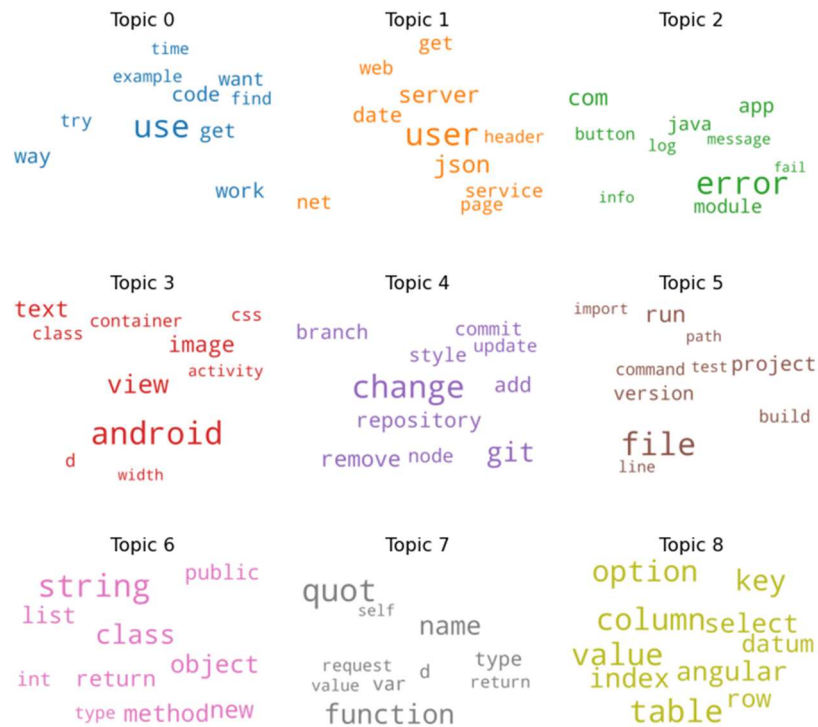
Application du modèle LDA aux données

En amont, il est nécessaire de créer un dictionnaire et un corpus pour entrainer le modèle.

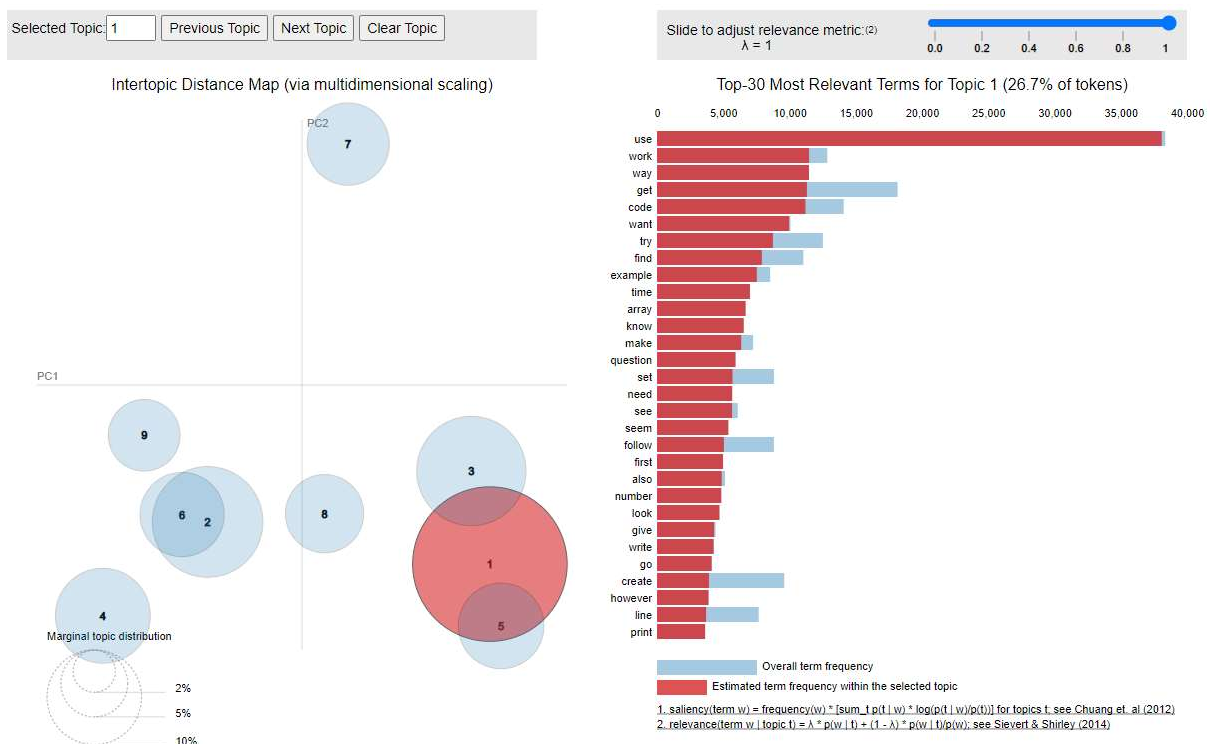
On entraîne ensuite plusieurs modèles pour choisir le nombre optimal de topics en utilisant le score de cohérence comme métrique. Ici l'optimal est **9 topics**.



On visualise enfin les **top words** pour chaque topics :



Le package pyLDAvis permet de visualiser les résultats du modèle LDA sur les 2 axes principaux d'une ACP. On peut distinguer les distances entre les topics, leur importance dans le corpus et le top30 des mots pour chaque topics.

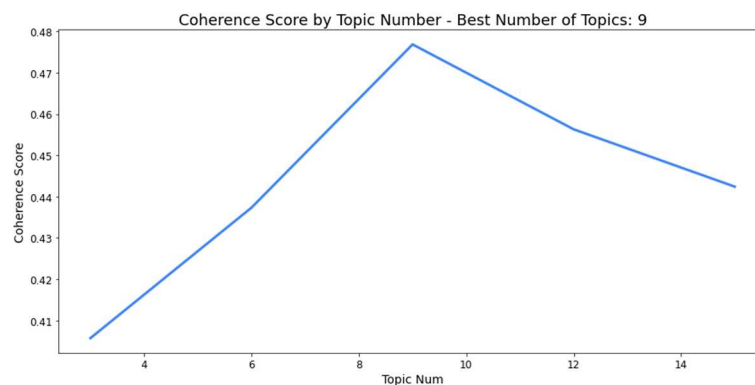


## Non-Negative Matrix Factorization

La NMF est donc une technique de réduction de dimension adaptée aux matrices creuses contenant des données positives. Le modèle classe les documents sur la base d'un vocabulaire de mots. Chaque document se décompose (coefficients positifs) sur des "facteurs" ou thèmes, eux-mêmes définis chacun par un sous-ensemble de ces mots.

### Application du modèle NMF aux données

Comme pour le LDA, on entraîne plusieurs modèles pour choisir le nombre optimal de topics en mesurant le score de cohérence. L'idéal est également de **9 topics**.



On utilise le modèle optimal pour effectuer des prédictions. Pour cela, on entraîne le modèle sur une part du dataset et on effectue des prédictions sur l'autre part.

On obtient pour chaque document du test set un topic ainsi qu'une liste de 10 mots (les top 10 du topics) qui peuvent être une prédiction de tags.

	content	preprocess_content	topic_num	topics
0	i m learning objective c and keep bumping into...	[learn, objective, keep, bump, symbol, use, di...	2	string convert character format replace number...
1	is there a simple way in a pretty standard uni...	[simple, way, pretty, standard, unix, environm...	1	file line directory command folder path open r...
2	without local access to the server is there an...	[local, access, server, way, duplicate, clone,...	5	table column row value select query name datab...
3	what are the best practices to consider when c...	[best_practice, consider, catch, exception, th...	0	use class function error method code get objec...
4	i am trying to insert into a table using the i...	[try, insert, table, use, input, table, entire...	5	table column row value select query name datab...
...	...	...	...	...
10177	after the latest update of php intelephense th...	[late, update, php, intelephense, get, today, ...	0	use class function error method code get objec...
10178	seems pretty googleable but haven t been able ...	[seem, pretty, googleable, able, find, online,...	0	use class function error method code get objec...
10179	i m starting to play with the create react app...	[start, play, create, react, app, index, load,...	3	quot name json class type file
10180	recently i upgraded the version of django fram...	[recently_upgrade, version, django, framework,...	1	file line directory command folder path open r...
10181	using suggested method this is the result a li...	[use, suggest, method, result, link, button, c...	7	android image text button css view element inp...

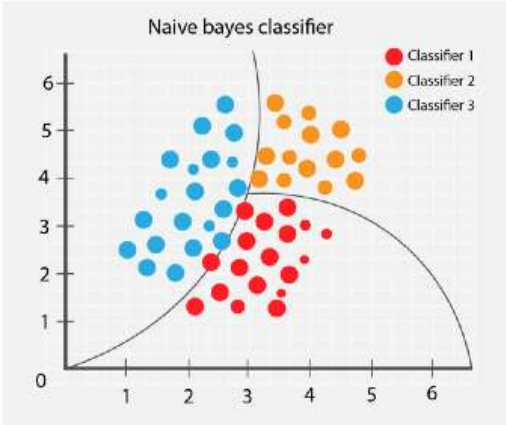
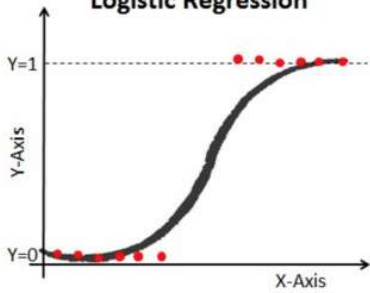
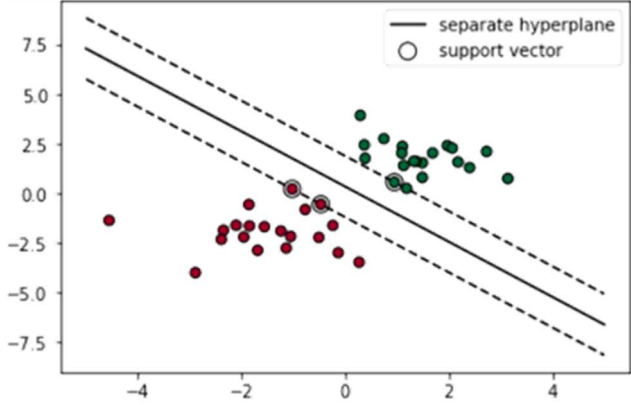


## Modèles supervisés

### Pre-process : TF-IDF

Tf-Idf signifie fréquence de terme-fréquence de document inverse, et au lieu de calculer le nombre de mots dans chaque document de l'ensemble de données, il calcule le nombre normalisé où chaque nombre de mots est divisé par le nombre de documents dans lesquels ce mot apparaît.

### Test sur différents algorithmes

Algorithme	Principe	Illustration	Résultats
Baseline	Classifier qui prédit la classe la plus fréquente. Permet de comparer les résultats avec les autres modèles		Accuracy = 0.12 F1_score = 0.02
Naive Bayes	Classifier qui suppose l'indépendance des variables entre elles	 A scatter plot titled 'Naive bayes classifier' showing three classes of data points: Classifier 1 (red), Classifier 2 (orange), and Classifier 3 (blue). The points are distributed in a 2D space with axes from 0 to 6. Three curved decision boundaries separate the classes.	Accuracy = 0.45 F1_score = 0.42
Logistic Regression	Classifier permettant d'assigner une probabilité d'appartenance à une classe (0/1) grâce à la fonction sigmoid	 A graph titled 'Logistic Regression' showing a sigmoid curve on a coordinate system. The Y-axis is labeled 'Y-Axis' and the X-axis is labeled 'X-Axis'. The curve starts near Y=0 and approaches Y=1. Red dots representing data points are plotted along the curve.	Accuracy = 0.47 F1_score = 0.46
Support Vector Machine	Classifier recherchant à maximiser les marges entre les échantillons les plus proches (vecteurs supports)	 A scatter plot illustrating a Support Vector Machine (SVM) decision boundary. It shows two classes of data points (red and green) separated by a solid line labeled 'separate hyperplane'. Dashed lines parallel to the hyperplane represent the margins. Two points on the margins are marked with circles and labeled 'support vector' in the legend.	Accuracy = 0.46 F1_score = 0.43

Random Forest	La forêt aléatoire se repose sur plusieurs modèles dits « faibles » d'arbres de décision puis les combinent en prenant la solution de la majorité.	<p><b>Random Forest Simplified</b></p> <pre> graph TD     Instance --&gt; RF[Random Forest]     RF --&gt; T1[Tree-1]     RF --&gt; T2[Tree-2]     RF --&gt; Dots[...]     RF --&gt; Tn[Tree-n]     T1 --&gt; CA[Class-A]     T2 --&gt; CB1[Class-B]     Dots --&gt; CB2[Class-B]     Tn --&gt; CB3[Class-B]     CA --&gt; MV[Majority-Voting]     CB1 --&gt; MV     CB2 --&gt; MV     CB3 --&gt; MV     MV --&gt; FC[Final-Class] </pre>	Accuracy = 0.456 F1_score = 0.46
---------------	--	--	-------------------------------------

## Conclusion

Le modèle de Régression Logistic apporte les meilleurs résultats pour développer une fonctionnalité de génération automatique de tags.