



stackoverflow

CATEGORISATION AUTOMATIQUE DES QUESTIONS

NLP



TABLE OF CONTENTS



01

RAPPEL DU CONTEXTE

Présentation du besoin et du support Stack Overflow

02

TRAITEMENT DES DONNEES

Filtrage des documents et pré-traitement

03

ENTRAINEMENT DES MODELES

Comparaison des approches supervisées et non supervisées

04

API

Développement de l'outil et mise en production





[01]

RAPPEL DU CONTEXTE



LE BESOIN



SUPPORT DE REFERENCE

- Création : 2008
- 100 millions visiteurs uniques / mois
- 4 questions par minute

DEMANDE

Un outil de suggestion de tags



Ask a public question











Title

Be specific and imagine you're asking a question to another person

e.g. Is there an R function for finding the index of an element in a vector?

Body

Include all the information someone would need to answer your question

B *I*









[Links](#)
[Images](#)
[Styling/Headers](#)
[Lists](#)
[Blockquotes](#)
[Code](#)
[HTML](#)
[Tables](#)
[More](#)

code

bold

italic

>quote

Tags

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

e.g. (java reactjs json)

[Review your question](#)

Step 1: Draft your question

The community is here to help you with specific coding, algorithm, or language problems.

Avoid asking opinion-based questions.

1. Summarize the problem

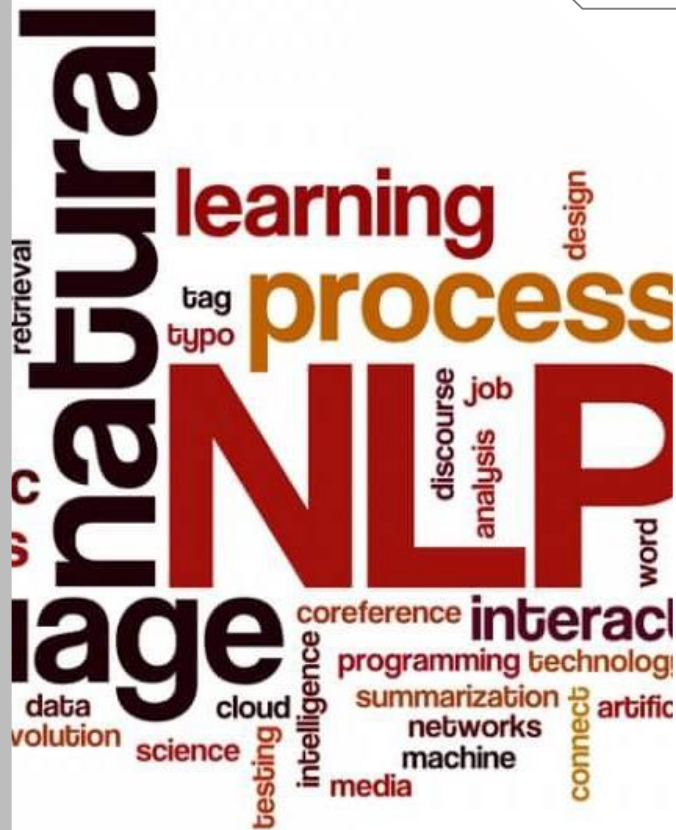
- Include details about your goal
- Describe expected and actual results
- Include any error messages

2. Describe what you've tried

3. Show some code

Have a non-programming question?

More helpful links



[02]

TRAITEMENT DES DONNEES





LA BASE DE DONNEES



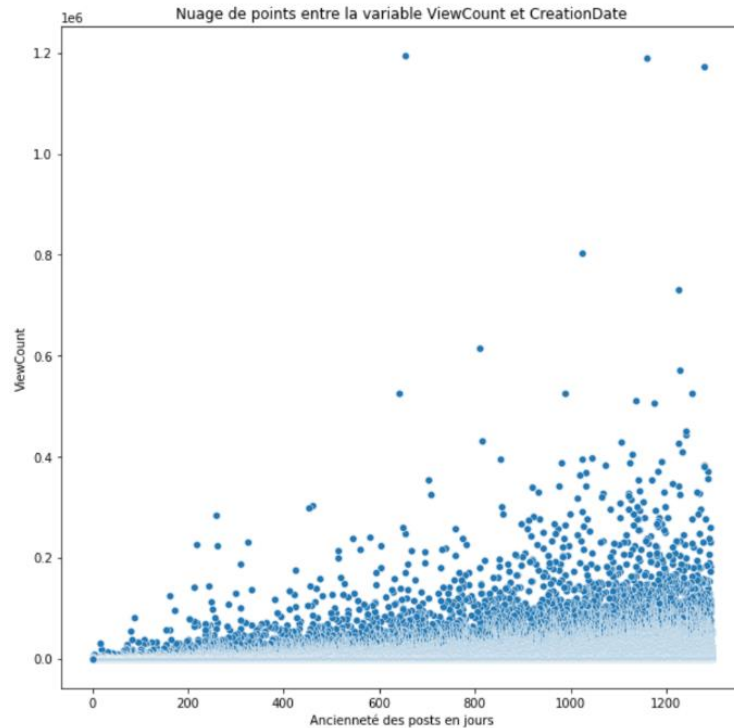
16,6 Go de données pour ma table posts

- Période : janvier 2018 à juillet 2021
- Score non nul
- Nombre de vues non nul
- Nombre de réponses non nul
- Nombre de commentaires non nul
- Nombre de mise en relation non nul





ANCIENNETE DES POSTS



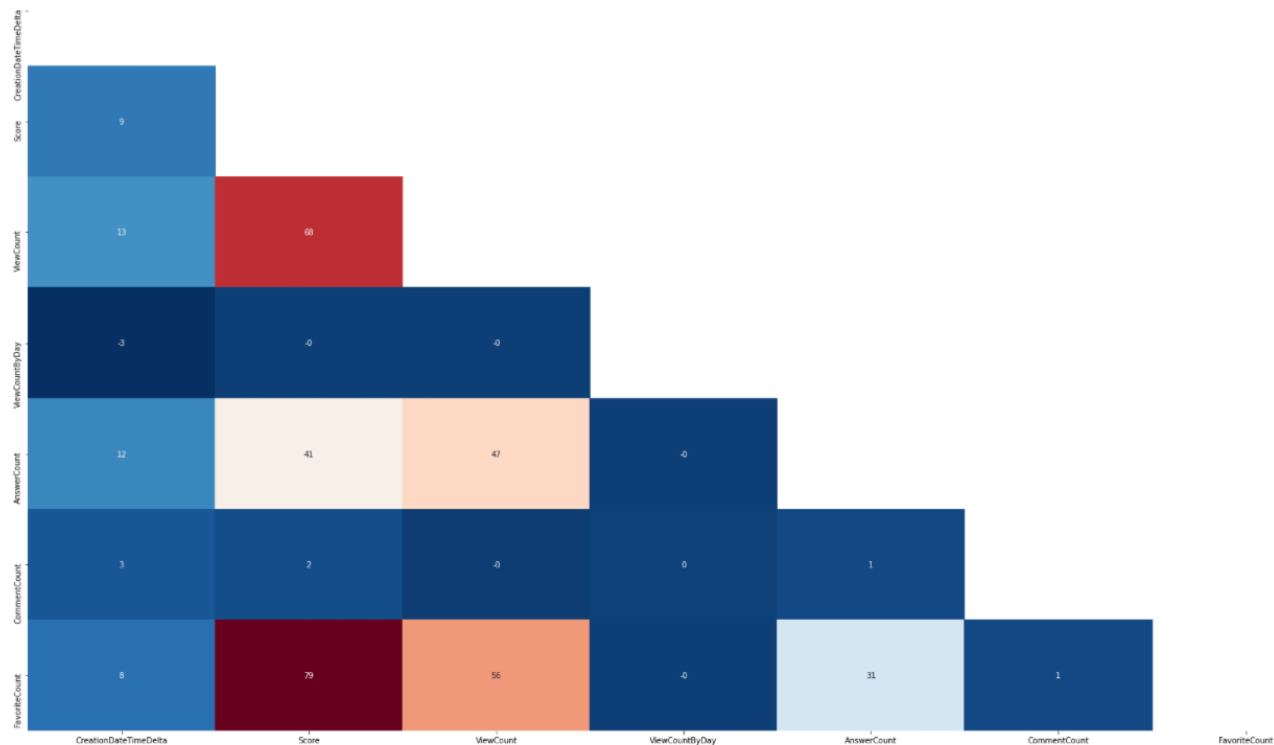
- Pénalisation des documents récents
- Obsolescence des thématiques





CORRELATION DE PEARSON

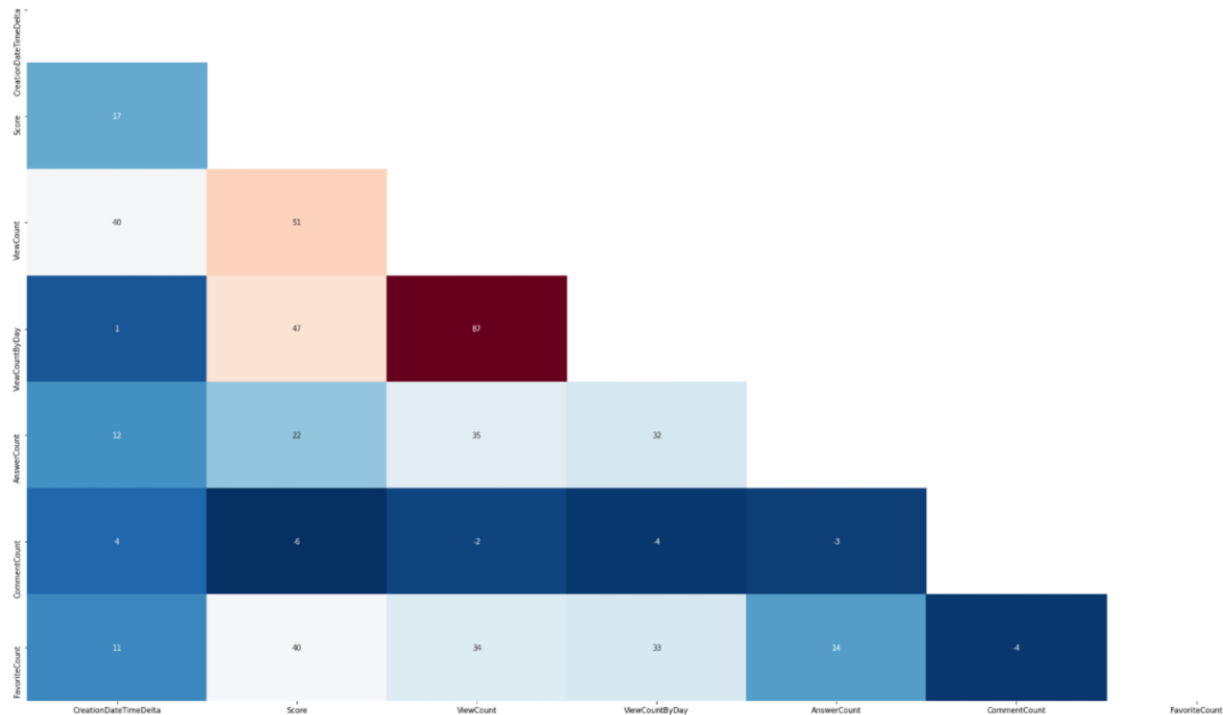
Matrice de corrélations de pearson en %





CORRELATION DE SPEARMAN

Matrice de corrélations de spearman en %



FILTRAGE DES DONNEES

Score	> 0
AnswerCount	> 0
CommentCount	> 0
FavoriteCount	> 0
ViewCountByDay	> 5

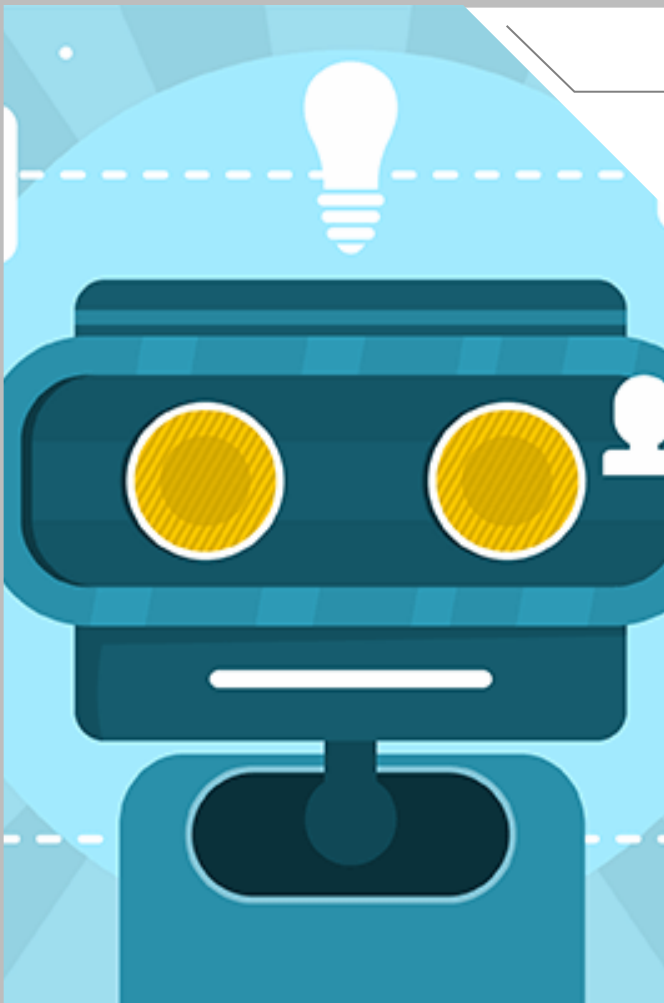
Données restantes	
Avant	Après
902 285	46 504

PRE-TRAITEMENTS DES DOCUMENTS

Traitement	Description
Suppression des balise HTML	Suppression des balises
Nettoyage du texte	<ul style="list-style-type: none">• Passage en minuscules• Filtrage des caractères non alphabétique• Filtrage des termes de moins de 3 caractères
Tokenisation	<ul style="list-style-type: none">• Découpage en tokens• Suppression des stop words
Filtrage à l'aide d'un modèle de POS (Part Of Speech) tagging	Filtrage des noms communs
Racinisation des tokens	Lemmatisation
Filtrage des valeurs vides	Suppression des documents au contenu vide après pré-traitement
Vectorisation du corpus	TF-IDF

Données restantes

Avant	Après
46 504	45 902



[03]

ENTRAINEMENT DES MODELES





APPROCHE SUPERVISEE

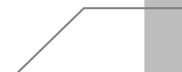


PRE-TRAITEMENTS SPECIFIQUE



Traitement	Description
Dédoublonnage des tags - labels	Nettoyage des déclinaisons de tags
Réduction des dimensions des prédicteurs	Réduction par ACP
Vectorisation des tags - labels	<ul style="list-style-type: none">• Découpage en tokens• Suppression des stop words

Données restantes	
Avant	Après
45902	43132



RESULTATS

	Micro precision	Micro recall	Micro F1 Score	Temps d'entraînement
KNN	0.687998	0.281438	0.399467	38.1 secondes
SVM	0.798641	0.342438	0.479345	22.5 secondes
Random Forest	0.868575	0.146893	0.251289	2 minutes 59 secondes
Gradient Boosting	0.505483	0.307369	0.382283	14 heures 49 minutes et 39 secondes



Modèle retenu : SVM

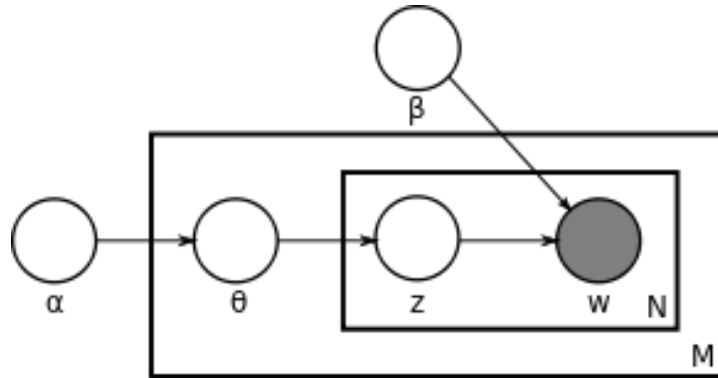


APPROCHE NON SUPERVISEE





LE MODELE LDA



α : Ensemble de tous les topics

β : Ensemble des mots de tous les documents

M : Ensemble des variables liée à un document

θ : Distribution d'un topic pour un document

N : Ensemble des variables liées à un mot

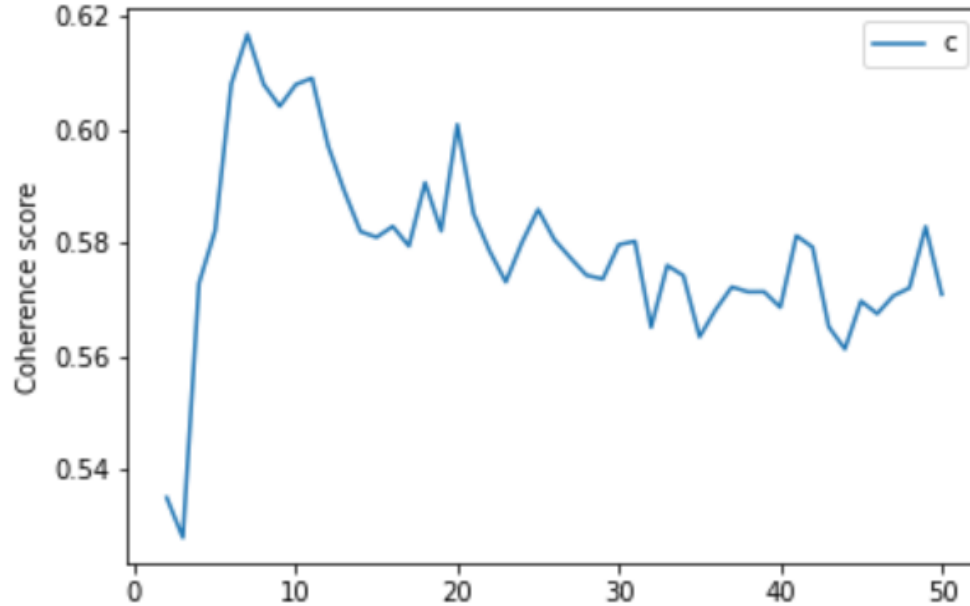
z : Distribution d'un topic pour un mot

w : Mot





SCORE DE COHERENCE





DISTRIBUTION DES TOPICS

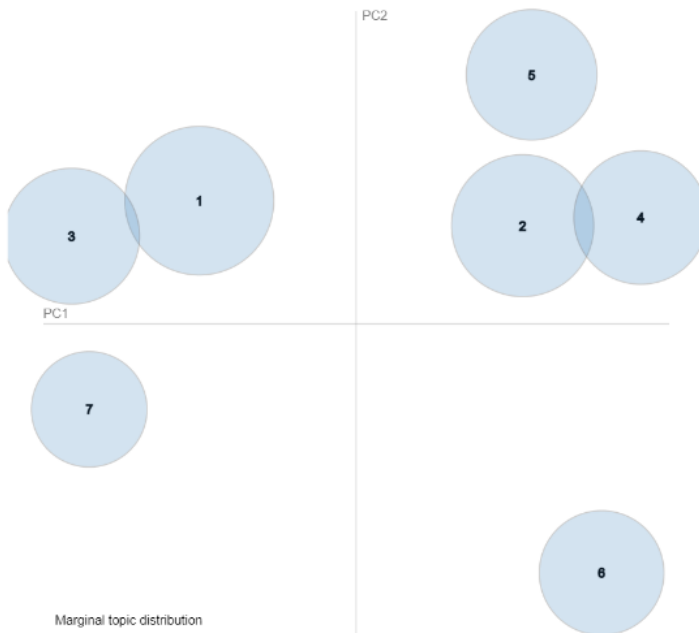
Numéro du topic	Top 10 mots associés	Nombre de document associés	Pourcentage de document associés
1	docker, client, http, access, server, password, service, request, connection, error	7186	15,66%
2	component, button, react, class, import, state, item, const, style, page	6361	13,86%
3	java, android, spring, version, class, boot, system, google, support, device	5260	11,46%
4	project, node, module, package, studio, version, error, json, file, core	6526	14,22%
5	python, file, line, import, model, command, site, install, error, print	6649	14,49%
6	Image, flutter, view, list, color, widget, text, child, context, screen	4462	9,72%
7	value, function, column, array, type, name, number, form, return, date	9458	20,60%





DISTRIBUTION SPATIALE DES TOPICS

Intertopic Distance Map (via multidimensional scaling)




lda_tfidf.html

Marginal topic distribution





COMPARAISON DES APPROCHES





COMPARAISON DES RESULTATS



Tags originaux	Prédiction supervisée (SVM)	Prédiction non supervisée (LDA)
Javascript, ecmaScript	javascript, code	code
xcode, macos, command, line, terminal	- (aucun tag retourné)	line, command, error, path
java, performance, benchmarking, bytecode	java	java, version, sytem, void, string
Info, plist, xcode	code	error, build, code
java (4 déclinaisons), javac	array, code, java	java, version, class, system, void, string
javascript	react, github	project, react
android, intellij, idea, kotlin, corda	github	error
angular (4 déclinaisons)	Juery, core, chrome, html, router, bootstrap, node, typescript, angular, google, eslint, github, json	project, node, package, version, error, json, core, config, build, index
javascript, type, conversion	javascript, code	Result, code



COMPARAISON DES APPROCHES



AVANTAGES

INCONVENIENTS

Approche supervisée

Des modèles et indicateurs connus

Besoin de pré-traitements supplémentaires

Plus d'éléments à maintenir

Approche non supervisée

Appréhension simplifiée

Pas de pré-traitements supplémentaires

Un modèle unique à maintenir

Difficile d'évaluer les performances du modèle





04

API



STACK TECHNIQUE



DEMONSTRATION DE L'INTERFACE



API for tags prediction on Stack Overflow posts 0.0.0 GA3

@param api json

Return tags related to a poste

default

Call / Root

POST /predict Get Prediction

Parameters Try it out

No parameters

Request body required application/json

Example Value | Schema

```
{
  "text": "string"
}
```

Responses

Code	Description	Links
200	Successful Response	No links
	Media type: application/json Content Accept Header:	
	Example Value Schema	
	<pre>"string"</pre>	
422	Validation Error	No links
	Media type: application/json	
	Example Value Schema	
	<pre>{ "detail": { "loc": ["string", "msg": "string", "type": "string"] } }</pre>	



MERCI!

Avez vous des questions?

cedricsoares@me.com
06 09 25 47 45



CREDITS: This presentation template was created
by **Slidesgo**, including icons by **Flaticon**, and
infographics & images by **Freepik**

