Rapport sur le projet

Catégorisez automatiquement des questions

OPENCLASSROOMS

Préparé par : Irina Maslowski

Plan

- Les objectifs du projet
- Résumé des tâches
- Présentation des données
- Catégorisation automatique des questions
  - Approche non-supervisée
  - Approche supervisée
- API (Application Programming Interface)
- Conclusion

Les objectifs du projet

- Objectif
  - Aider la communauter de StackOverflow

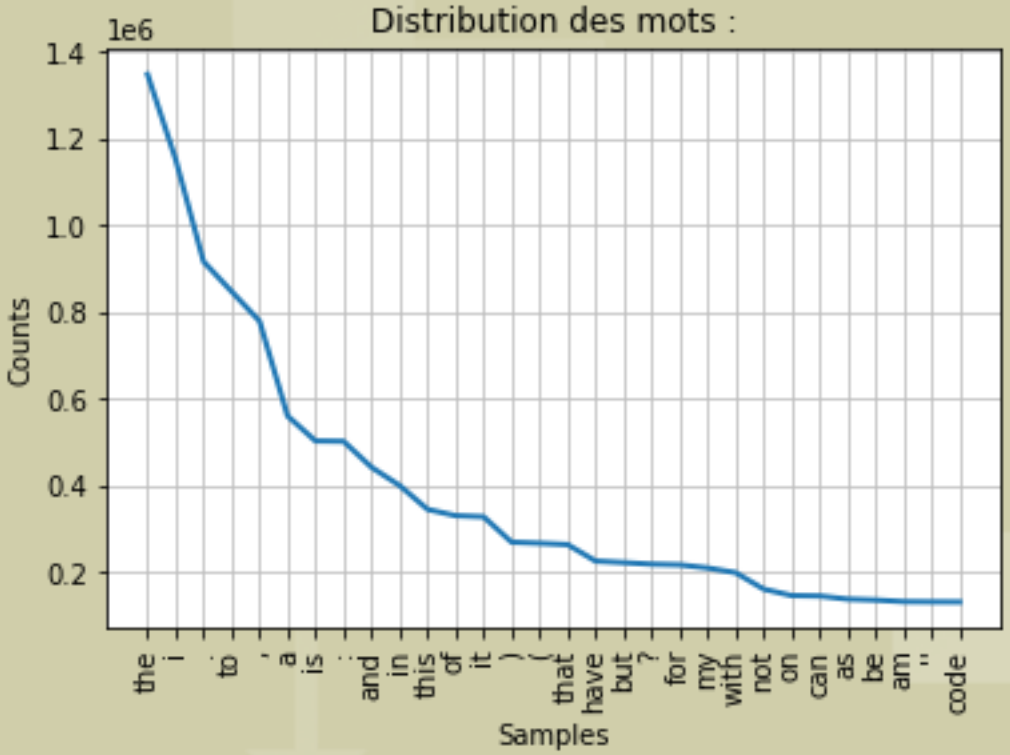
  de trouver facilement les mots clés pour les questions à poser
- Résultat attendu
  - Une application web proposant des mots clés (API)

Résumé des tâches

- Développer **un système de suggestion de tags**
  - Utiliser une approche supervisée ou non
- Evaluer le système
- Développer une API

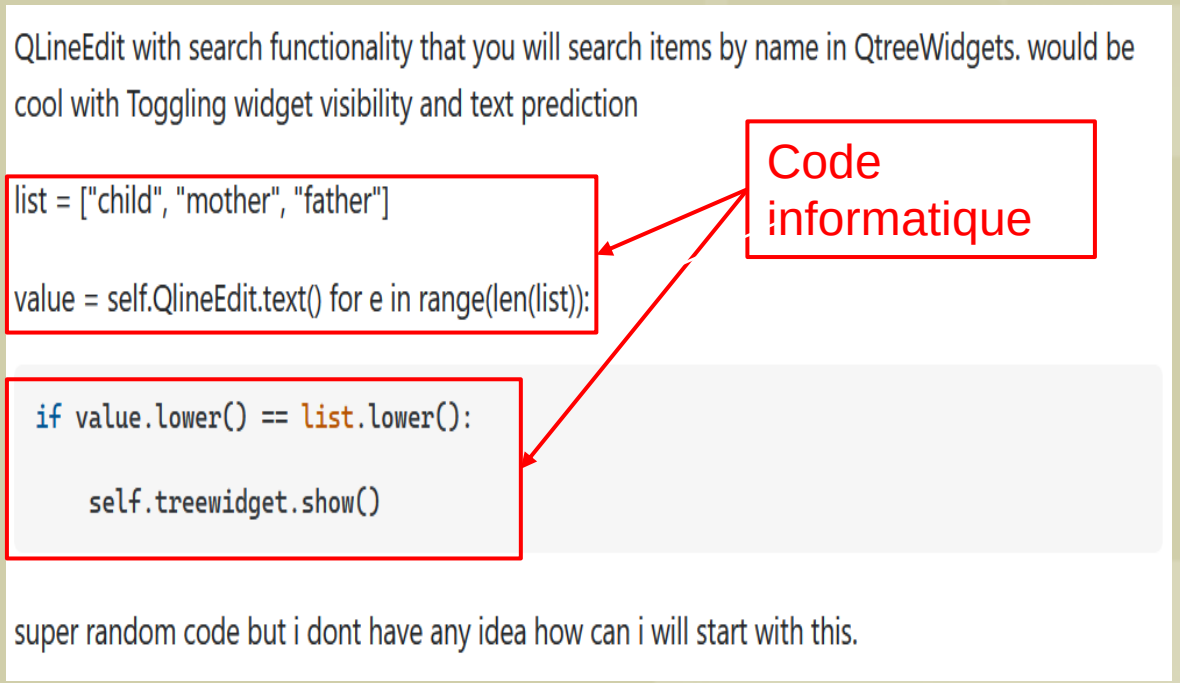
Présentation des données

- **Période temporelle**
  - Corpus d'apprentissage: 01/10/2020 et 01/10/2021
  - Corpus de test: octobre 2021
- **Caractéristiques quantitatives**
  - 252 420 publications
  - 27 535 825 mots
  - 93 mots/ question en moyenne
  - 1,7% de mots différents



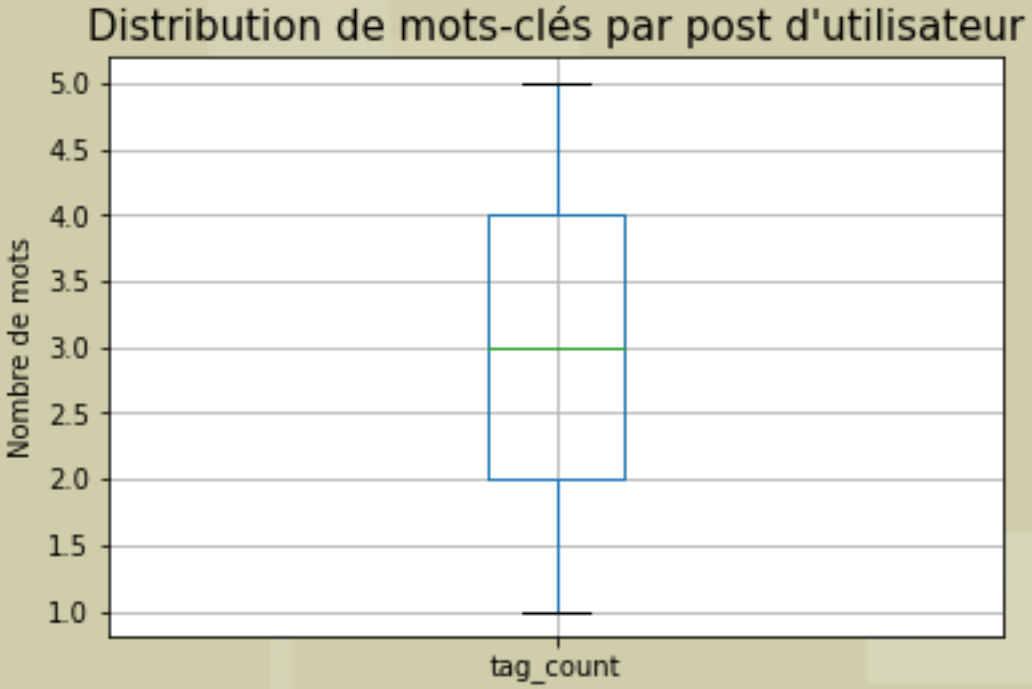
Présentation des données

- Spécificités de données
  - code informatique (n'es pas exploité)
  - termes métier → problems à la lemmatization ou racinisation ('css' -> 'cs', 'https' -> 'http')



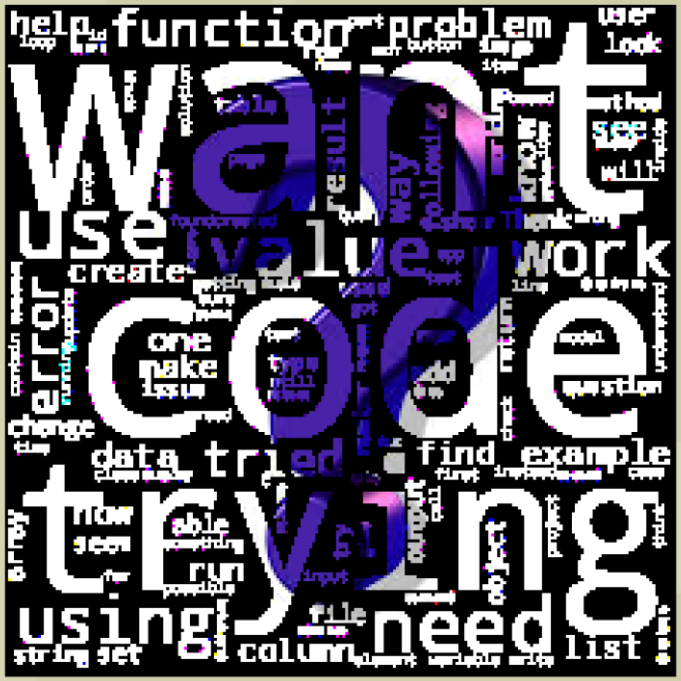
Présentation des données

- **Tags**
  - 776 026 tags
  - 3% tags distincts
  - 3 tags par question en moyenne



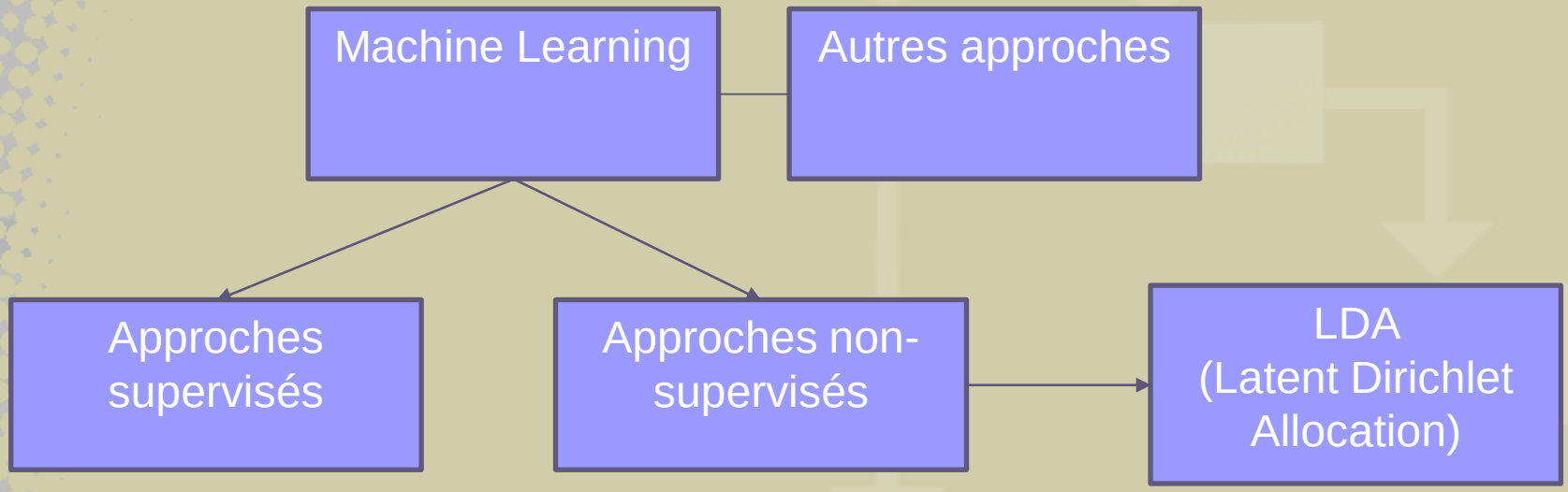
Présentation des données

- Quelles approches existent pour extraire des mots clés de ce type de données ?

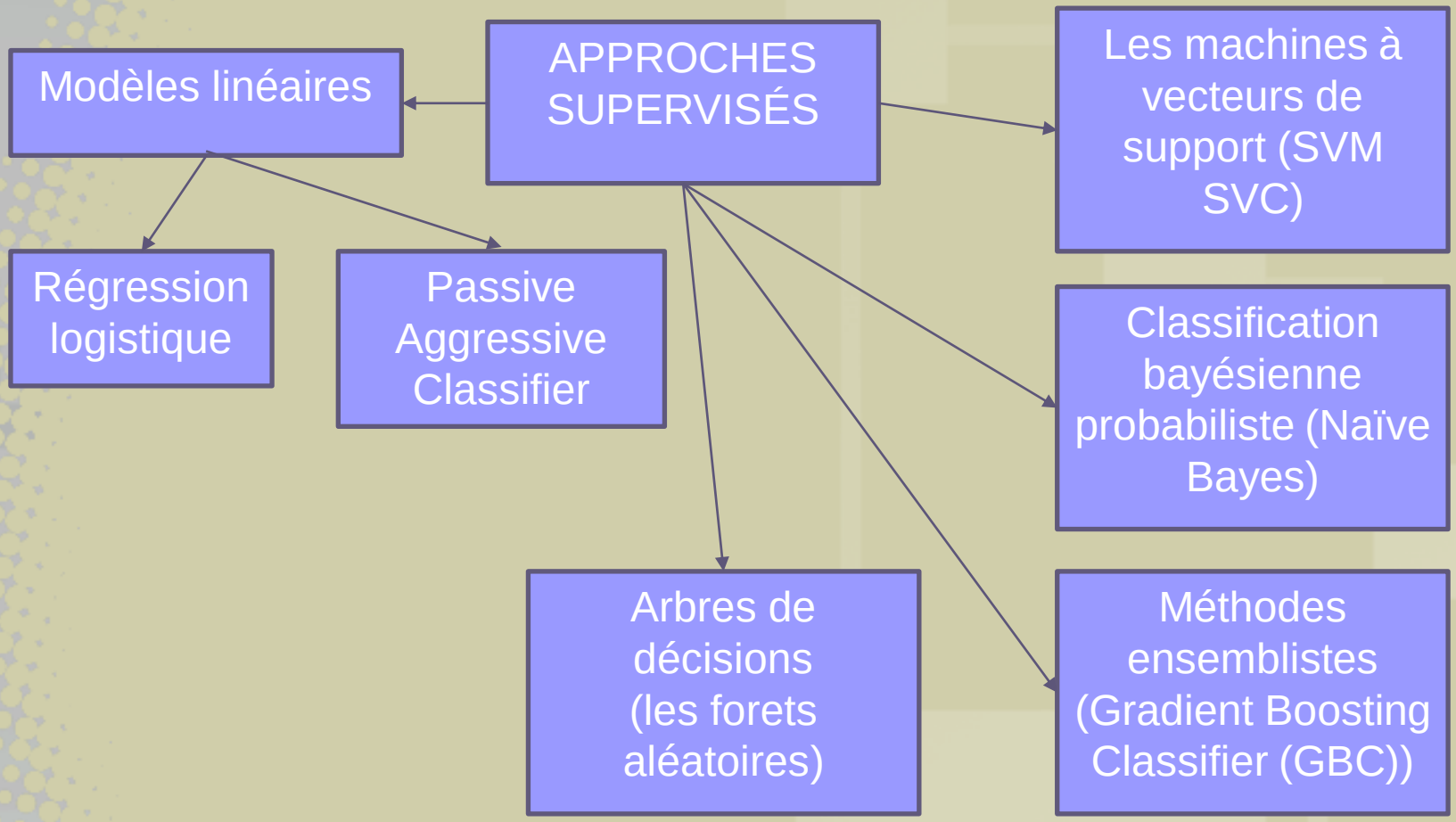


Catégorisation automatique des questions

= extraction de mots clés



Catégorisation automatique des questions



Catégorisation automatique des questions

- Méthodes de la réduction dimensionnelle
  - LDA
  - LSA (analyse sémantique latente)
  - NMF (factorisation matricielle non négative)

Approche non-supervisée

cohérence	poids de titre, n (titre x n)	max_df	min_df	nombre de thèmes
0,83	2	0,40	8	5
0,81	1	0,60	10	10
0,78	3	0,60	9	10
0,77	**1**	**0,60**	**9**	**20**

Approche non retenue

Approche supervisée

Modèle	Temps moyen d'entrainement (sec)	TF-IDF		Score de Jaccard
		max_df	min_df	
Dummy Classifier	159,7	0,7	10	0,008
GBC + NMF	3270,6	0,6	10	0,23
Logistic Regression	746,7	0,6	10	0,33
Passive Aggressive Classifier	**434,3**	**0,6**	**9**	**0,45**

Approche supervisée

Modèle	Score de Jaccard d'entrainement	Score de Jaccard de test	Précision	**Rappel**	**F-score**
Passive Aggressive Classifier	0,45	0,37	0,48	0,75	0,58

API (Application Programming Interface)

Predict Tags

Question Title

```
What programming language to choose?
```

Question Text

```
<p>Hi! What programming language to choose between Python, Java and C++?</p>
```

Conclusion

- Analyse de corpus "Questions StackOverflow 2020 - 2021«
- Analyse des approches existantes
- Approche choisie: supérvisée