Projet 5 - Catégoriser automatiquement des questions

Aurélien Corroyer-Dulmont, PhD

Ingénieur imagerie médicale

Centre Baclesse
L'excellence pour vaincre votre cancer

Rappel de l'appel à projet

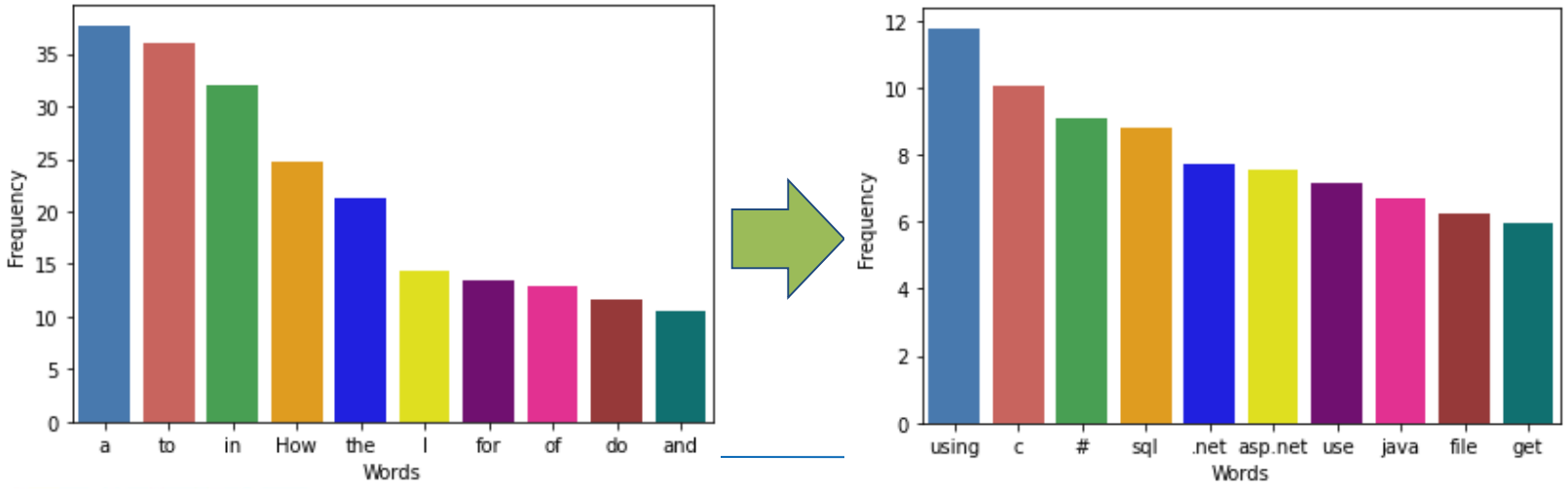
- Problématique :
  - Pour poser une question sur ***stack overflow*** il faut proposer plusieurs tags
  - Pour de nouveaux utilisateurs ce n'est pas forcément évident
- Objectif :
  - Développez un système de suggestion de tags pour le site utilisant un algorithme de machine learning qui assignera automatiquement plusieurs tags à une question posée.
- Données :
  - Base de données d'anciennes publications sur le sites comportant une question, un titre et des tags.

Centre Baclesse
L'excellence pour vaincre votre cancer

Nettoyage des données

Natural Language Analysis with Python NLTK

- Requête SQL sélectionnant les meilleurs post, avec des tags
- Nettoyage des informations textuelles non-intéressantes :
  - Utilisation des “*regex*” pour supprimer le texte inutile (*ex : “\”,”\n”,”\xa0”,”\s+”...*)
  - Suppression des “*stop-words*” (*ex : “the”, “a”, “an”, “in”...*)
  - Lemmatisation pour représenter les mots sous leur forme canonique (*“am”, “are”, “is”* => *“be”*)
  - Probabilité de lien entre les mots : utilisation des *bigram*
  - Suppression des balises html par *beautifulsoup*

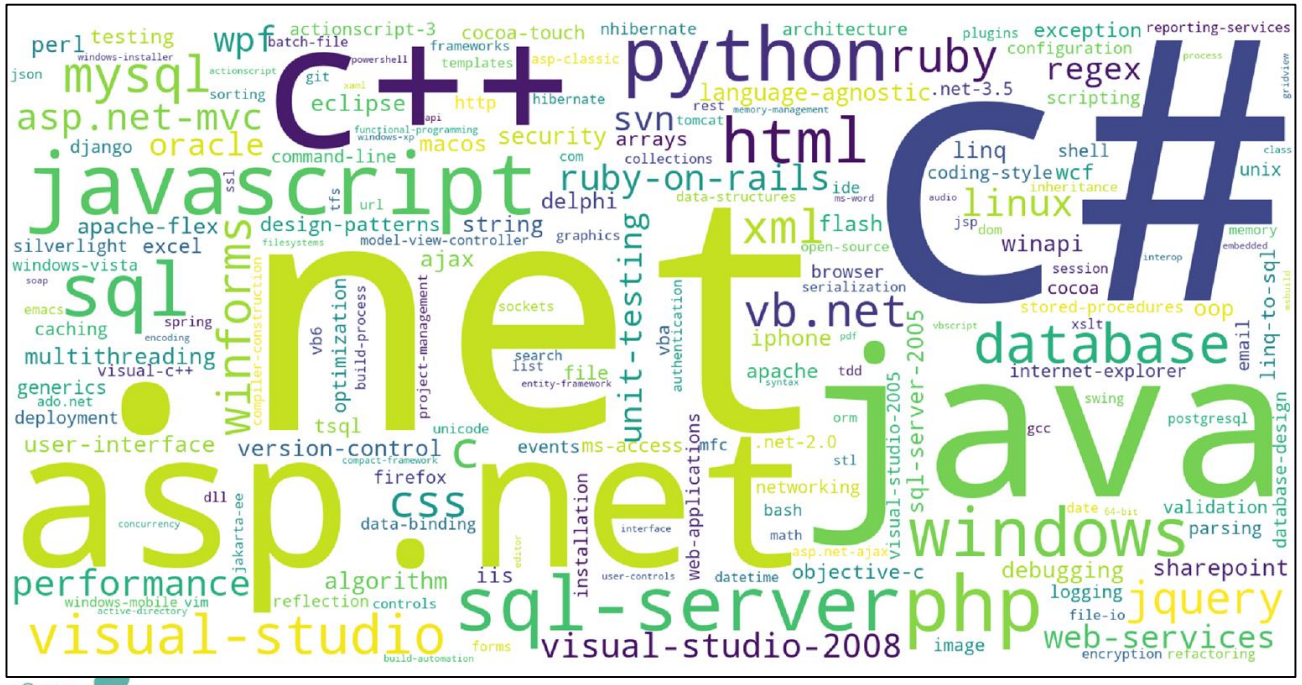


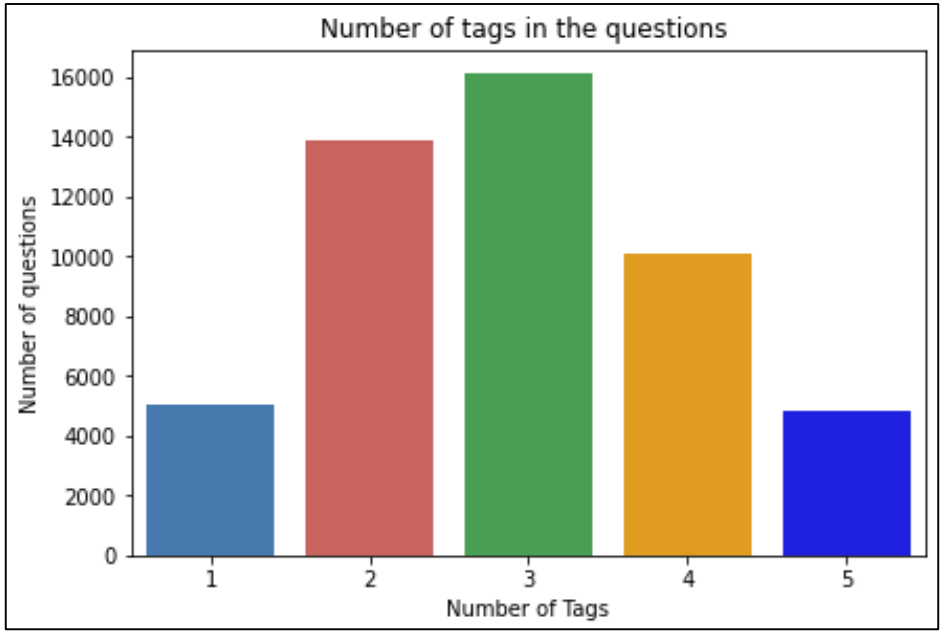
“Aggressive JavaScript caching, I've run into a problem where I make changes to a few JavaScript files that are referenced in an HTML file, but the browser doesn't see the changes”

“aggressive javascript cach've run problem make changes javascript files referenced html file browser”

Exploration des données

- Exploration globale des variables d'intérêt
  - Les mots clés qui ont la plus grande fréquence d'apparition sont *".net", "C#", "java"*
  - Le nombre de tags moyen par question est autour de 2-3





Centre Baclesse
L'excellence pour vaincre votre cancer

Modélisation

- Choix des modèles étudiés et méthodes d'extraction des features :

Non-supervisé :

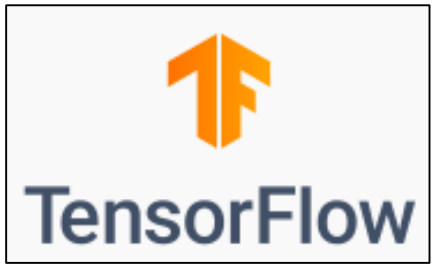
- LdaMulticore()

Supervisés :

- LogisticRegression()
- SGDClassifier()
- MultinomialNB()
- LinearSVC()
- RandomForestClassifier()

Extraction features:

- Bag-Of-Word()
- Word2Vec()
- BERT()
- USE()



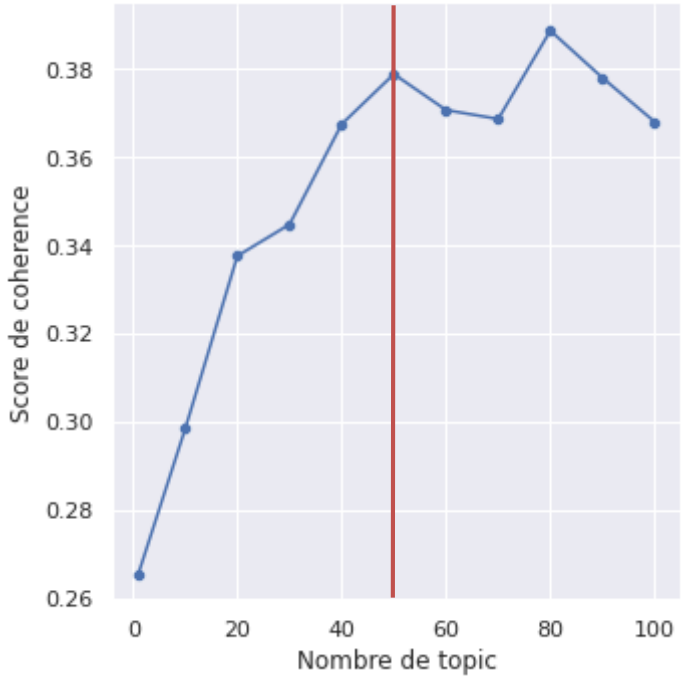
Centre Baclesse
L'excellence pour vaincre votre cancer

Modélisation non-supervisée

- Modèle *LdaMulticore* pour extraction de mots-clés :

  - Evaluation du **nombre de topics** représentatifs du corpus de mots

  - Etude du score de cohérence :



Un nombre de topic de 50 semble être assez représentatif du corpus de question

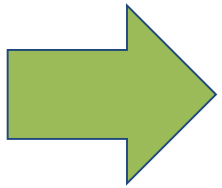
Centre Baclesse
L'excellence pour vaincre votre cancer

Modélisation non-supervisée

- Modèle *LdaMulticore* pour extraction de mots-clés :

Extraction des probabilités d'appartenance à des mots clés

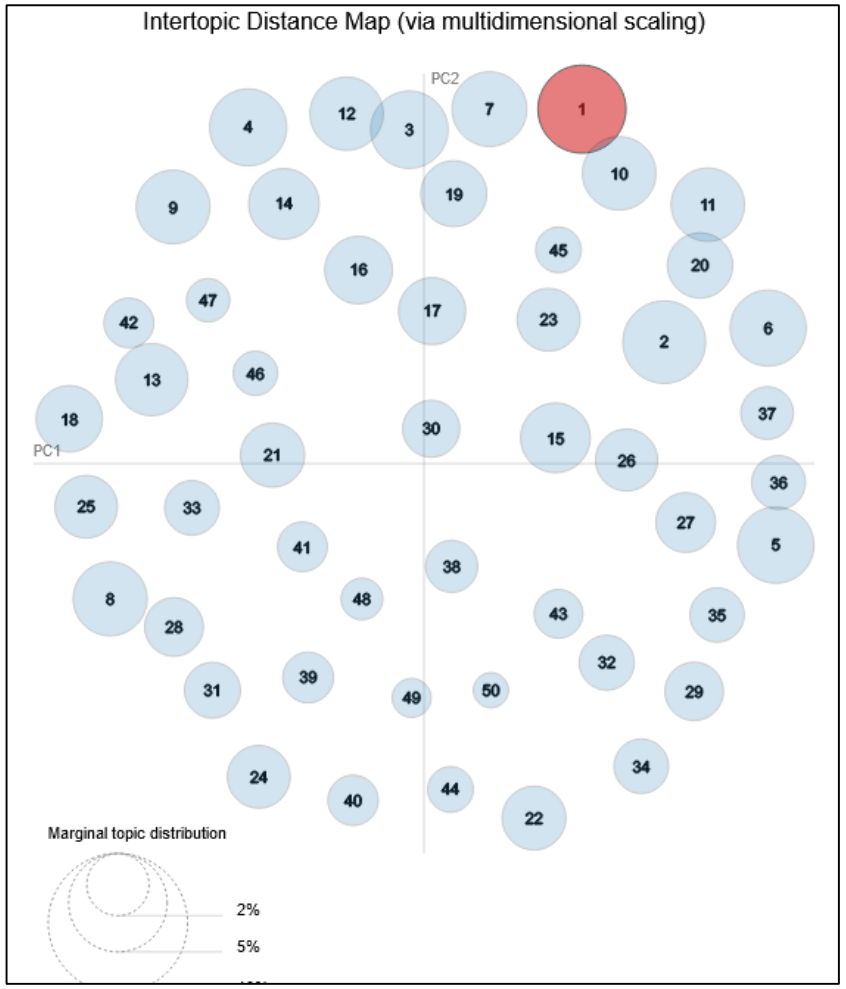
(0.15883656, 'code'),
(0.04904402, 'string'),
(0.04549977, 'return'),
(0.027419357, 'write'),
(0.026723403, 'would'),
(0.025701026, 'output'),
(0.024962518, 'list'),
(0.023175692, 'find'),
(0.021499356, 'value'),
(0.021286612, 'method'),
(0.017863216, 'call'),
(0.01754858, 'base'),
(0.01726231, 'select'),
(0.016704714, 'array'),
(0.016365414, 'group'),
(0.016351862, 'stre'),
(0.015215631, 'date'

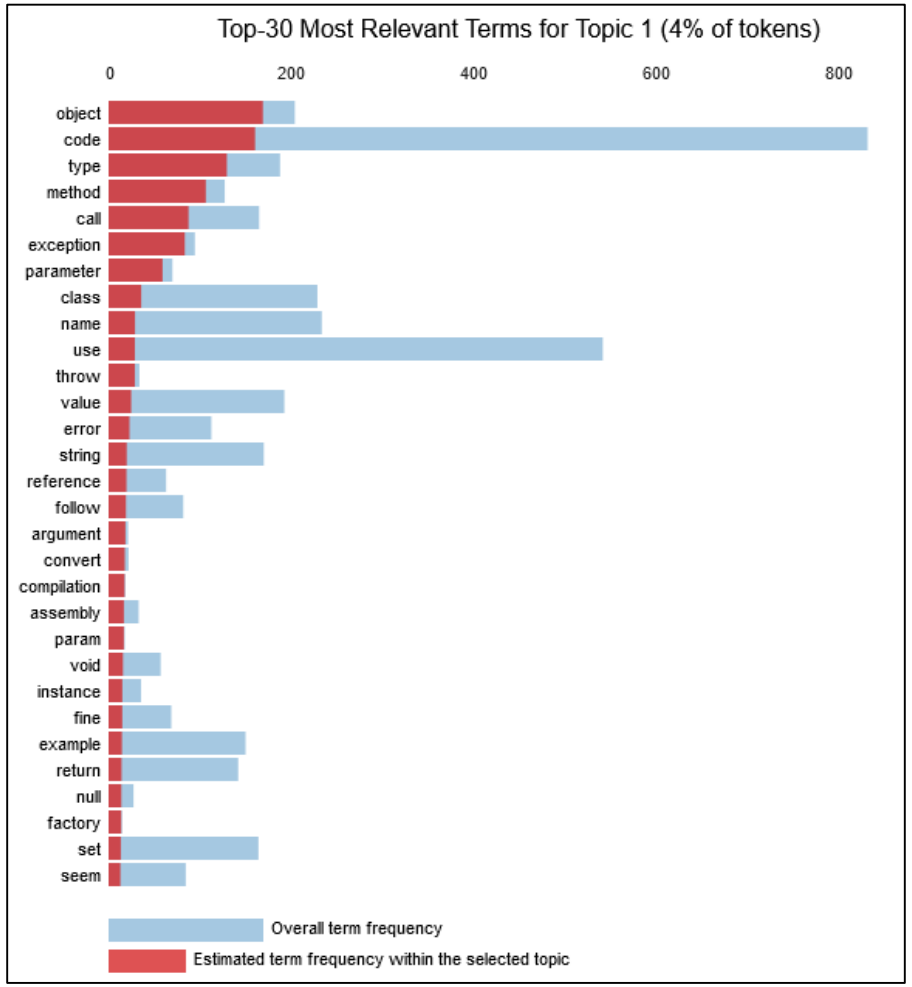


n°_Topic	Keyword_1	Keyword_2	Keyword_3
1	page	problem	control
2	see	index	much
3	code	string	return
4	function	message	open
5	name	code	query
6	com	question	stackoverflow
7	code	define	interface
8	process	bar	foo
9	application	net	use
10	access	develop	software

Modélisation non-supervisée

- Modèle *LdaMulticore* pour extraction de mots-clés :





Les résultats ne sont pas incongrus mais un modèle supervisée pourrait peut-être faire mieux

Centre Baclesse
L'excellence pour vaincre votre cancer

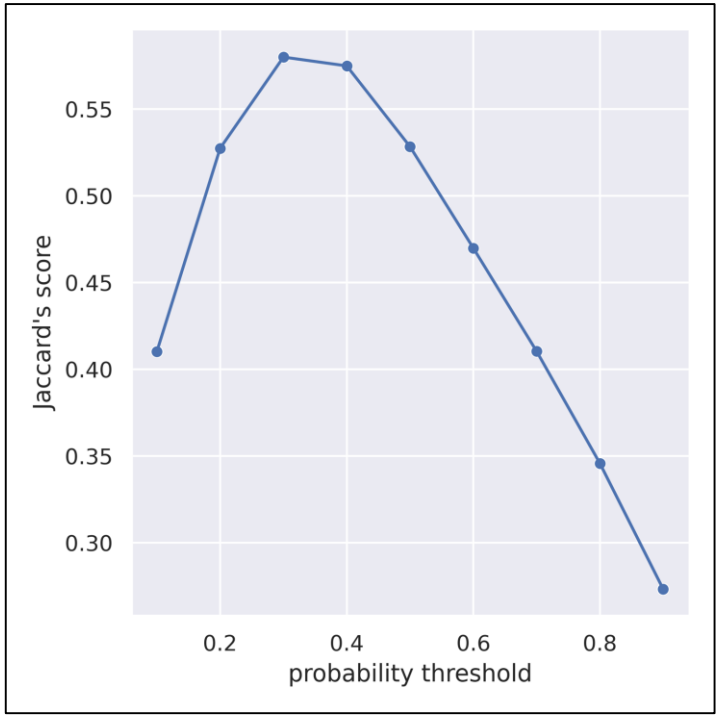
Modélisation supervisée

- Préparation des données pour les modélisations supervisées :
  - Sélection des tags les plus fréquents (50 sur les 4286 initiaux) :
    - Ces 50 tags apparaissent 1264 à 89 fois dans le dataset
  - Création d'un dictionnaire de mots compris dans le corpus:
    - Renvoi un dictionnaire de mots fréquents de 3962 mots
  - Split des données d'entraînement (*document* et *tags*)
  - Binarisation des données de tags
  - Extraction des features par méthode de Bag-Of-Words

Centre Baclesse
L'excellence pour vaincre votre cancer

Modélisation supervisée

- Entraînement des modèles supervisés :
  - Utilisation d'un pipeline incluant "*TF-IDF transforme*"
  - Evaluation de l'impact du paramètre "*predict-proba*" sur l'efficacité des modèles :

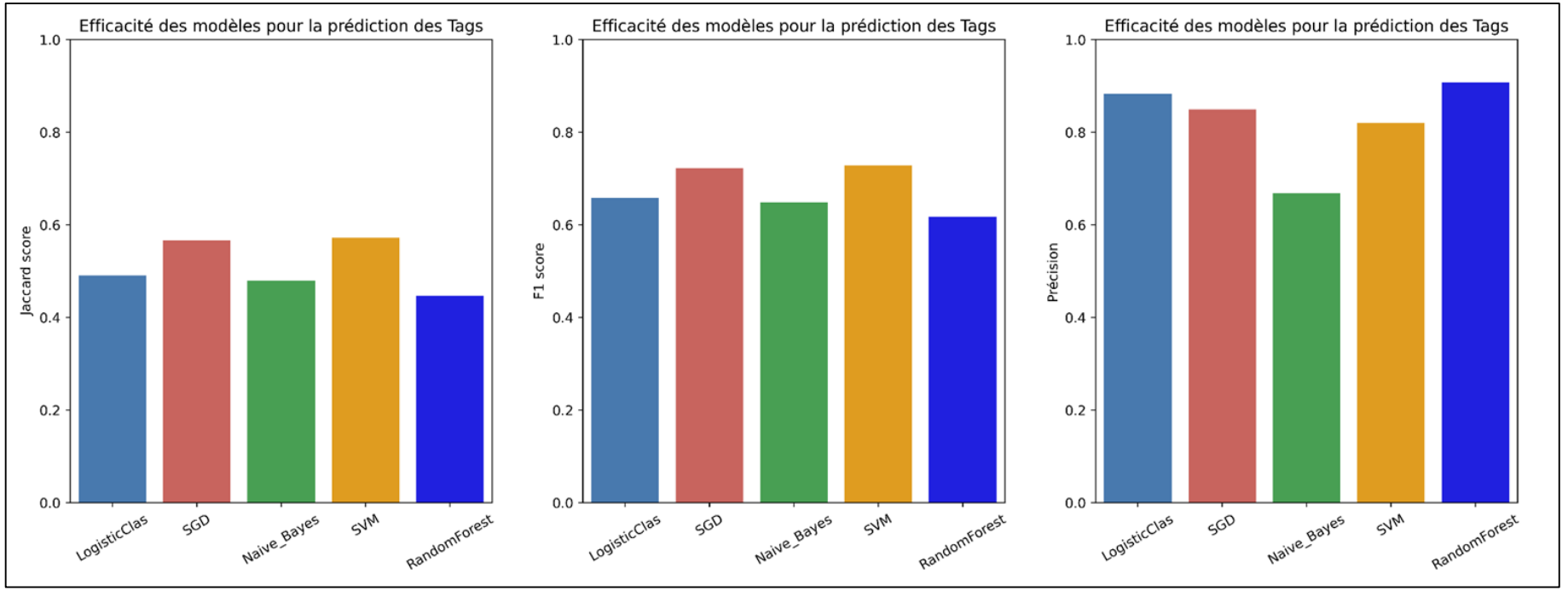


Un seuil à 0.3 semble pertinent dans notre projet

Centre Baclesse
L'excellence pour vaincre votre cancer

Modélisation supervisée

- Entraînement des modèles supervisés :
  - Recherche des hyperparamètres pour chacun des modèles
  - Evaluation de l'efficacité des modèles :

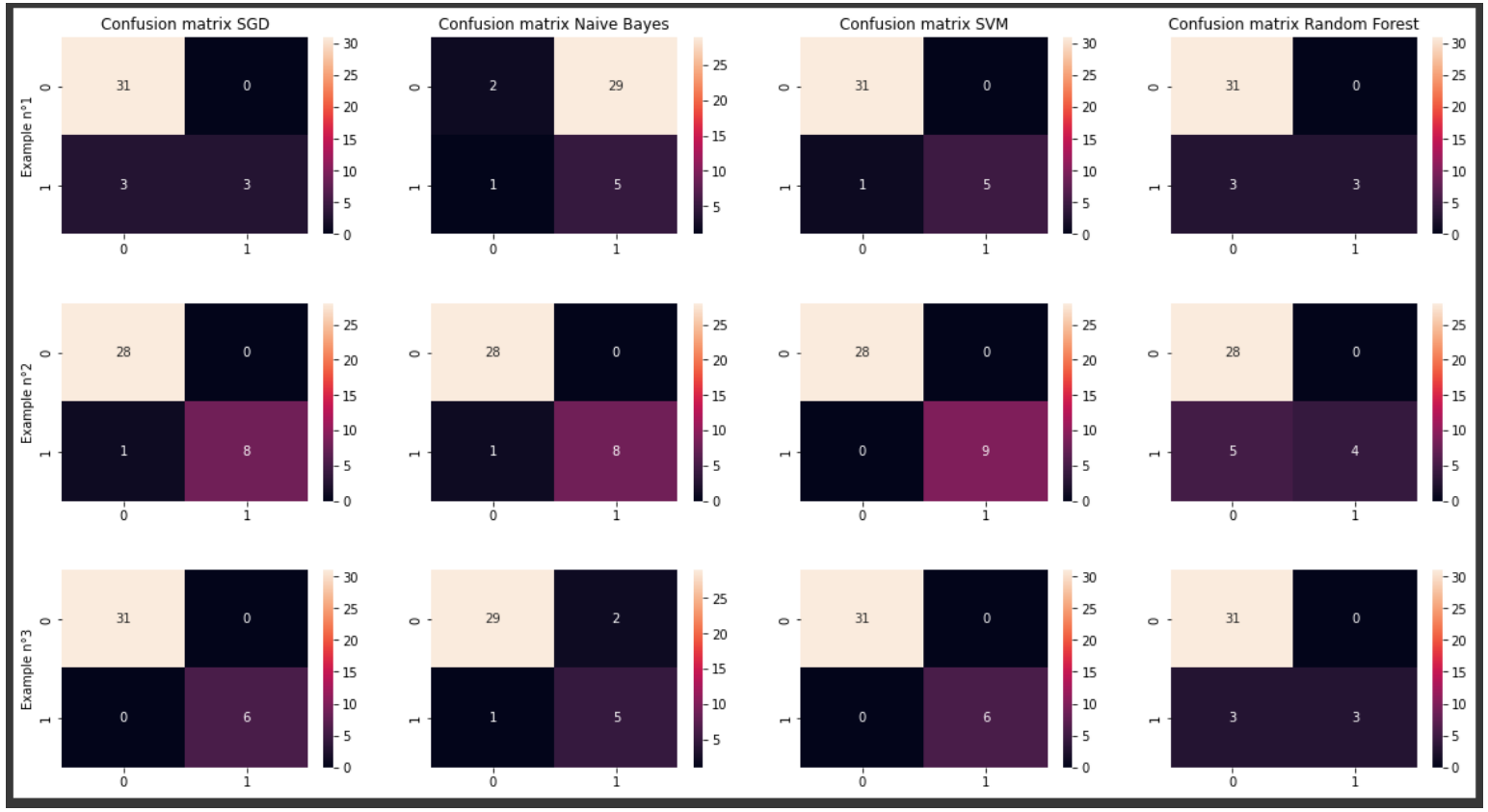


Les modèles SGD et SVM semblent être les plus performants

Centre Baclesse
L'excellence pour vaincre votre cancer

Modélisation supervisée

- Entraînement des modèles supervisés :
  - Evaluation de l'efficacité des modèles : matrice de confusion



Centre Baclesse
L'excellence pour vaincre votre cancer

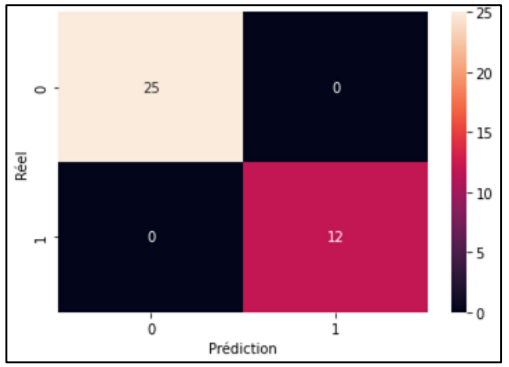
Modélisation supervisée

- Entraînement des modèles supervisés :
  - Evaluation de l'efficacité du modèle SGD: matrice de confusion

Fonctionne bien :

Pour la phrase : '*aggressive javascript cach'ved run problem make changes javascript files referenced html file browser*'

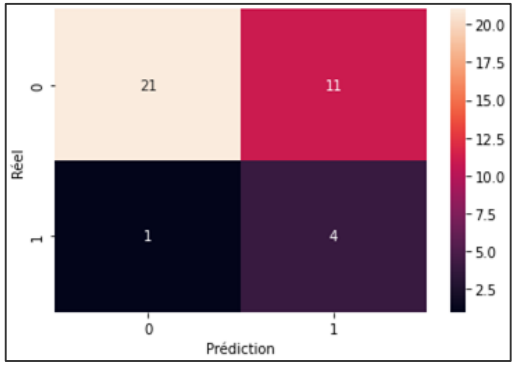
Le tag proposé est : [**'javascript'**]



Fonctionne moins bien :

Pour la phrase : '*detect whether user control running ide debug mode released exuser control building.*'

Le tag proposé est : [**'c#'**]



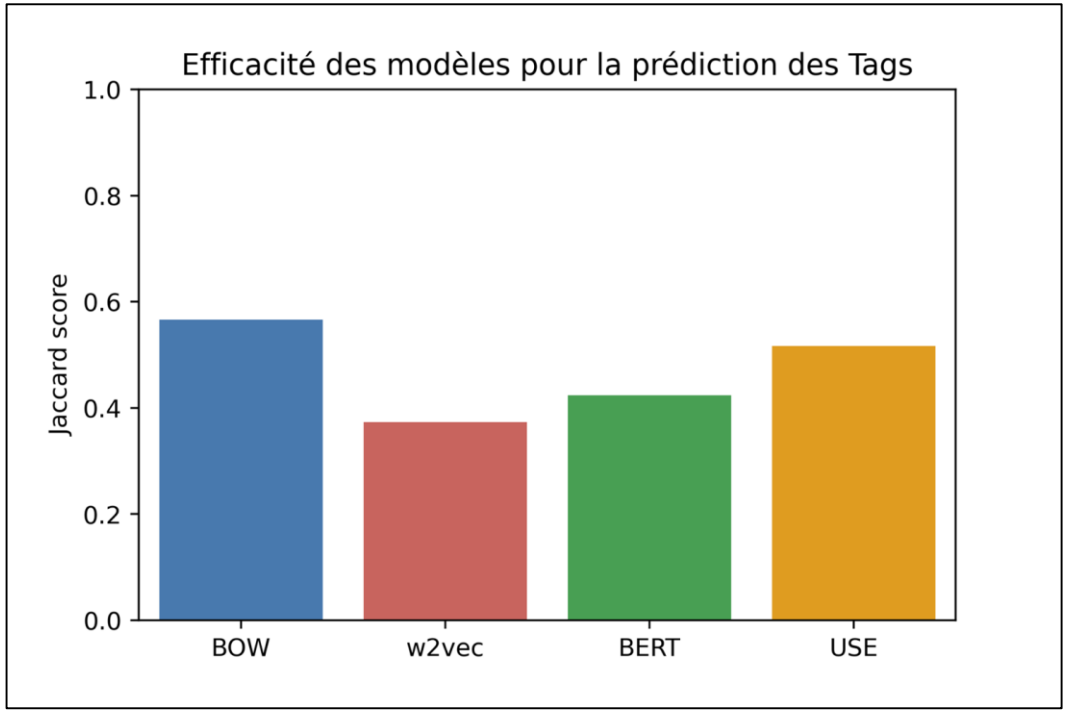
Centre Baclesse
L'excellence pour vaincre votre cancer

Modélisation supervisée

- Impact de la méthode d'extraction des features :
  - Evaluation de l'efficacité du modèle SGD avec les différentes méthodes d'extraction :

Extraction features:

- Bag-Of-Word()
- Word2Vec()
- BERT()
- USE()

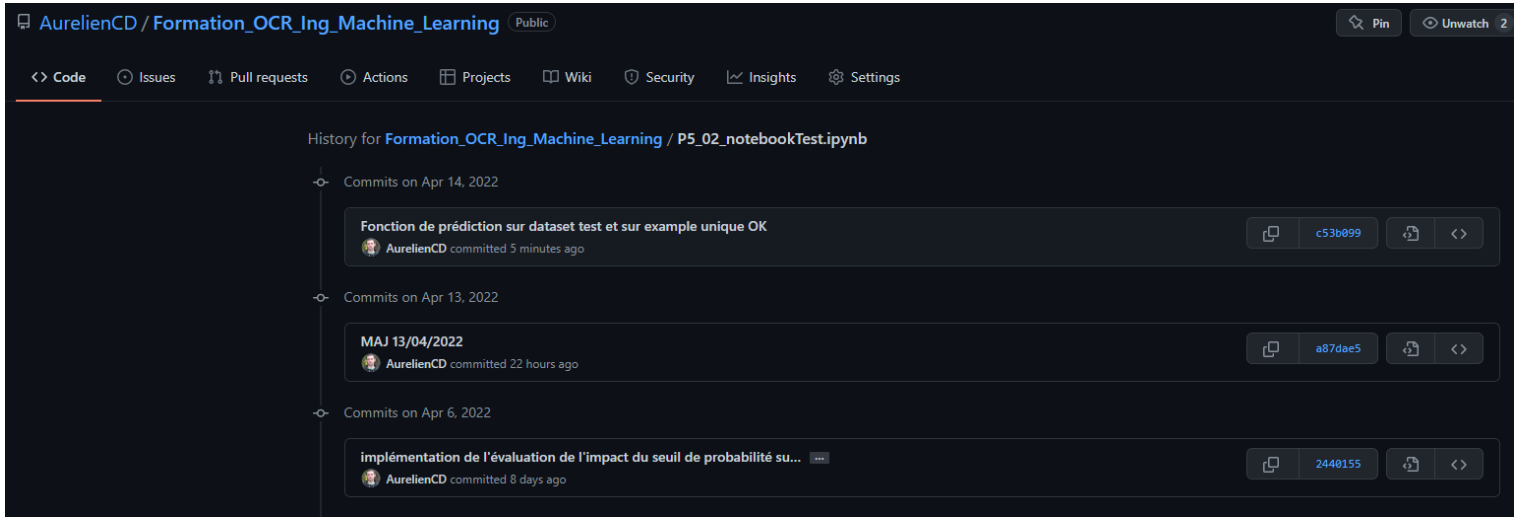


Centre Baclesse
L'excellence pour vaincre votre cancer

Code final à déployer

- Développement d'un code regroupant les fonctions de nettoyage, de préparation des données et de prédiction
  - Utilisation de la librairie "*joblib*" pour sauvegarder/importer les variables (dictionnaire, binariser) et fonctions de nettoyage/prédiction nécessaires

- Utilisation d'un logiciel de gestion de versions (GitHub) pour suivre les modifications du code final à déployer.

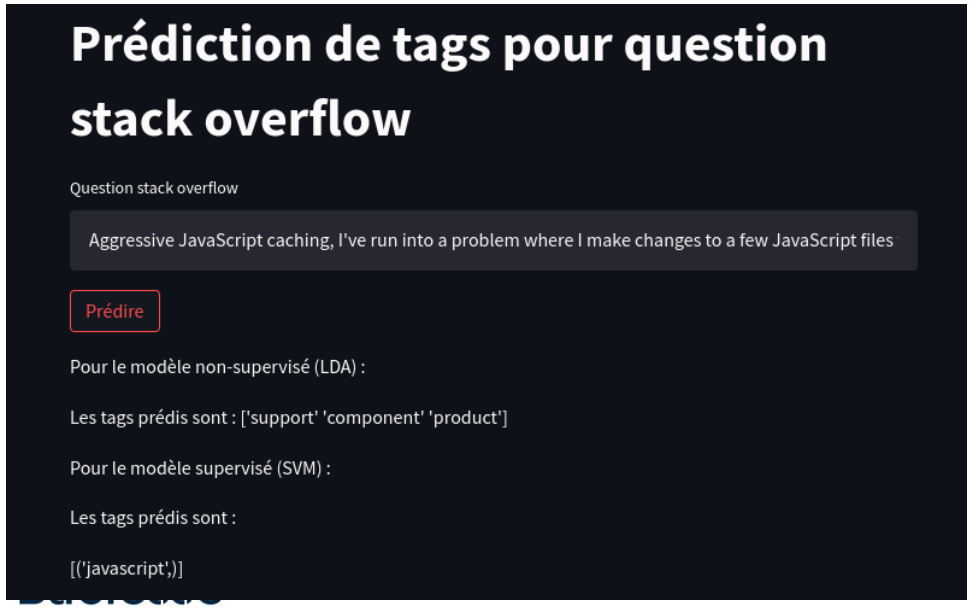


Centre Baclesse
L'excellence pour vaincre votre cancer

Mise en production : API

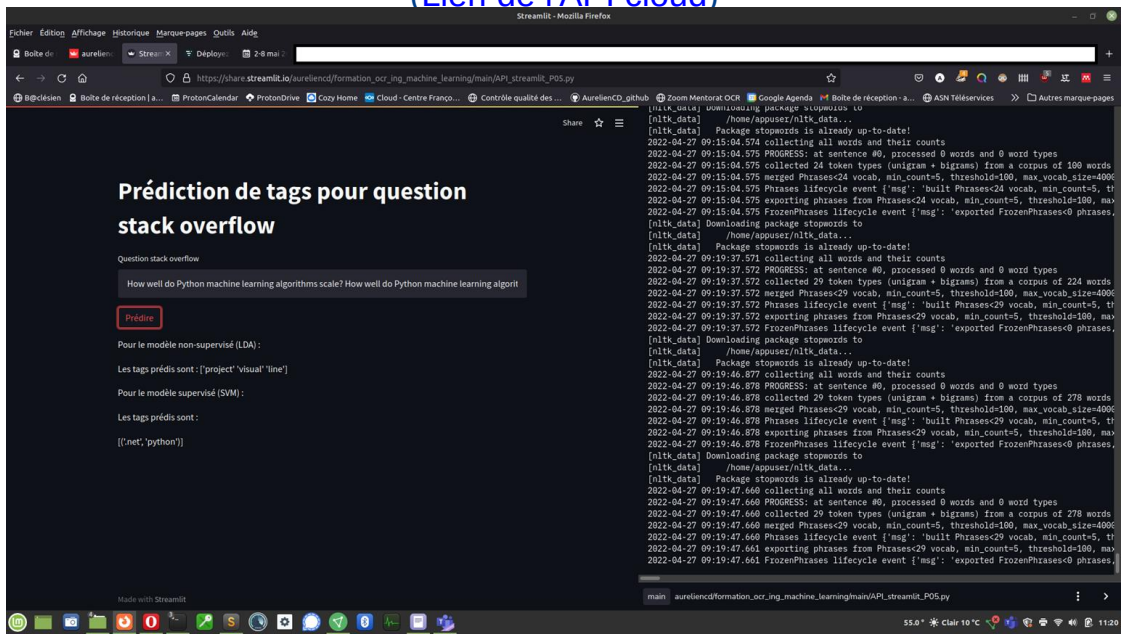
- Création d'un point d'entrée pour une API :
  - Développement du code en utilisant :
    - *"MLFlow"* pour la mise en production du modèle
    - *"Streamlit"* pour le déploiement :

- **En local** :



- **Sur le cloud** :

(Lien de l'API cloud)



https://stackoverflow.com/questions/8284373/how-well-do-python-machine-learning-algorithms-scale
https://stackoverflow.com/questions/72027373/how-to-get-a-removed-reference-tuple-type-without-using-an-instance-in-c

CONCLUSION

- Rappel de la problématique :
  - Développer un système de suggestion de tags vis-à-vis de questions sur le site stack overflow

- Résultats :
  - Une modélisation non-supervisée donne des résultats moyens mais permet d'obtenir des mots clés par question qui ne sont pas incohérents
  - Les méthodes supervisées après optimisation des hyperparamètres donnent de meilleurs résultats
  - Les modèles SGD et SVM présentent les meilleurs résultats et permettent une mise en production pour répondre à la problématique

Centre Baclesse
L'excellence pour vaincre votre cancer

Projet 5 - Catégoriser automatiquement des questions

Aurélien Corroyer-Dulmont, PhD

Ingénieur imagerie médicale

Centre Baclesse
L'excellence pour vaincre votre cancer