

# Catégorisation automatique des questions



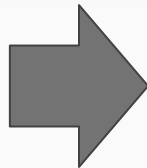
# Sommaire

1. **Contexte**
2. **Projet**
3. **Données et nettoyage**
4. **Exploration**
5. **Modélisations**
6. **API**
7. **Conclusion**



- site web proposant des questions et réponses, thèmes concernant la programmation informatique.
- Il a été lancé le 15 septembre 2008 par Jeff Atwood et Joël Spolsky.
- En août 2015, Stack Overflow revendique plus de 10 000 000 questions.
- L'utilisation de "Tags" ("étiquettes").

**BDD Stack Overflow**



**Questions / Tags**

1. **Nettoyage et pré-traitement des données textuelles.**
2. **Feature engineering**
3. **Modélisations supervisées et non-supervisées.**
4. **Point d'entrée API (web app)**

**Python / Notebook Jupyter / Colab / CloudReady**

# DONNÉES

# StackExchange

```
SELECT id, PostTypeId, Title, Body,  
Tags, CreationDate, Score  
FROM Posts  
WHERE CreationDate >= '2019/01/01'  
AND PostTypeId = 1  
AND Score > 2  
ORDER BY CreationDate
```

Limitation 50000 obs

Multiples requêtes successives  
par dates.

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

**Formatting**  
B I Hide formatting tips

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

**Tags**  
Add up to 5 tags to describe what your question is about  
e.g. (.net ruby json)

Review your question

## Préparation:

- Retrait balises code
- Suppression ponctuation et stopwords
- Tokenisation et Lemmatisation

<p>This is actually for the thread on <a href="https://stackoverflow.com/questions/49162667/unknown-error-call-function-result-missing-value-for-selenium-send-keys-even">unknown error: call function result missing &#39;value&#39; for Selenium Send Keys even after chromedriver upgrade</a> but I guess my reputation isn't high enough to participate(lame).</p>

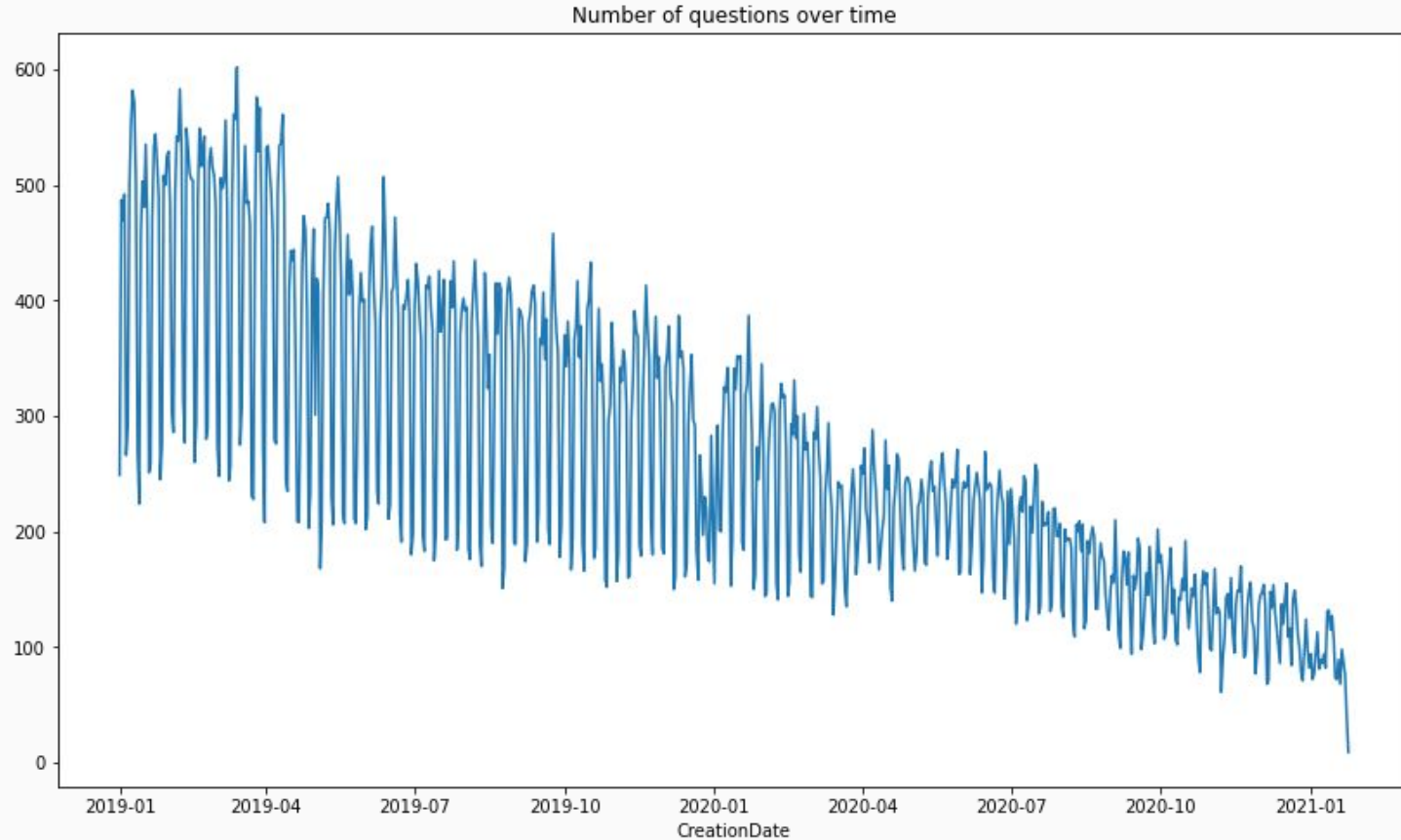


'actually', 'thread', 'unknown', 'error', 'call', 'function', 'result', 'missing', 'value', 'selenium', 'send', 'key', 'even', 'chromedriver', 'upgrade', 'guess', 'reputation', 'high', 'enough', 'participate', 'lame'

# ANALYSE EXPLORATOIRE

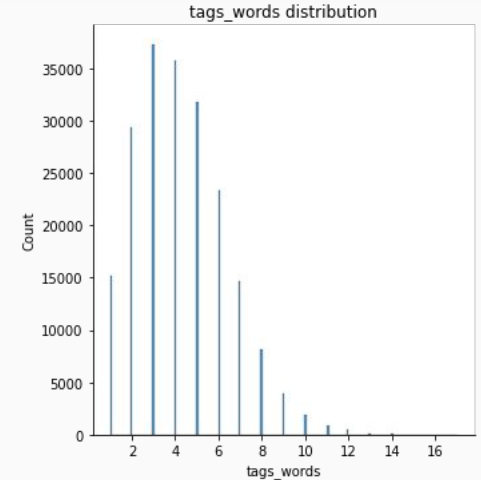
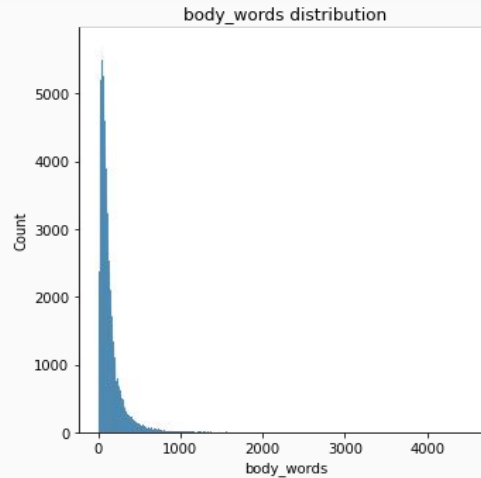
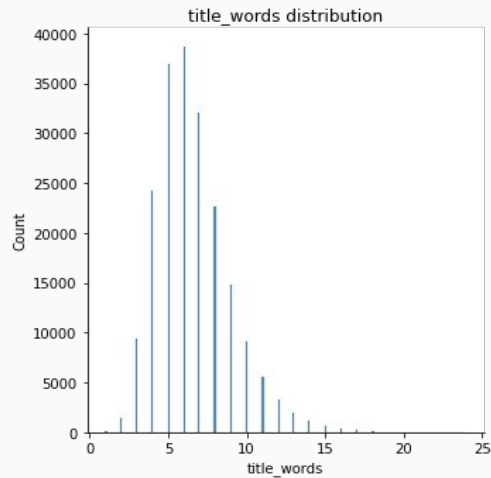


# Questions



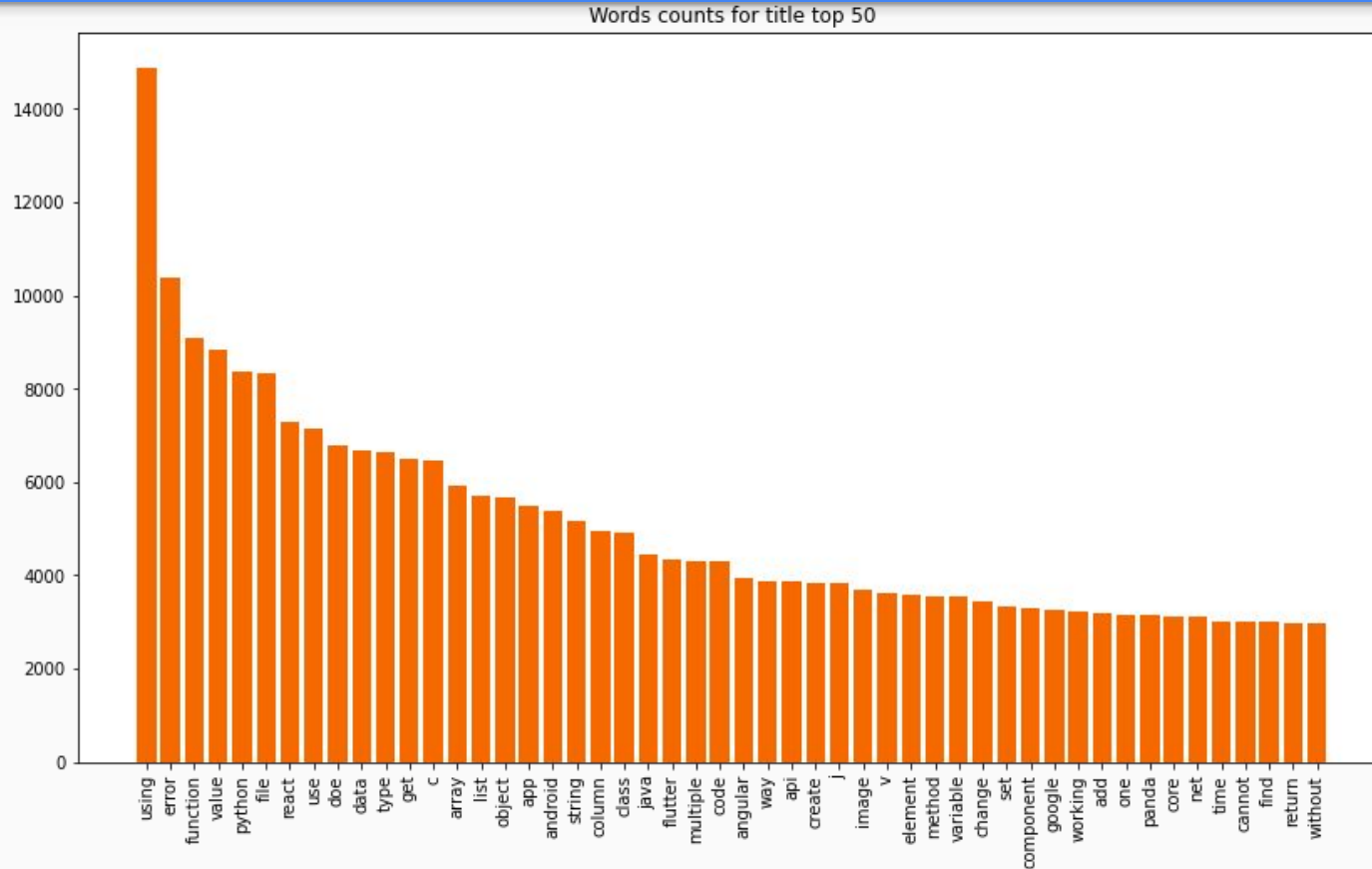
**202 861 questions**

# Tendances centrales



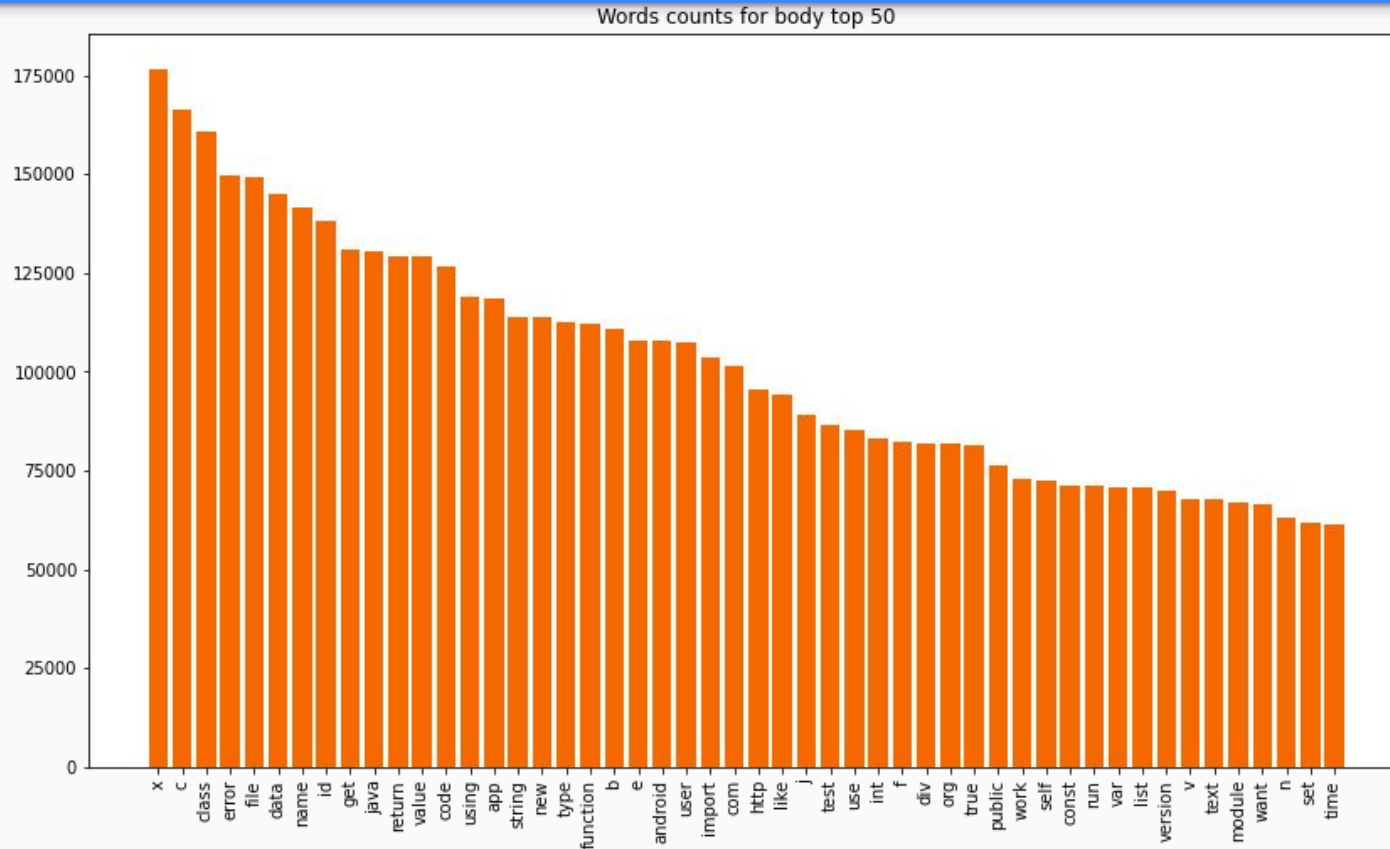
	body_words	title_words	tags_words
count	202861.000000	202861.000000	202861.000000
mean	150.741715	6.614480	4.277436
std	200.680904	2.374406	2.131873
min	1.000000	1.000000	1.000000
25%	55.000000	5.000000	3.000000
50%	96.000000	6.000000	4.000000
75%	171.000000	8.000000	6.000000
max	4560.000000	24.000000	17.000000

# Fréquences : Title



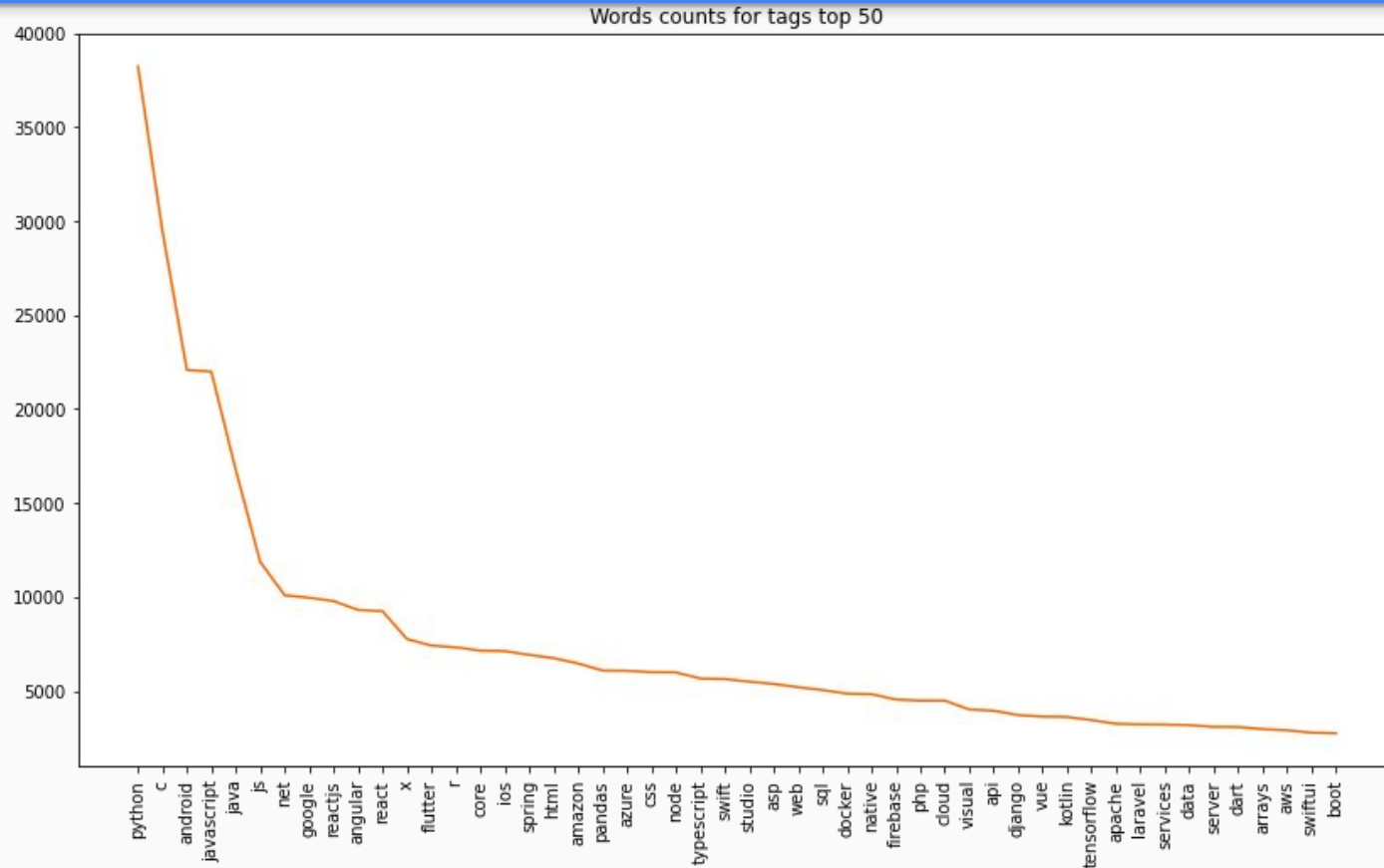
**4 114 mots  
uniques**

# Fréquences : Body



**541 371 mots  
uniques**

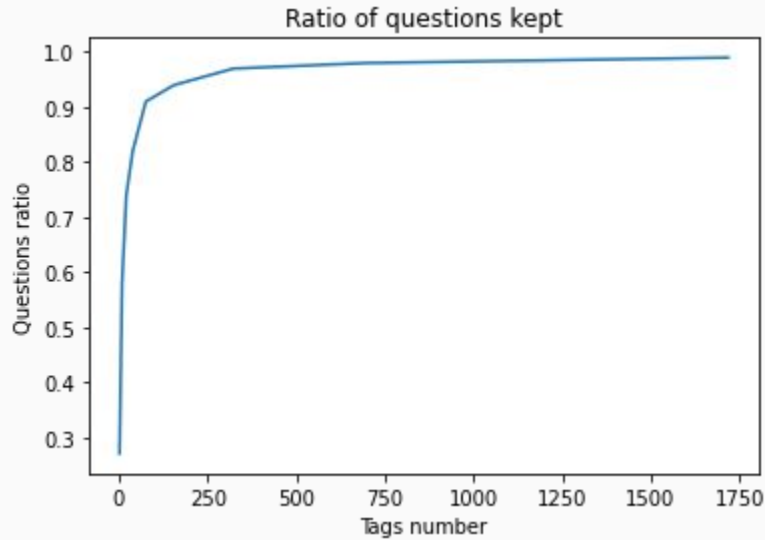
# Fréquences : Tags



**14 966 mots  
uniques**

# MODÉLISATION (supervisée)

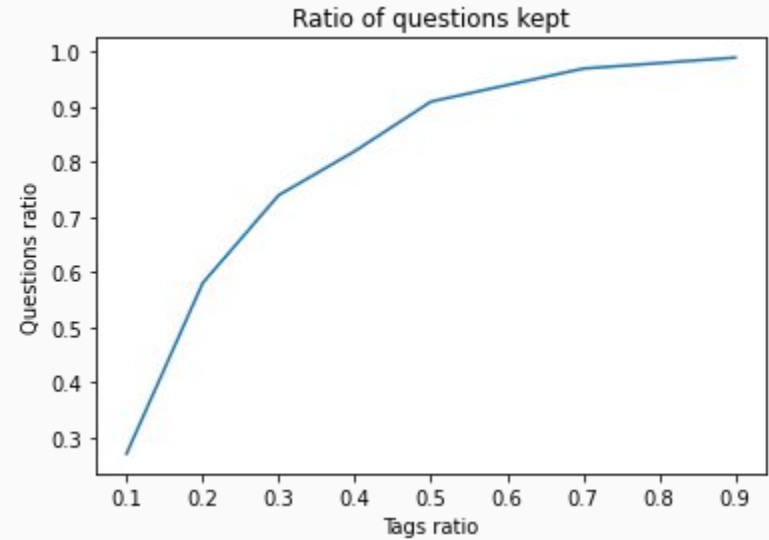
# Target : Tags



**76 Tags**

**Suppression 0**

tags



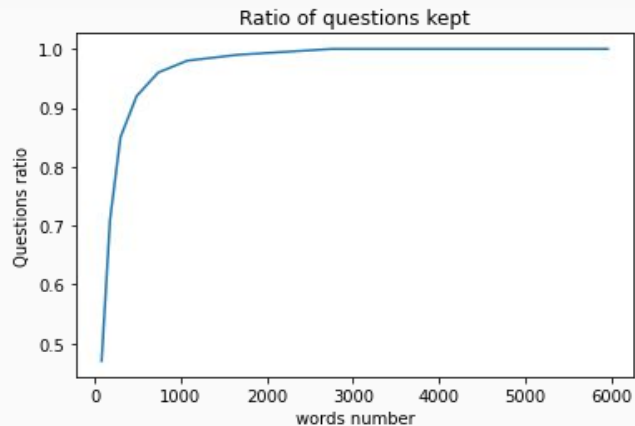
**classification multi-étiquettes (OvR)**

**Jaccard Score**

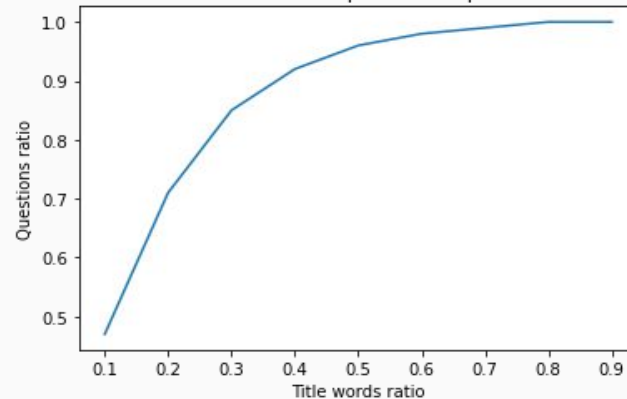
# Features

Title

mots

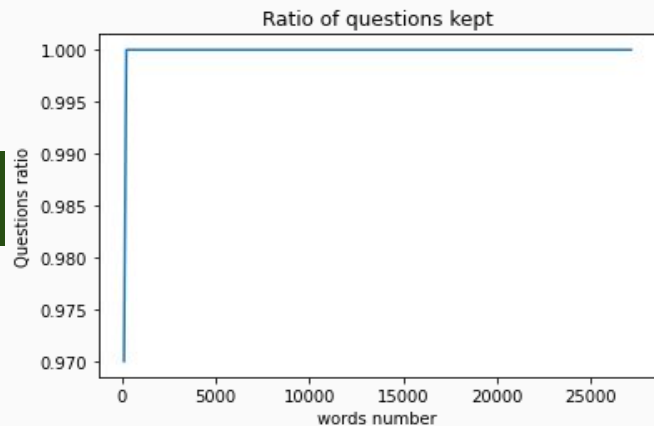


Number of questions kept

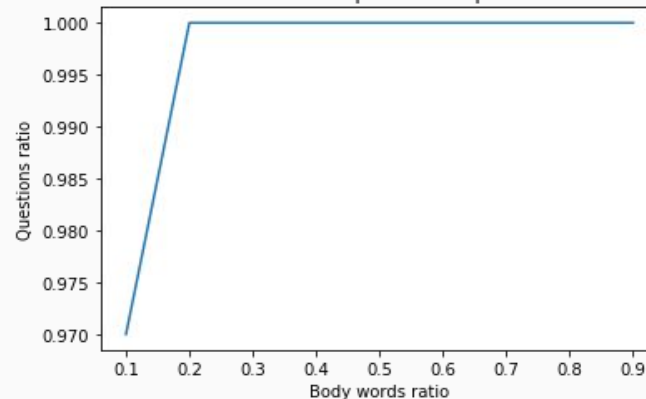


750 mots

Body



Ratio of questions kept



250 mots



# Approche et performances

Title

Body

Input BOW

Input tf-idf

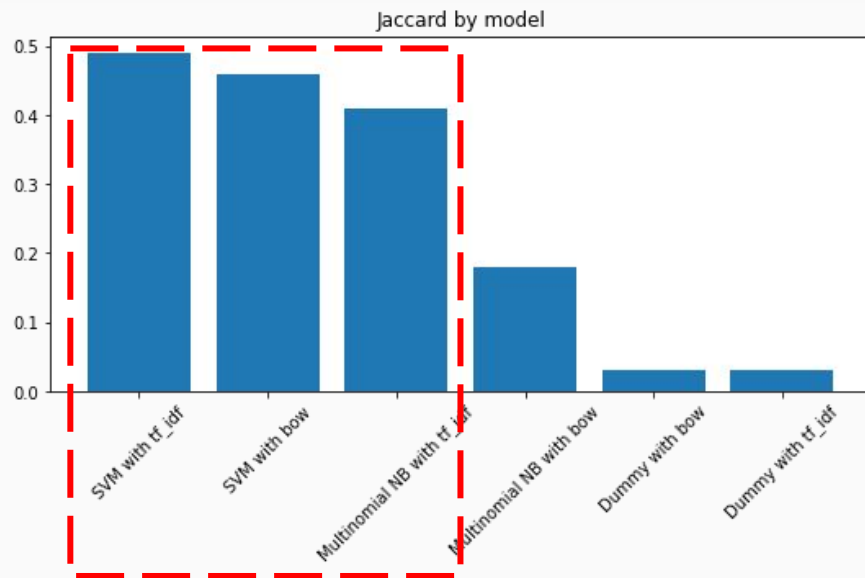
Multinomial NB classifier

Linear SVC

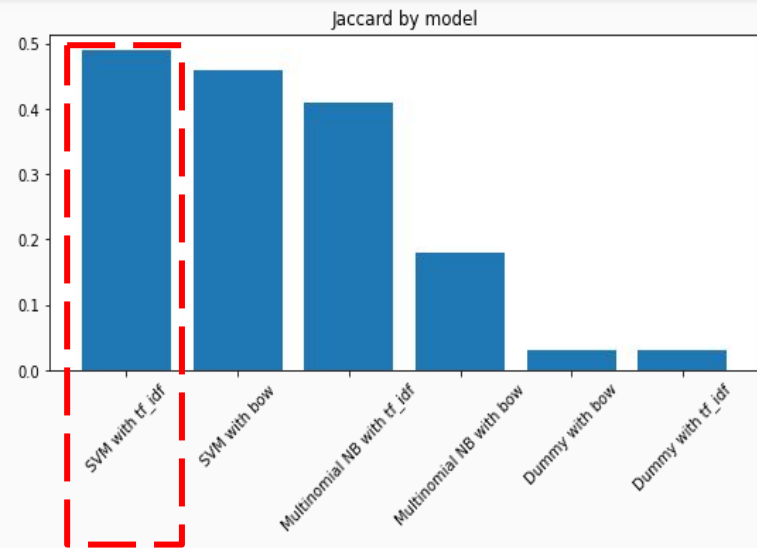
classification multi-étiquettes (OvR)

Jaccard Score

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$



# Gridsearch

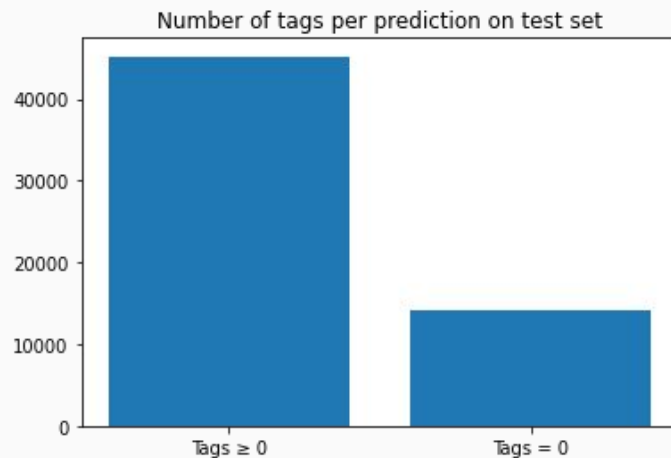


Input tf-idf

Linear SVC

Jaccard

Score



Gridsearch with Linear SVC and tf-idf

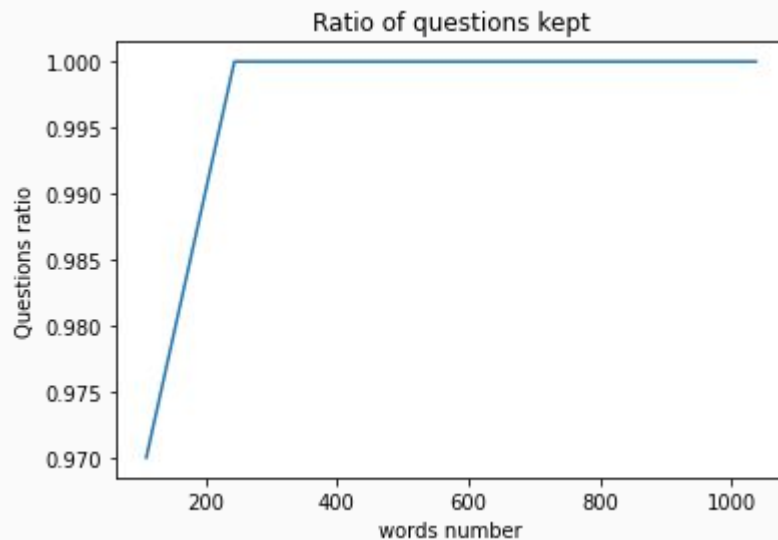
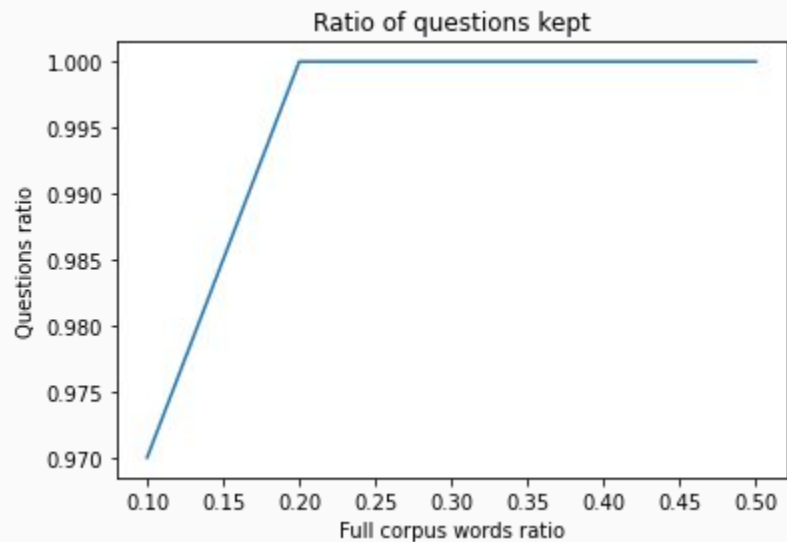
Gridsearch best params: {'estimator\_\_C': 10, 'estimator\_\_penalty': 'l2'}

Gridsearch Jaccard score on test split: 0.49

Gridsearch Jaccard score on train split: 0.51

# MODÉLISATION (non supervisée)

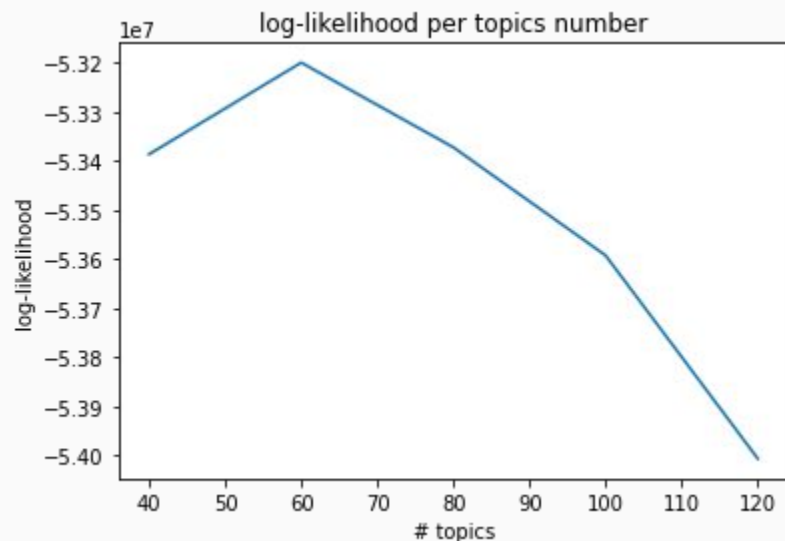
## Title + Body



250 mots

## Input BOW

## Latent Dirichlet Allocation



**60 topics**

**4 mots / tags**

```
Topic 0:
info task debug level
Topic 1:
size input output max
Topic 2:
true false state prop
Topic 3:
service net core token
Topic 4:
aws location resource permission
Topic 5:
end next vector double
Topic 6:
std template foo bar
Topic 7:
server client val stream
Topic 8:
option na select cell
Topic 9:
view io firebase fun
Topic 10:
map child flutter context
Topic 11:
db database sql spark
Topic 12:
java org internal gradle
```

# API

Dash




Démarche :

Si tags = 0

Modèle supervisé

Modèle non-supervisé

**GitHub** →  **heroku**

<https://stack-overflow-auto-tag.herokuapp.com/>

# CONCLUSION





Ask a public question

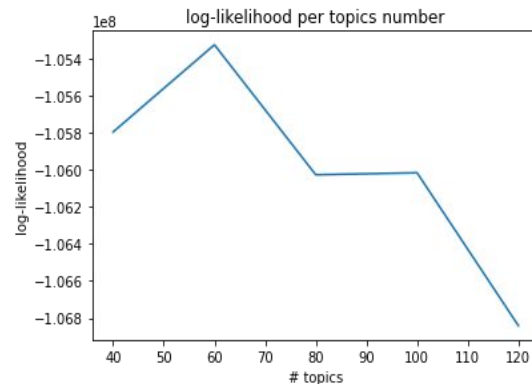
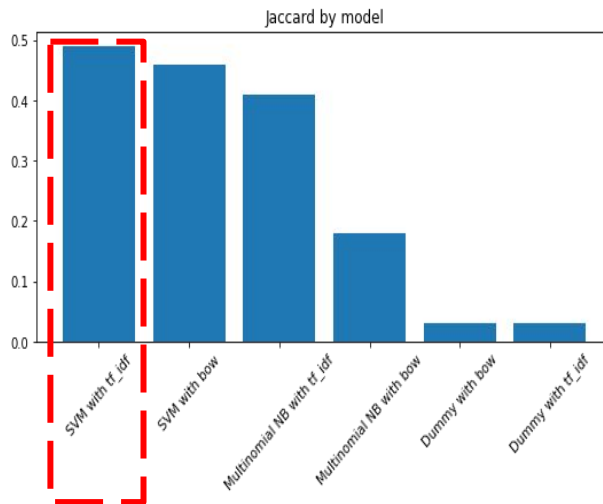
**Title**  
Be specific and imagine you're asking a question to another person.  
e.g. Is there an R function for finding the index of an element in a vector?

**Body**  
Include all the information someone would need to answer your question.

**Links** **Images** **Styling/Headers** **Lists** **Blockquotes** **Code** **HTML** **Tables** **More**

**Tags**  
Add up to 5 tags to describe what your question is about.  
e.g. (.net ruby json)

[Review your question](#)



# PERSPECTIVES

- Nettoyage termes spécifiques contexte ( ex : X, y). Bruit.
- Accès direct Stack Exchange API pour suggestions plus fines des tendances récentes.
- Utilisation réseaux de neurones.
- Exploration word embedding (word2vec)
- Augmenter le nombre tags conservés.
- n grams

# MERCI !



@xavbarbier



<https://www.linkedin.com/in/barbierxavier/>



<https://github.com/xavierbarbier/>



[contact@xavierbarbier.com](mailto:contact@xavierbarbier.com)