



Catégorisation automatique des questions



FABRICE DEPREZ- 07/2023

FORMATION INGENIEUR MACHINE LEARNING

PLAN

01 CONTEXTE

Présentation de stack
Overflow et du besoin

02 TRAITEMENT

Filtrage des données et pré-
traitement

03 MODELES

Elaboration des modèles

Supervisés et non supervisés

04 API

Présentation de l'API

Comparaison des modèles

Méthodologie et déploiement

Conclusion

01 Contexte



La plateforme incontournable pour les questions et les réponses en informatique !

Avec une vaste communauté de développeurs passionnés, Stack Overflow est le lieu idéal pour obtenir des réponses rapides et fiables à vos problèmes de programmation !



Comment faciliter la recherche et améliorer l'efficacité des utilisateurs ?

01 CONTEXT - besoin

"Optimisez son expérience sur Stack Overflow grâce à un algorithme de machine learning de pointe qui attribue automatiquement les tags pertinents à chaque question, permettant de trouver les solutions rapidement et efficacement."



"Faciliter les recherches sur Stack Overflow ! Présenter un système de suggestion de tags alimenté par l'intelligence artificielle pour obtenir rapidement les réponses dont on a besoin."

02 TRAITEMENT - récupération

StackExchange Data Explorer

Home Queries Users Compose Query

Editing Query

Enter a title for your query

edit description

```
1 SELECT TOP 50000 Title, Body, Tags,
2       Id, Score, ViewCount,
3       FavoriteCount, AnswerCount
4 FROM Posts
5 WHERE CreationDate BETWEEN CONVERT(datetime, '2023-01-01') AND CONVERT(datetime, '2023-01-01')
6 AND ViewCount > 10
7 AND FavoriteCount IS NOT NULL
8 AND Score > 10
9 AND AnswerCount > 0
10 AND LEN(Tags) - LEN(REPLACE(Tags, '<', '')) >= 5
```

Database Schema

Posts	
Id	int
PostTypeId	tinyint
AcceptedAnswerId	int
ParentId	int
CreationDate	datetime
DeletionDate	datetime
Score	int
ViewCount	int
Body	nvarchar (max)

Revisions

2132917	anonymous	may 27 at 7:12
---------	-----------	----------------

hide sidebar >>

Run Query Cancel Options: ☐ Text-only results ☐ Include execution plan

StackExchange dataExplorer est un outil puissant
mis à disposition par Stack Overflow pour
extraire les données du site via des requêtes SQL



02 TRAITEMENT - données

Extraire un jeu de données pertinent et conséquent

```
SELECT TOP 50000 Title,  
                  Body,  
                  Tags,  
                  Id,  
                  Score,  
                  ViewCount,  
                  FavoriteCount,  
                  AnswerCount  
FROM Posts  
WHERE ViewCount > 10  
      --AND FavoriteCount > 10  
AND Score > 10  
AND AnswerCount > 0  
AND LEN(Tags) - LEN(REPLACE(Tags, '<', '')) >= 5
```

500 000 enregistrements
Périodicité 2008 - Now

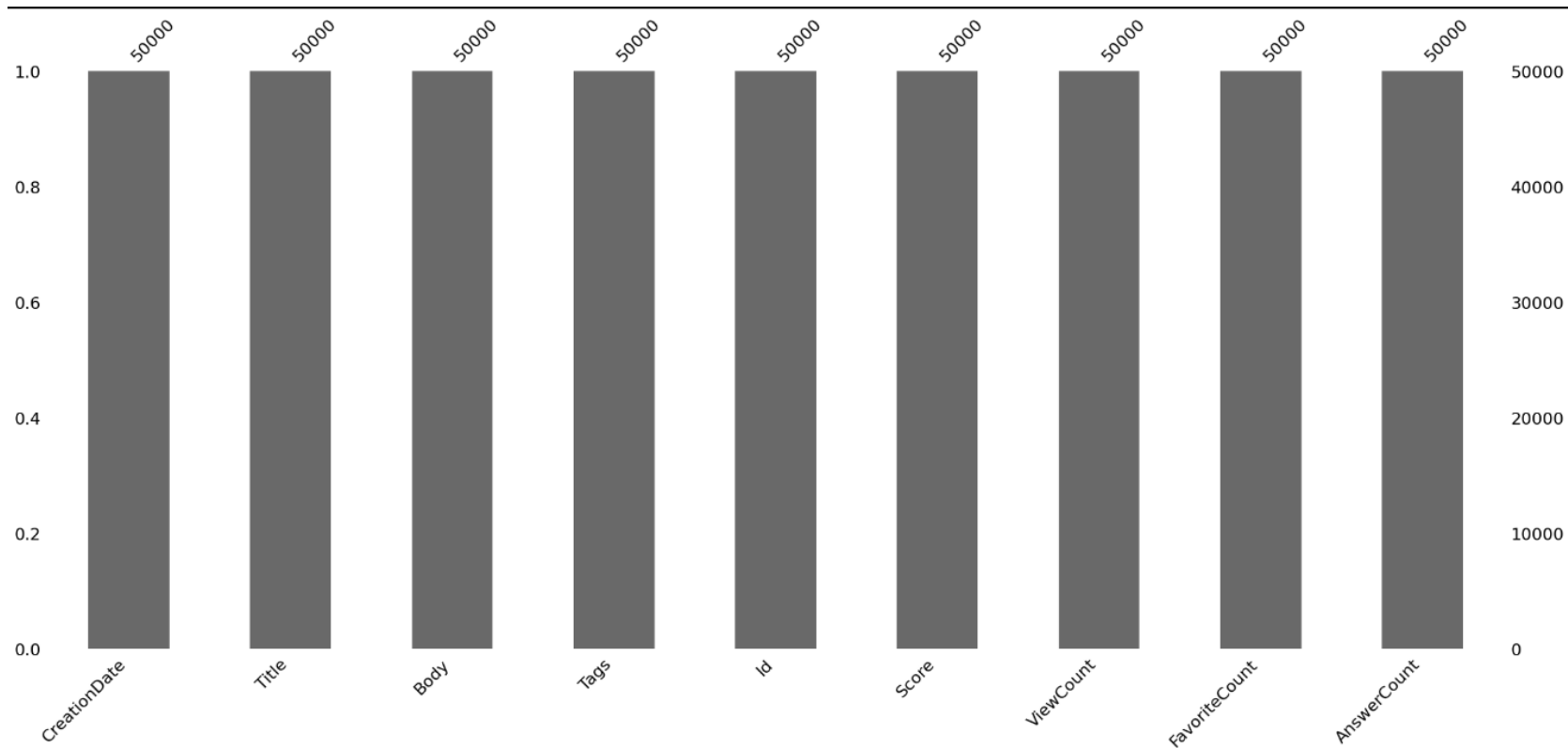
Au moins 5 tags

Score honorable

Visibilités et favoris

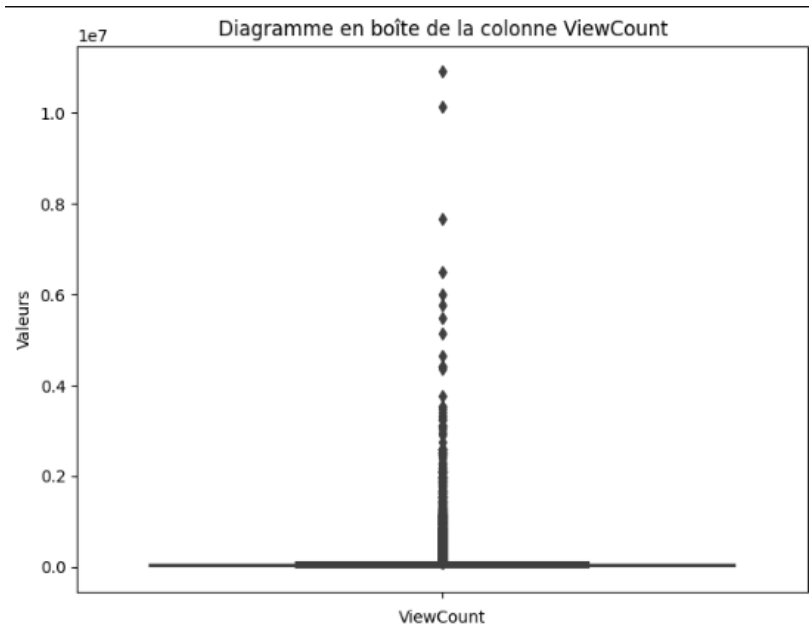
02 TRAITEMENT - données

Un ensemble de données homogène et non vide, facilitant la phase de nettoyage

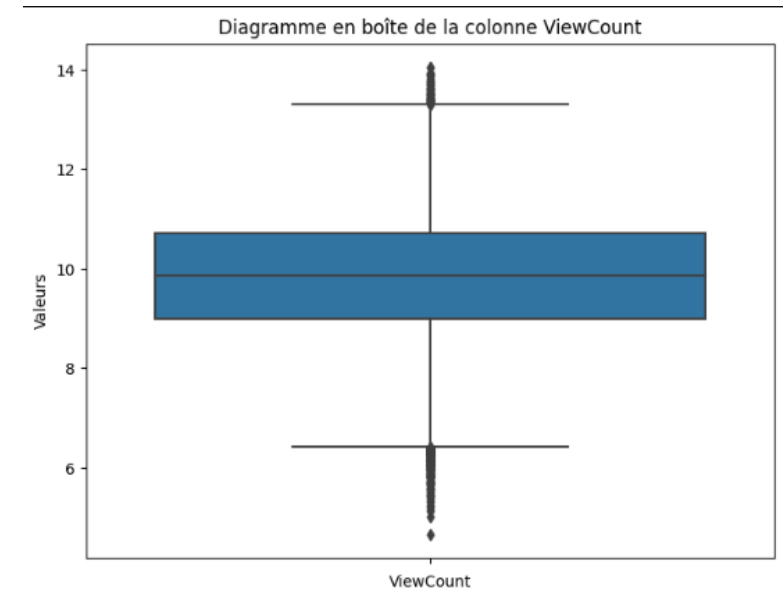


02 TRAITEMENT - Netoyage

Identification des outliers

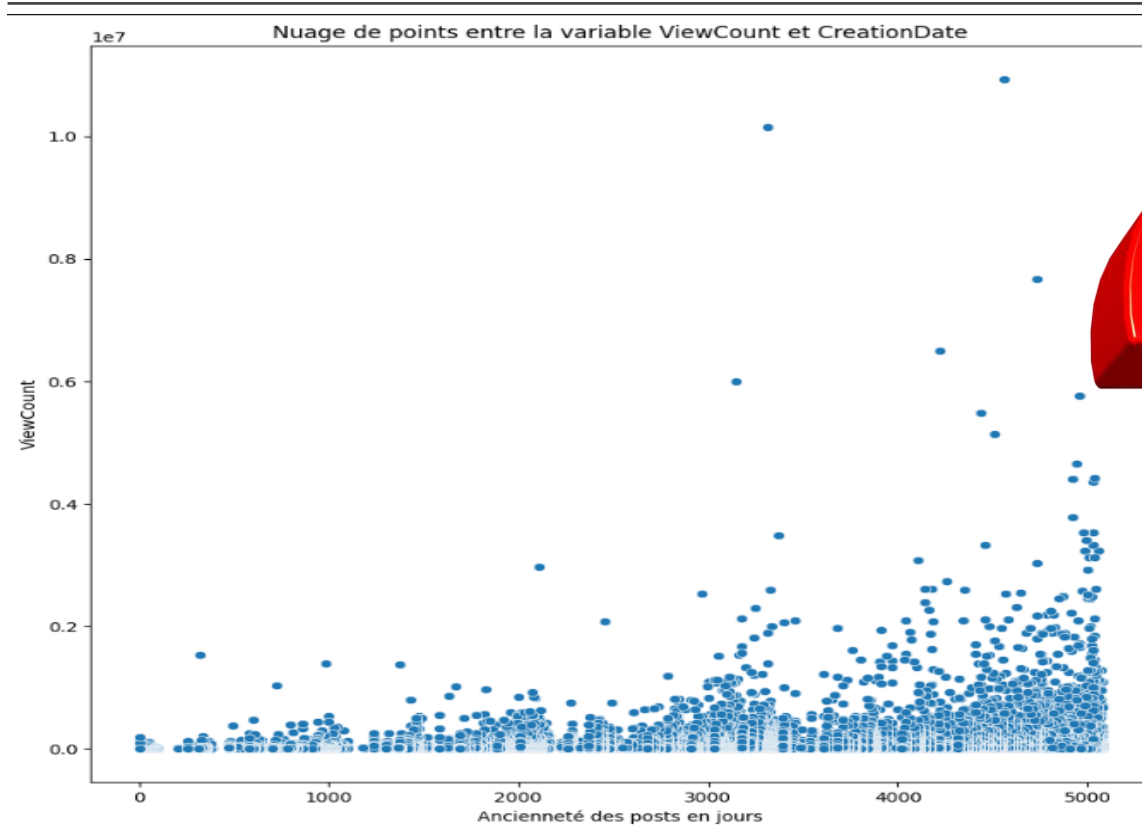


Transformation logarithmique



02 TRAITEMENT - données

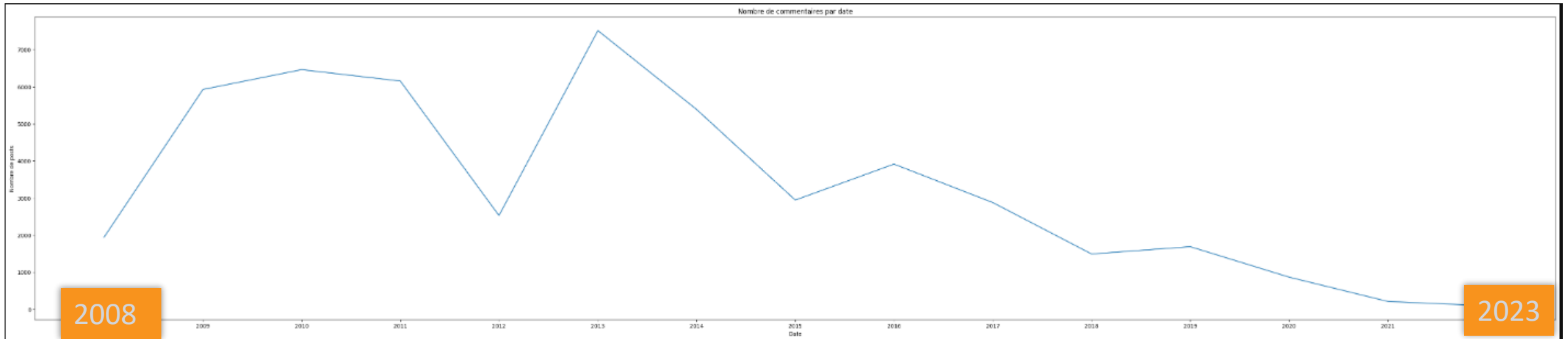
Analyse de l'ancienneté des posts de 2008 à 2022



Une baisse de la fréquentation et des posts sur les dernières années peut-elle potentiellement fausser l'analyse ?

02 TRAITEMENT - Filtrage

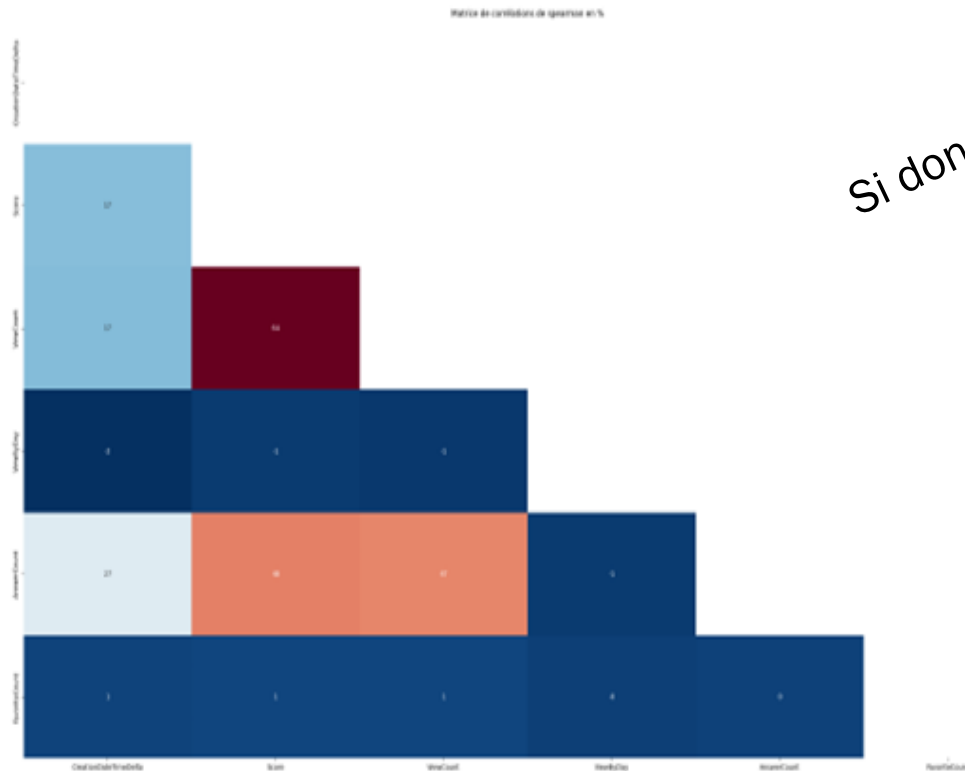
Analyse de l'ancienneté des posts de 2008 à 2022



Graphe plus parlant

02 TRAITEMENT - Corrélations

Pearson et Spearman



Si données linéaires

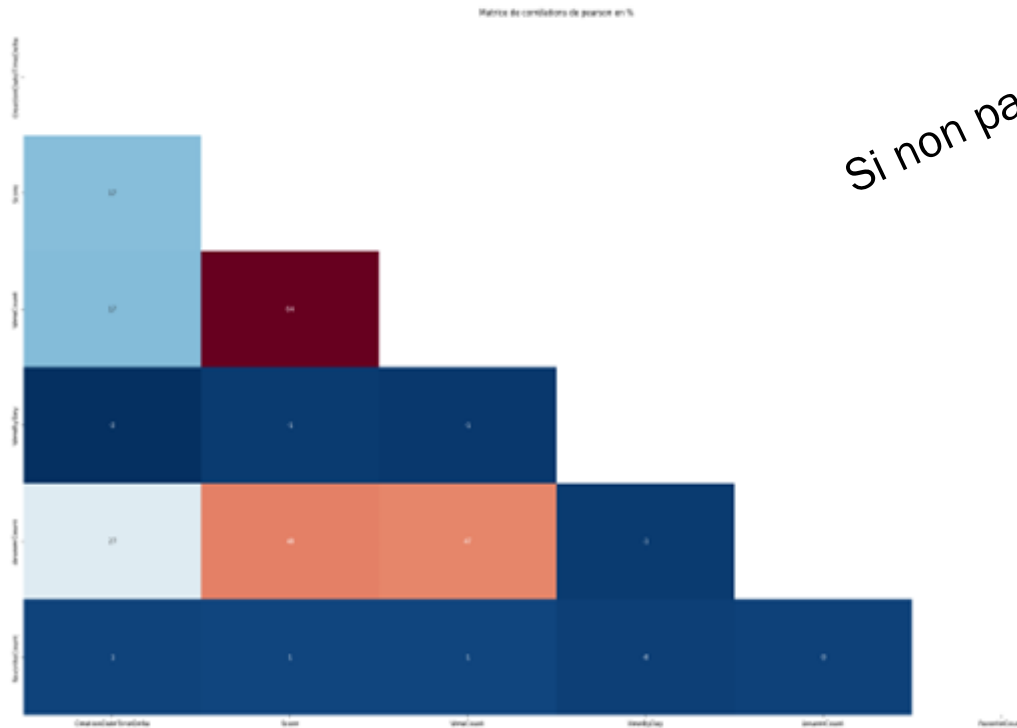
Pearson (*100)

Nb vues et Score	:	64
Score et Réponses	:	48
Réponses et Nb Vues	:	47
Réponses et Ancienneté	:	27

Nb vues/jour et reste	:	-1
Favoris et rest	:	-1

02 TRAITEMENT - Corrélations

Pearson et Spearman



Si non paramétriques

Spearman (*100)

Nb vues et Score	:	64
Score et Réponses	:	48
Réponses et Nb Vues	:	47
Réponses et Ancienneté	:	27

Nb vues/jour et reste	:	-1
Favoris et rest	:	-1

03 MODELES

récupération

Chargement des données pré-traitées

```
: # on ne va garder que les colonnes qui nous intéressent (id title body et tags)
data = pd.read_csv('./data/cleaned_data.csv', usecols=['Id','Title', 'Body', 'Tags'], index_col='Id')
data.reset_index(inplace=True)
data.drop(columns='Id', inplace=True)
data.head(10)
```

	Title	Body	Tags
0	Integer value comparison	<p>I'm a newbie Java coder and I just read a v...	<java><integer><int><equals><autoboxing>
1	How do you handle deploying rails applications...	<p>I recently turned a couple of my plugins in...	<git><plugins><capistrano><deployment><git-sub...
2	Golang converting from rune to string	<p>I have the following code, it is supposed t...	<string><parsing><go><unicode><rune>
3	Java substring: 'String index out of range'	<p>I guess I'm getting this error because the ...	<java><string><substring><indexoutofboundsexce...
4	Number of threads used by Go runtime	<p>How many threads can the Go runtime (schedu...	<multithreading><go><docker><concurrency><goro...
5	C# List Comprehensions = Pure Syntactic Sugar?	<p>Consider the following C# code:</p>\n\n<pre>	<c#><linq><optimization><compiler-construction...
6	Error occurred while decoding OAEP padding	<p>While decrypting text using <code>RSACrypto...	<c#><encryption><rsa><digital-signature><rsacr...
7	Is it possible to share an enum declaration be...	<p>Is there a way to share an enum definition ...	<c#><c++><enums><native><managed>
8	Powermock (With EasyMock) no last call on a mo...	<p>I am trying to just run a simple test case...	<java><unit-testing><junit><easymock><powermock>
9	Breakpoint for "Warning: Attempt to present * ...	<p>Sometimes it happens that - from different ...	<ios><objective-c><swift><uiviewcontroller><br...

03 MODELES

Pre-traitement

CORPUS

```
data['Post'] = data.apply(lambda x: x['Title'] + ' ' + x['Body'] if x['Title'] != x['Title'] else x['Body'], axis=1)
corpus      = data['Post'].to_list()
tags        = data['Tags'].to_list()
```

Occurences dans le corpus: 48926
Occurences dans les tags : 48926

Nettoyage du texte

- Elimination des balises HTML
- Suppression de la ponctuation
- Conversion en minuscules
- Remplacement des termes spécifiques
- Suppression des mots vides (stop words)
- Lemmatisation

03 MODELES

Lemmatization ou Stemming ?



Original	Stemming	Lemmatization
New	New	New
York	York	York
is	is	be
the	the	the
most	most	most
densely	dens	densely
populated	popul	populated
city	citi	city
in	in	in
the	the	the
United	Unite	United
States	State	States

Stemming (Racinisation) :

- Processus rapide et simple, souvent basé sur des règles heuristiques.
- Réduit les mots à leur racine ou forme de base (stem), qui n'est pas nécessairement un mot réel dans la langue.
- Par exemple, les mots "running", "runs" et "runner" peuvent être réduits à la racine "run".

Lemmatization (Lemmatisation) :

- Processus plus complexe et plus lent qui prend en compte le contexte linguistique.
- Réduit les mots à leur lemme, qui est leur forme de base selon le dictionnaire.
- Utilise l'information morphologique pour trouver la forme canonique des mots.
- Par exemple, "better" est transformé en "good", "am/are/is" devient "be".

03 MODELES

Tokenization du corpus et des tags



```
['integer',  
'value',  
'comparison',  
'newbie',  
'java',  
'coder',  
'read',  
'variable',  
'integer',  
'class',  
'described',  
'three',  
'different',  
'vay',  
'i',  
'flowing',  
'eto',
```

1.Tokenisation : Le processus de division des textes en petites parties appelées "tokens".

1. Utilisé pour décomposer le texte en unités plus petites pour une meilleure compréhension du contexte.

2.Application : La tokenisation a été appliquée à notre corpus de textes et à nos tags.

3.Résultat :

1. Nous obtenons des listes de tokens pour chaque texte et chaque ensemble de tags.
2. Chaque token représente un mot ou un élément de texte/tag distinct.

4.Pourquoi faire cela ?

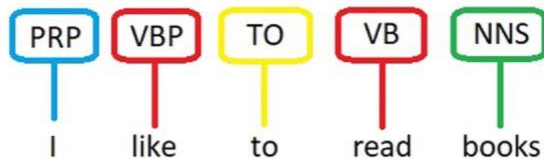
1. La tokenisation est une étape clé dans la préparation des données pour de nombreux modèles de traitement du langage naturel.
2. Elle permet de travailler avec des unités de texte plus petites qui ont un sens en soi.

```
'cable',  
'goal',  
'figure',  
'see',  
'value',
```


03 MODELES

Filtrage par POS Tagging du corpus

POS Tagging



1.Part-of-Speech (Pos) Tagging : Processus d'attribution de balises grammaticales aux mots d'un texte.

- Les balises peuvent indiquer des noms, des verbes, des adjectifs, des adverbes, etc.

2.Application : Nous avons utilisé le Pos Tagging pour filtrer et ne garder que les noms dans notre corpus.

- Les noms sont souvent très informatifs dans le contexte de l'analyse de texte.

3.Résultat :

- Nous obtenons un corpus où chaque texte est une liste de noms.
- Cela simplifie et focalise notre analyse sur les entités les plus pertinentes.

4.Pourquoi faire cela ?

- Le filtrage par Pos Tagging permet de réduire la taille du corpus et de se concentrer sur les aspects les plus pertinents du texte.
- Cela est particulièrement utile dans des domaines tels que l'analyse des sentiments et l'extraction d'entités.

03 MODELES

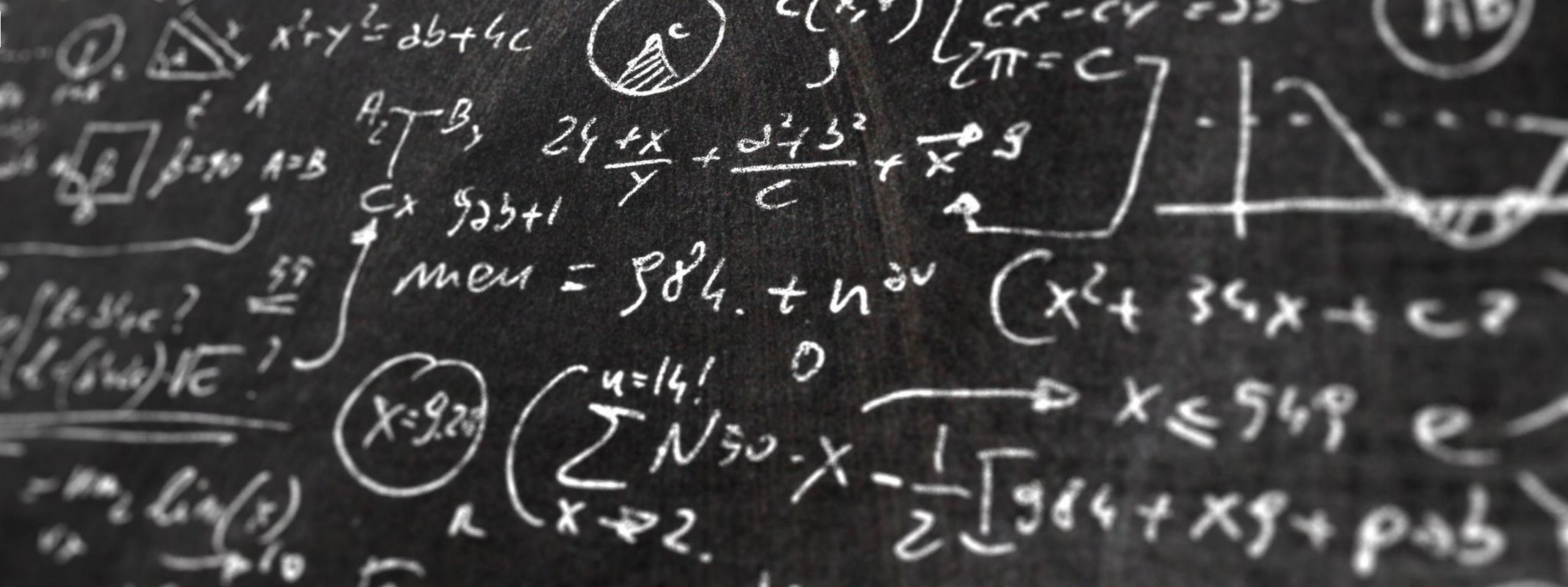
Fréquence de distribution

Nombre de tokens du corpus 17793
Affichage des 10 tokens les plus utilisés

Frequency	
Word	
system	3273
class	2536
code	2505
file	1935
application	1896
value	1824
type	1741
use	1738
way	1565
name	1560
project	1447
work	1428
int	1408
c	1403
version	1394
error	1370
method	1359
return	1288
window	1206
test	1169

Top 10 des tags les plus utilisés

Frequency	
Word	
net	3976
core	222
bit	128
64	100
3d	75
htaccess	69
32bit	35
64bit	35
2d	30
standard	24
assembly	24
32	22
framework	11
cross	11
internet	10
explorer	10
google	10
embed	9
domain	9
youtube	8



03 MODELES

Méthode supervisée

03 MODELES

T-IDF ou Bag of Words ?



Pour pouvoir entraîner nos modèles, nous devons transformer les listes de tokens lemmatisés en vecteurs. Deux méthodes couramment utilisées sont les suivantes :

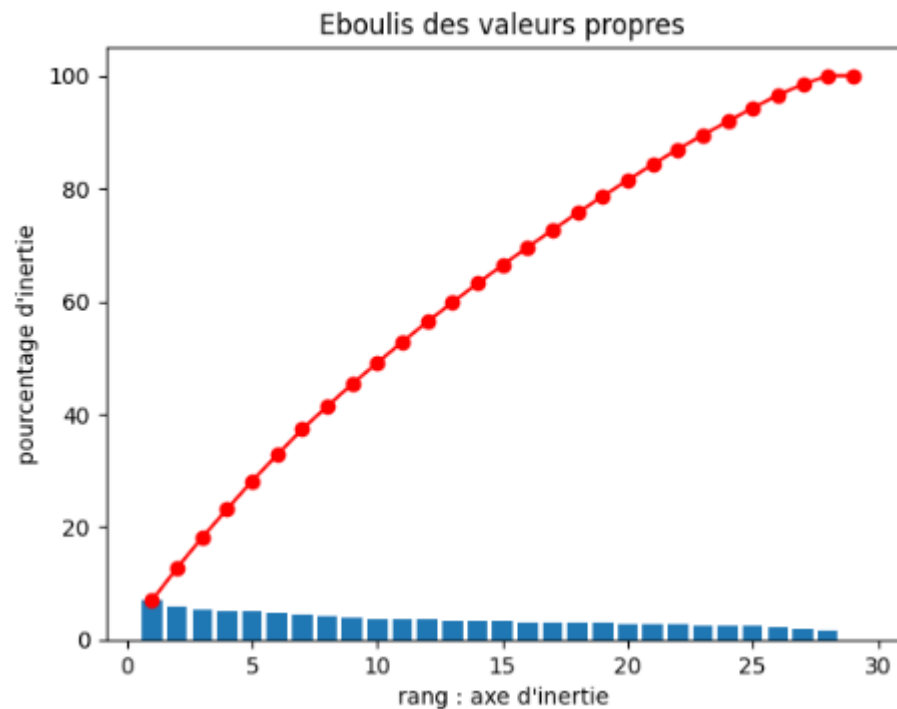
- Bag of Words (BoW) : Chaque liste de tokens (appelée document) est convertie en un vecteur indiquant la fréquence brute de chaque terme du corpus dans le document.
- TF-IDF : Cette méthode remplace la fréquence brute d'un token par un indicateur composé de sa fréquence d'apparition dans le document et de la fréquence inverse du nombre de documents où le token apparaît. Cette méthode permet de minimiser l'importance des tokens présents dans un grand nombre de documents et de normaliser la taille des documents.

Termes génériques
Longueurs variables

03 MODELES

Réduction / ACP

```
X_train, X_test, y_train, y_test = train_test_split(tfidf_data, dedoub_tags, test_size=0.2, random_state=42)
```

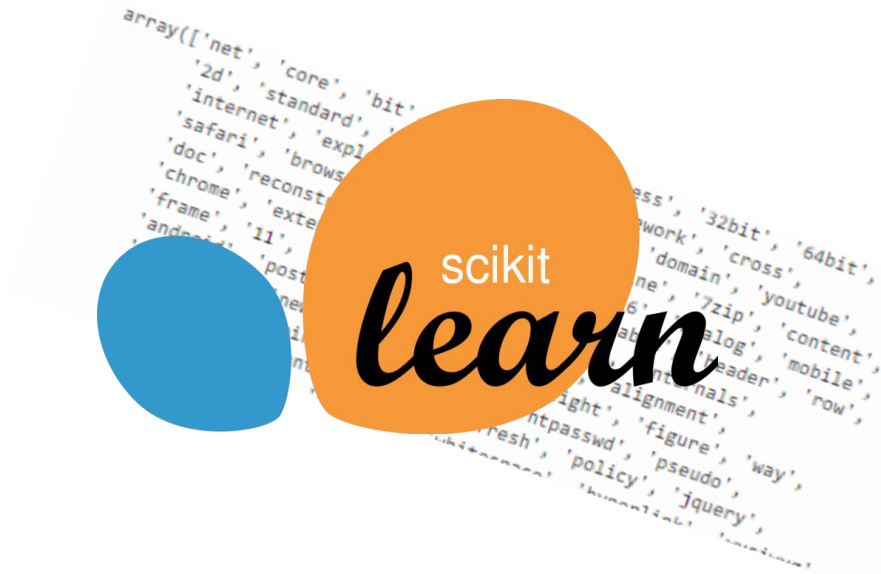


```
pca = PCA(n_components=0.85, random_state=42)
pca.fit(X_train)
X_train_transformed = pca.transform(X_train)
X_test_transformed = pca.transform(X_test)
print(f"Nombre de composantes principales: {pca.components_.shape[0]}")

filename_pca_model = './models/pca_model.pkl'
pickle.dump(pca, open(filename_pca_model, 'wb'))
```

Nombre de composantes principales: 22

03 MODELES



1.Vectorisation : Processus de transformation des étiquettes textuelles en un format numérique que le modèle peut comprendre et manipuler.

2.MultiLabelBinarizer : Un outil de scikit-learn utilisé pour transformer les étiquettes multivaluées en un format binaire.

3.Application : Nous avons utilisé le MultiLabelBinarizer pour transformer nos étiquettes d'entraînement et de test en un format numérique.

1. Le modèle a été ajusté sur l'ensemble d'entraînement et utilisé pour transformer les deux ensembles.

4. Résultat :

1. Nous obtenons des ensembles d'étiquettes d'entraînement et de test en format binaire.
2. Ces formats sont maintenant prêts à être utilisés par un modèle de machine learning.

5. Pourquoi faire cela ?

1. La vectorisation est nécessaire car les modèles de machine learning ne peuvent pas traiter directement les données textuelles.
2. Elle permet également de manipuler et comparer facilement les étiquettes.

03 MODELES

Comparaison : KNN, SVM, RF, GB

	micro_precision	micro_recall	micro_f1
knn	0.904403	0.753669	0.822184
svm	0.907975	0.775681	0.836631
Random Forest	0.889429	0.767296	0.823860
Gradient Boosting	0.781548	0.772537	0.777016

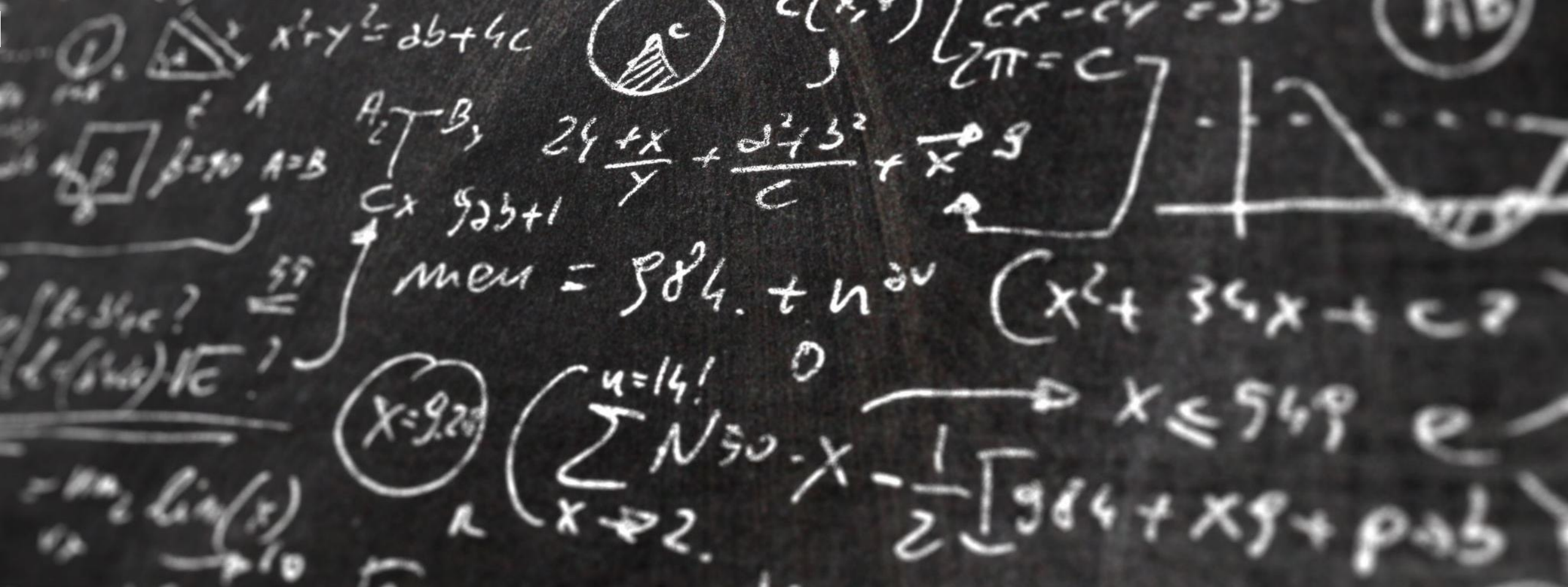
- **knn** : Ce modèle a une précision de 90.4%, un rappel de 75.3% et un score F1 de 82.2%. Cela signifie qu'il est assez précis, mais a du mal à identifier toutes les instances positives (comme le montre le rappel).

- **svm** : Ce modèle a une précision de 90.8%, un rappel de 77.6% et un score F1 de 83.7%. Ce modèle est légèrement plus performant que le knn en termes de rappel et de score F1.

- **Random Forest** : Ce modèle a une précision de 88.9%, un rappel de 76.7% et un score F1 de 82.4%. Il a un rappel légèrement plus élevé que le knn, mais une précision inférieure.

- **Gradient Boosting** : Ce modèle a une précision de 78.1%, un rappel de 77.3% et un score F1 de 77.7%. Il est le moins précis des quatre, mais son rappel est comparable à celui des autres.





03 MODELES

Méthode non supervisée

03 MODELES

LDA ou NMF ?



Latent Dirichlet Allocation (LDA) :

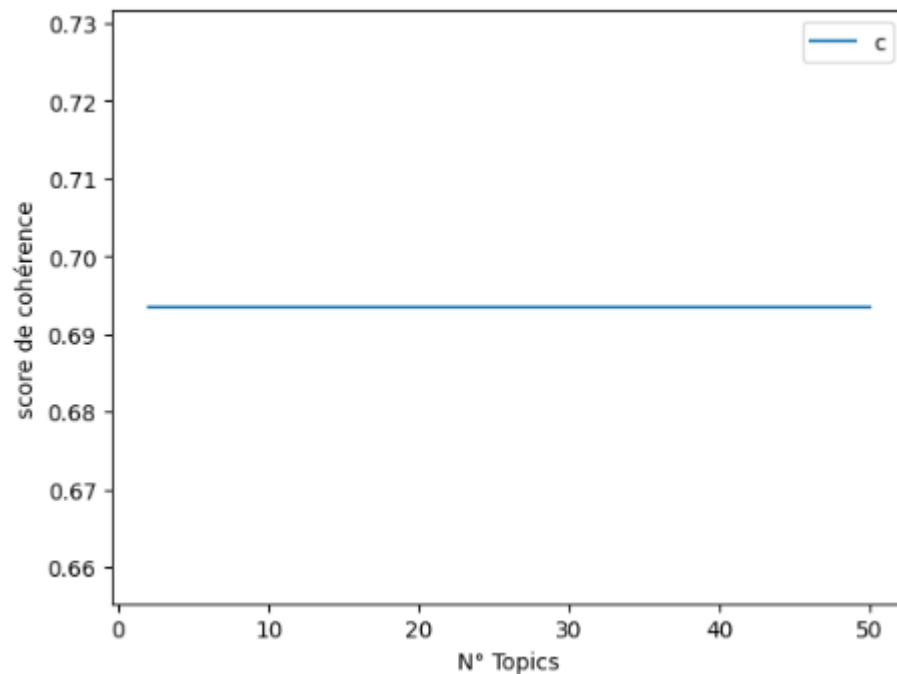
- Technique de modélisation de thèmes
- Chaque document est une combinaison de sujets
- Chaque sujet est une distribution de mots
- Non supervisée (pas de labels requis)

Factorisation Matricielle Non-négative (NMF) :

- Méthode de décomposition de matrices
- Utile pour la réduction de dimensionnalité
- Chaque document est une combinaison de sujets
- Chaque sujet est une combinaison de mots
- Non supervisée (pas de labels requis)

03 MODELES

LDA : 0,6934



Cohérence constante à travers différentes valeurs pour le nombre de sujet !

- 1.Profondeur du corpus:** Si les sujets sont très similaires les uns aux autres, ou si chaque document contient un mélange de nombreux sujets différents, il se peut que le score de cohérence soit constant parce que le modèle LDA ne parvient pas à trouver des distinctions claires entre différents ensembles de sujets.
- 2.Prétraitement:** un problème peut-être par exemple : le nettoyage du texte, la suppression des mots vides, la lemmatisation), il faut vérifier si cela a été fait correctement. S'il y a beaucoup de bruit ou d'irrégularités dans le corpus, cela pourrait également influencer la cohérence du modèle.
- 3.Paramètres du modèle:** Les paramètres utilisés pour notre modèle LDA, comme le nombre de passes, peuvent également affecter la cohérence. Il faut essayer d'ajuster certains de ces paramètres pour voir si cela a un impact.
- 4.Métrique de cohérence:** Il convient également de noter que la cohérence du sujet n'est pas la seule mesure de la qualité d'un modèle de sujet. Il est possible que le modèle LDA fonctionne bien pour notre tâche spécifique, même si le score de cohérence est constant.

03 MODELES

Score de cohérence et perplexité

Perplexity: -2.4272944530275447

Coherence Score: 0.6934482532730719

```
[(0, '0.276*"way" + 0.276*"code" + 0.236*"class" + 0.212*"use"'),  
(1, '0.938*"code" + 0.038*"use" + 0.018*"class" + 0.005*"way"'),  
(2, '0.870*"class" + 0.053*"code" + 0.041*"way" + 0.037*"use"'),  
(3, '0.937*"use" + 0.030*"code" + 0.029*"way" + 0.005*"class"'),  
(4, '0.940*"way" + 0.056*"code" + 0.002*"use" + 0.001*"class"'),  
(5, '0.480*"class" + 0.425*"use" + 0.086*"code" + 0.009*"way"'),  
(6, '0.439*"code" + 0.251*"way" + 0.196*"use" + 0.114*"class"')]
```

Perplexité : 2,43 – Cohérence : 0,69

Les mots les plus importants pour chaque sujet semblent être très similaires, ce qui peut suggérer que le modèle a du mal à distinguer clairement différents sujets.

03 MODELES

Attribution des sujets principaux pour chaque document

LDA

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text	Original_keywords
0	0.0	0.0	0.1429	way, code, class, use	How can I configure Entity Framework to automa...	NaN
1	1.0	3.0	0.5710	use, code, way, class	How to perform a binary search on IList<T>? <p...	NaN
2	2.0	0.0	0.1429	way, code, class, use	Anonymous methods and delegates <p>I try to un...	NaN
3	3.0	4.0	0.6255	way, code, use, class	How to Unit Test Asp.net Membership? <p>I am n...	NaN
4	4.0	0.0	0.1429	way, code, class, use	Win32 Console app vs. CLR Console app <p>I'm w...	NaN
5	5.0	0.0	0.1429	way, code, class, use	Why can't I drag execution point in IntelliJ (...	NaN
6	6.0	2.0	0.5942	class, code, way, use	Why C# is not allowing non-member functions li...	NaN
7	7.0	0.0	0.1429	way, code, class, use	I need some clarification on the MVC architect...	NaN
8	8.0	0.0	0.1429	way, code, class, use	mex binding error in WCF <p>I am using VSTS 20...	NaN
9	9.0	3.0	0.5710	use, code, way, class	Regex for Money <p>I have <code>asp:TextBox</c...	NaN

03 MODELES

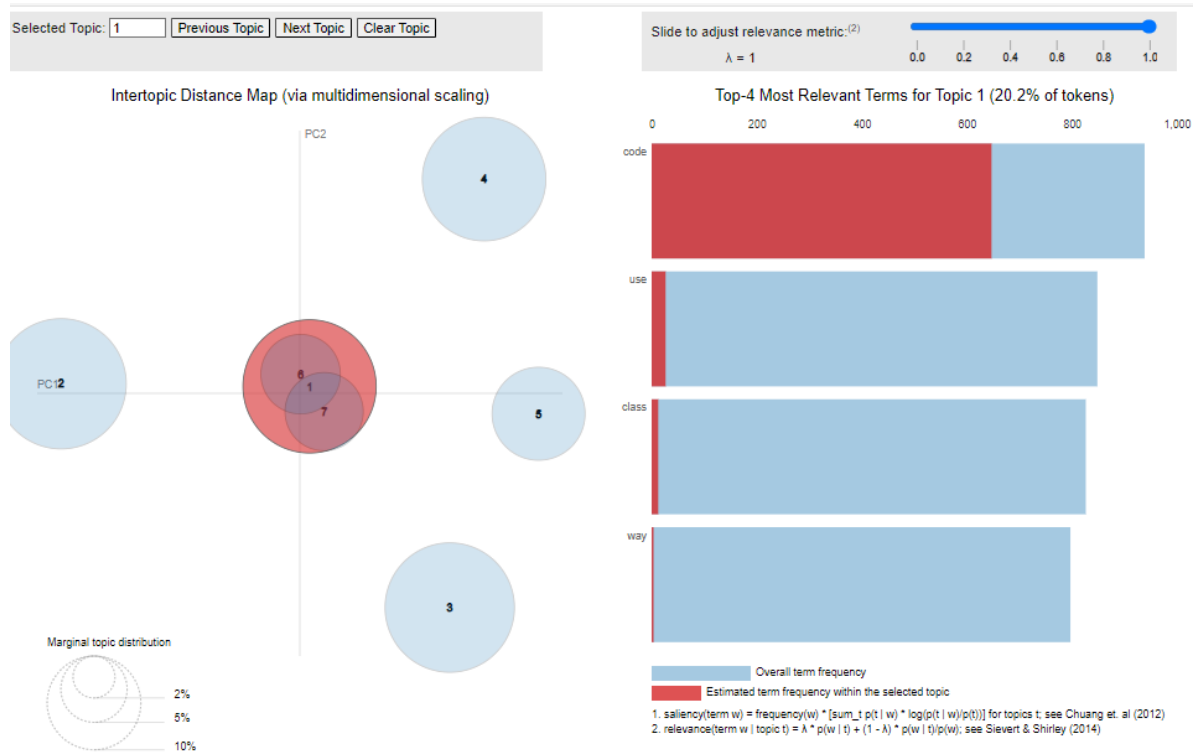
Attribution des sujets principaux pour chaque document

NMF

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text	Original_keywords
0	0.0	0.0	0.0405	code, pre, string, public, new, class, int, me...	How can I configure Entity Framework to automa...	NaN
1	1.0	0.0	0.0805	code, pre, string, public, new, class, int, me...	How to perform a binary search on IList<T>? <p...	NaN
2	2.0	0.0	0.0522	code, pre, string, public, new, class, int, me...	Anonymous methods and delegates <p>I try to un...	NaN
3	3.0	0.0	0.0614	code, pre, string, public, new, class, int, me...	How to Unit Test Asp.net Membership? <p>I am n...	NaN
4	4.0	3.0	0.0799	net, strong, com, href, http, application, mic...	Win32 Console app vs. CLR Console app <p>I'm w...	NaN
5	5.0	3.0	0.0285	net, strong, com, href, http, application, mic...	Why can't I drag execution point in IntelliJ (...)	NaN
6	6.0	0.0	0.0180	code, pre, string, public, new, class, int, me...	Why C# is not allowing non-member functions li...	NaN
7	7.0	1.0	0.1103	li, ul, ol, strong, em, code, href, noreferrer...	I need some clarification on the MVC architect...	NaN
8	8.0	2.0	0.2288	gt, lt, pre, list, public, binding, div, ifram...	mex binding error in WCF <p>I am using VSTS 20...	NaN
9	9.0	0.0	0.0705	code, pre, string, public, new, class, int, me...	Regex for Money <p>I have <code>asp:TextBox</c...	NaN

03 MODELES

Visualisation interactive des sujets par LDA



« Code » = forte association au sujet sélectionné

03 MODELES

Prédiction et Vérification : LDA

Document 7:

Publication originale:

I need some clarification on the MVC architecture and the three-tier architecture <p>I've been reading the book Pro ASP NET MVC Framework and I'm getting really confused with a lot of things. I've been trying to do some research but I'm finding th at with so many different approaches and concepts being thrown at me, it's just making things worse.
So I have a few questions:</p>

```
<ol>
<li><p>I know MVC is supposed to split the functionality into three main things: Model -> Controller -> View. Is the MVC a different approach than the three-tier architecture? Or am I still supposed to be thinking of creating a Data Access Layer a nd a Business Logic Layer in my project?</p></li>
<li><p>What exactly are Repositories? It is what acts as my Data Access Layer? Where/How do Repositories fit into the MVC?</p></li>
<li><p>The book talks about using LINQ to SQL to interact with the database but yet it states that LINQ to SQL will not be supported in the future and that Microsoft is dropping it for the Entity Framework. Where does the Entity Framework fit into the MVC and how do I interact with it?</p></li>
</ol>

<p>Thanks in advance for your help!<br/>
Matt</p>
```

tags pré-traités utilisés par l'utilisateur: ['tier']

prédits par le modèle supervisé: ['net']

tags prédits par le modèle non supervisé: []

03 MODELES

Prédiction et Vérification : NMF

Document 7:

Publication originale:

I need some clarification on the MVC architecture and the three-tier architecture <p>I've been reading the book Pro ASP NET MVC Framework and I'm getting really confused with a lot of things. I've been trying to do some research but I'm finding that with so many different approaches and concepts being thrown at me, it's just making things worse.
So I have a few questions:</p>

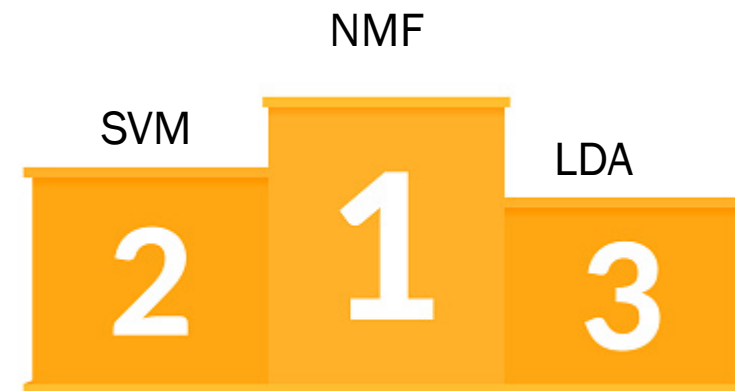
```
<ol>
<li><p>I know MVC is supposed to split the functionality into three main things: Model -> Controller -> View. Is the MVC a different approach than the three-tier architecture? Or am I still supposed to be thinking of creating a Data Access Layer and a Business Logic Layer in my project?</p></li>
<li><p>What exactly are Repositories? It is what acts as my Data Access Layer? Where/How do Repositories fit into the MVC?</p></li>
<li><p>The book talks about using LINQ to SQL to interact with the database but yet it states that LINQ to SQL will not be supported in the future and that Microsoft is dropping it for the Entity Framework. Where does the Entity Framework fit into the MVC and how do I interact with it?</p></li>
</ol>
```

```
<p>Thanks in advance for your help!<br/>
Matt</p>
```

tags pré-traités utilisés par l'utilisateur: ['tier']

prédits par le modèle supervisé: ['net']

tags prédits par le modèle non supervisé: ['net', 'project']




```
opacity:1;*top:-2px;*left:-5px;  
opacity:1\0/;top:-4px\0/;left:-6px\0/;rig  
-moz-inline-box;display:inline-block;fo  
gmoz{display:block;list-style:none;  
play:inline-block;line-height:27px;padd  
cursor:pointer;display:block;text-de  
ion:relative;z-index:1000
```

04 API

04 API

Intérêt ?

Facilité d'intégration

Abstraction

Utilisation du modèle
par des clients

Mise à jour facile

Sécurité



Flask

04 API

Déploiement



Gestion d'un serveur

Evolutivité


Support intégré pour Flask

Peu couteux

Base de données

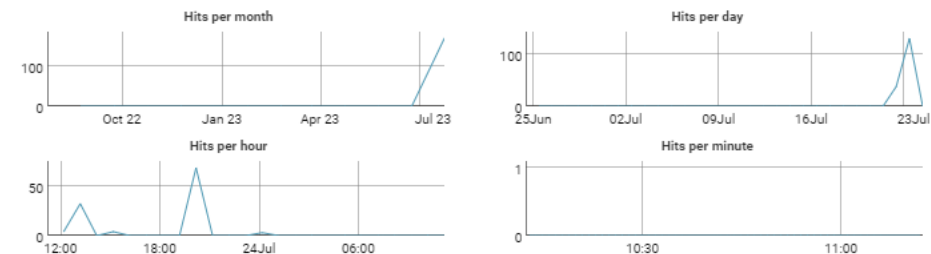
Configuration for kaneda.pythonanywhere.com 

Reload:

 Reload kaneda.pythonanywhere.com


Traffic:

How busy is your site?



Code:

What your site is running.

Source code:	/home/kaneda/api	Go to directory
Working directory:	/home/kaneda/	Go to directory
WSGI configuration file:	/var/www/kaneda_pythonanywhere_com_wsgi.py	
Python version:	3.9 	

04 API

Entry point



The screenshot shows a web browser window with the URL kaneda.pythonanywhere.com/static/index.html. The page features the OpenClassrooms logo and the text "Catégorisez automatiquement des questions". Below this, there are two text input fields. The first field contains the text: "something like this? Or do you think there should not be non-member functions and every method should belong to some class?</p><p>My opinion is to have non-member function support and it helps to avoid polluting class's interface.</p><p>Any thoughts..?</p>". Below the first field is a button labeled "Prédire LDA". The second field contains the text: "<p>My opinion is to have non-member function support and it helps to avoid polluting class's interface.</p><p>Any thoughts..?</p>". Below the second field is a button labeled "Prédire NMF".

Texte :

something like this? Or do you think there should not be non-member functions and every method should belong to some class?</p><p>My opinion is to have non-member function support and it helps to avoid polluting class's interface.</p><p>Any thoughts..?</p>

Prédire LDA

Texte :

<p>My opinion is to have non-member function support and it helps to avoid polluting class's interface.</p><p>Any thoughts..?</p>

Prédire NMF

Résultats de la prédiction LDA :

Tags non supervisés : class,code, Tags supervisés :

Résultats de la prédiction NMF :

Aucune suggestion

<https://kaneda.pythonanywhere.com/static/index.html>



Conclusion

- Le traitement du langage nécessite une **approche mixte** pour des résultats optimaux
- **NMF, SVM et LDA** se sont avérés efficaces, avec le NMF ayant les meilleures performances
- La **qualité des données d'entrée** et les étapes de **prétraitement** sont cruciales pour la performance des modèles
- Les **défis** existent pour interpréter avec précision le langage humain, surtout quand il contient du code comme sur Stack Overflow
- Malgré ces défis, nous avons fait des progrès significatifs grâce aux outils de **machine learning** disponibles aujourd'hui
- En continuant à développer ces outils, nous nous rapprochons d'un système capable de comprendre et d'interpréter le langage humain **presque aussi efficacement d'un humain**



Questions