Projet 5 : Suggestion de tags

Ali-higo Ebo Adou

September 19, 2022

Outline

Outline

Problématique

Natural Language Processing (NLP)

- Traduction
- Analyse de sentiment (classification)
- Analyse sémantique
- etc..

Classification multi-label en NLP

- Classification de commentaires toxique et/ou haineux et/ou criminel etc..
- Tagging pour la classification de CV

Approches pour le NLP

En NLP, les *mots* sont encodés au format numérique (ou vectorisés) pour qu'ils puissent être interprété par l'ordinateur, par exemple à l'aide d'un encodage *One-hot*, *Bag-of-Word* ou *Embedding*.

Machine Learning

- *Classification*
- *Topic Modelling*

Deep Learning: *State-of-the-art*

- *Recurrent neural network*
- *Transformers*

Métriques pour la classification multi-label

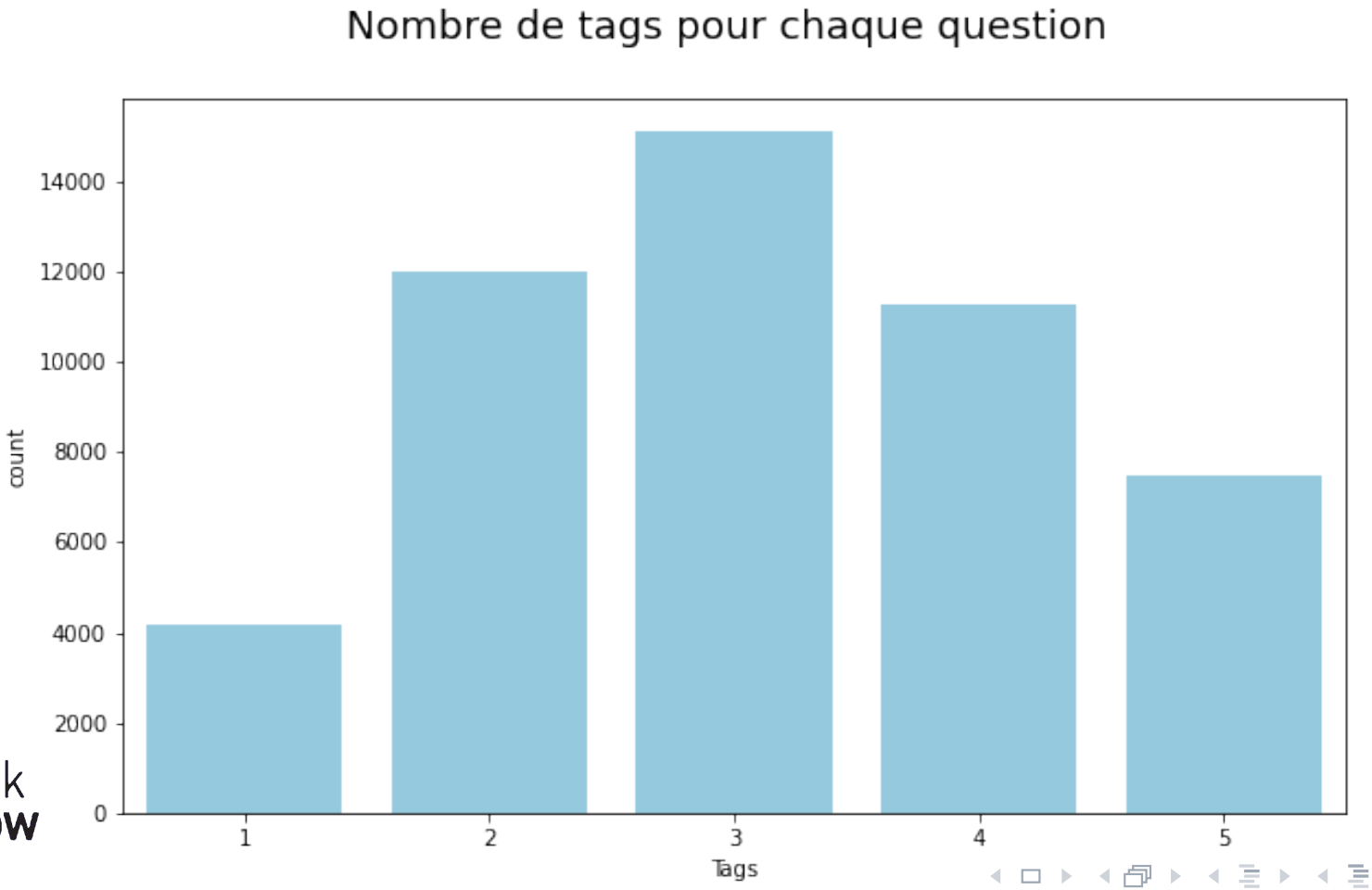
- Forte Précision / Fort Rappel :
  → Situation idéale !
- Forte Précision / Faible Rappel :
  → Le modèle prédit globalement peu et bien.
- Faible Précision / Fort Rappel
  → Le modèle prédit globalement beaucoup et mal.
- Faible Précision / Faible Rappel :
  → Le modèle peu et mal.

Outline

Le jeu de données original 1/2

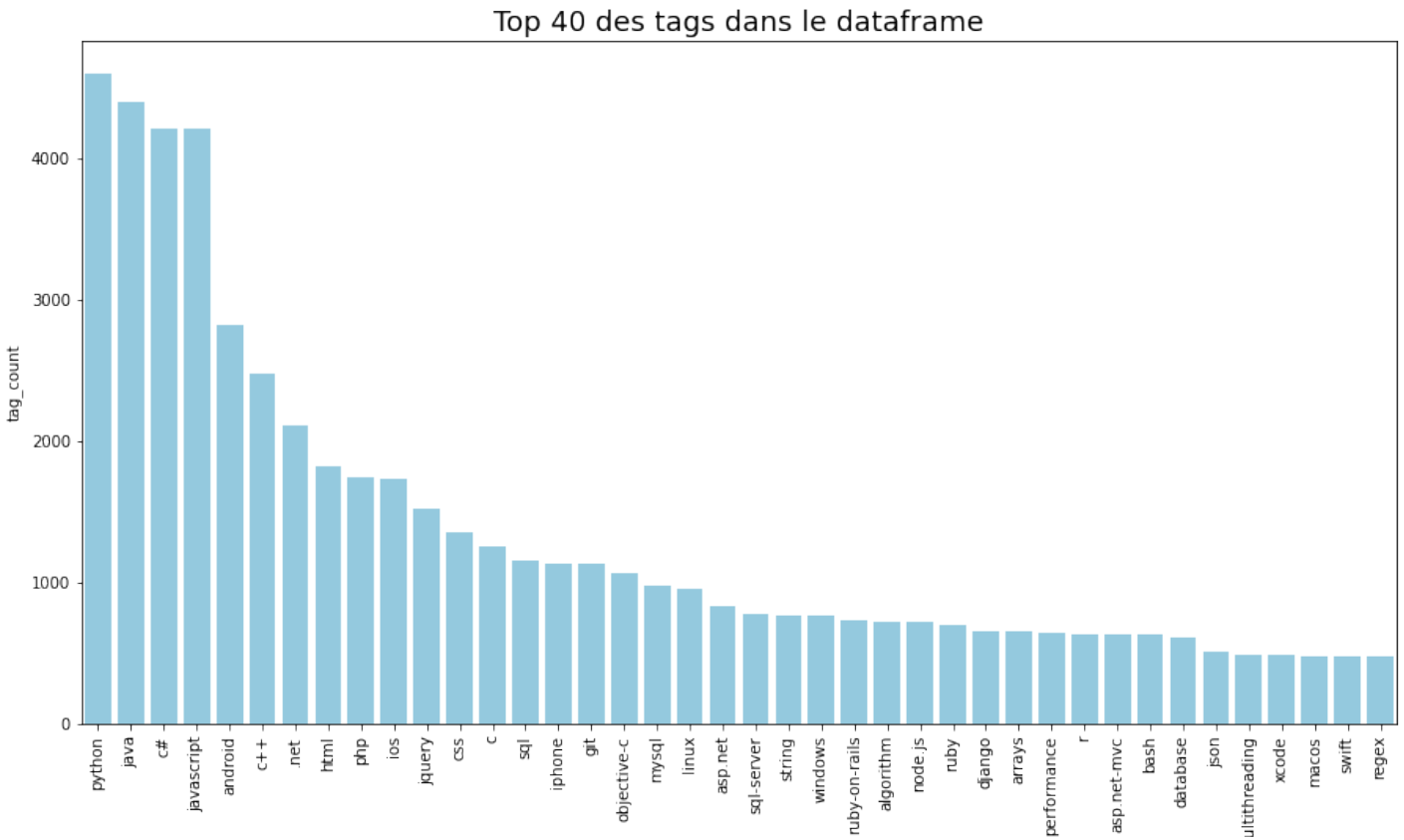
Critères de selection :

• plus de vues • plus de réponses • ayant au plus 5 tags



Le jeu de données original 2/2

- environ 50 000 lignes
- plus de 11 700 tags uniques



Nettoyage du dataset

Pre-Cleansing

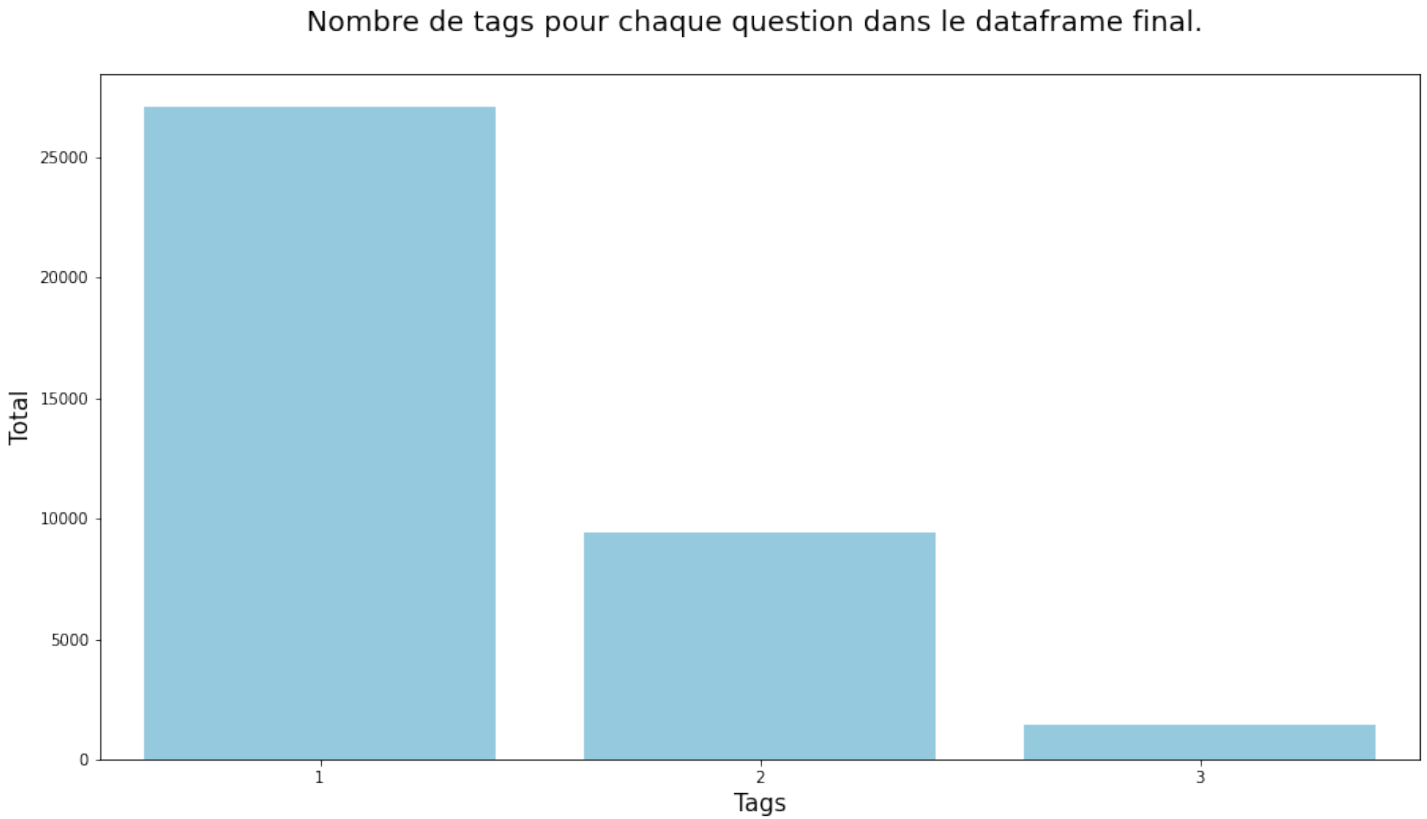
Retrait des balises HTML, adresses url, etc...

Tokenization & Lemmatization

Réduction/troncature de la taille du corps du texte à 128 caractères maximum.

Dataset final

- Réduction du nombre de tags à 35.
- Réduction du dataset original d'environ 25%.



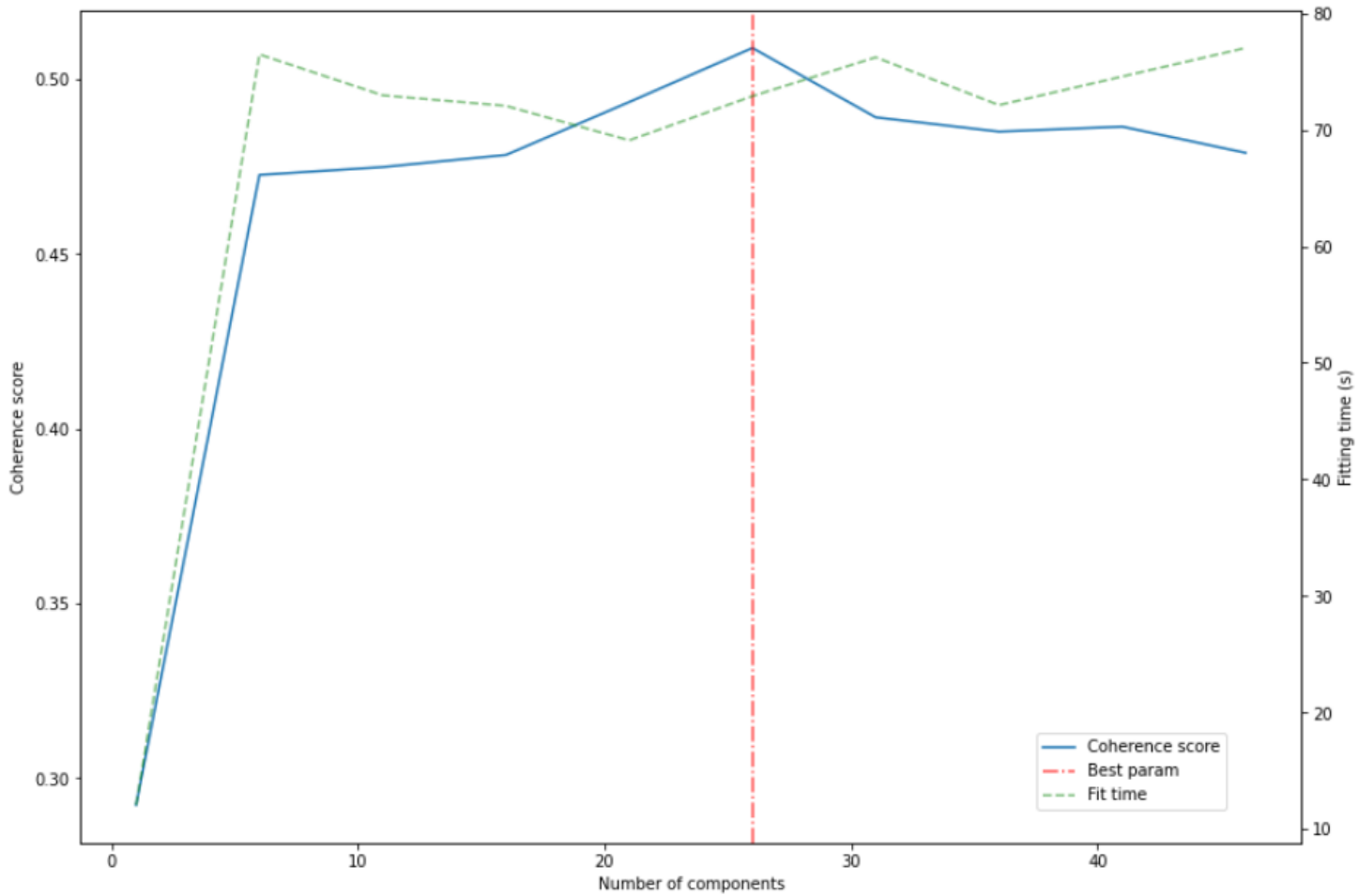
Outline

Latent Dirichlet Allocation

- Entraînement: Actualisation des
  - distributions de probabilités des topics pour chaque document du corpus,
  - distributions de probabilités des mots du vocabulaire pour chaque topic.

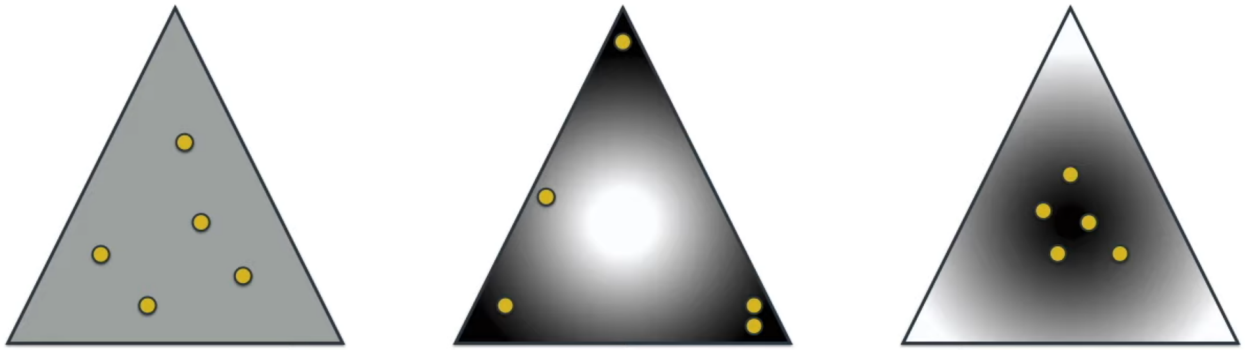
$\Longrightarrow$ *Embedding* des documents dans l'espace des topics.

Nombre de topics optimal



Apprentissage suppervisé "customisé"

- Paramètre α de la LDA



$\alpha = 1$ $\alpha < 1$ $\alpha > 1$

Fonction score

$$s = \Pi_{i=1}^{N} p(w_i/t)p(t/d)$$

- $p(w_i/t)$: probabilité du mot w_i pour le topic dominant t
- $p(t/d)$: probabilité du topic dominant t pour le document d

overflow

"Inférence" du système customisé

	precision	recall	f1-score	support
git	0.826923	0.811321	0.819048	53.0
c++	0.436782	0.413043	0.424581	92.0
css	0.260870	0.807692	0.394366	52.0
html	0.285714	0.426667	0.342246	75.0
asp.net	0.319149	0.348837	0.333333	43.0
python	0.224274	0.469613	0.303571	181.0
c	0.235955	0.411765	0.300000	51.0
php	0.339286	0.267606	0.299213	71.0

Outline

Embedding Word2Vec

• Initialisation aléatoires des poids, puis ajustement par rétropropagation du gradient.

"Combiaison linéaire" entre mots sémantiquement proches

- (King - Man) + (Queen - Woman) = 0 + ϵ
- Le nombre de dimensions est un hyperparamètre ajustable.

$\Longrightarrow$ Utilisation du vocabulaire d'entraînement.

Transfer Learning

BERT

- Transformers avec 12 couches d'encodeurs.
- Vecteur de sortie de dimension 768.

USE

- *Sentence Embdedding*
- Vecteur de sortie de dimension 512.

Résultat: *weighted average*

	Precision	Recall	F1-Score	Support
W2C	0.8483	0.5867	0.6650	1989
BERT	0.8209	0.6515	0.7158	1989
USE	0.6549	0.5057	0.5408	1989

Outline

Résultats

- Word2Vec est le meilleur compromis temps de calcul / espace mémoire.
- Nombre de tags uniques suggérés différents selon les modèles.
- Explicabilité du système customisé.

Merci de votre attetion.