Classification des jeux de données

« Détection de SPAM »

« Détection de polarité dans les tweets français »

Chengwanli YANG

Introduction

Classification automatique

- Catégorisation algorithmique des objets
- Attribution d’une catégorie à chaque objet à classer
- Appel à l'apprentissage automatique
- Nos tâches :
  - « Détection de SPAM » :

    SPAM et HAM
  - « Détection de polarité dans les tweets français » :

    Sentiment négatif et positif

Plan

- Jeux de données
- Méthode
- Résultat
- Discussion
- Bibliographie

Jeux de données

« Détection de SPAM »

- Source : https://raw.githubusercontent.com/Balakishan77/Spam-Email-Classifier/master/spamham.csv
- 2 classes : « ham » valeur 0 et « spam » valeur 1
- 5728 instances, au total 37303 mots
- 1368 instances dans la classe « spam » et 4360 instances dans la classe « ham »
- 4009 instances à entraîner et 1719 instances à tester
- Exemple de « spam » : *Subject: learn to play texas hold ' em and other poker classics on the most popular free site . - earn $ 100 bonus from partypoker . visit here…*
- Exemple de « ham » : *Subject: thank you. I would like to take this brief opportunity to thank you all for inviting me to visit enron …*

Jeux de données

« Détection de polarité dans les tweets français »

- Source : https://www.kaggle.com/hbaflast/french-twitter-sentiment-analysis
- 2 classes : « négatif » valeur 0 et « positif » valeur 1
- 1526724 instances, au total 252741 mots
- 771604 instances dans la classe « négatif » et 755120 instances dans la classe « positif »
- 1068706 instances à entraîner et 458018 instances à tester
- Exemple de « négatif » : *C'est le pire des pires. Je voudrais juste savoir où il se trouve. Si triste, je suis à la maison malade avec la grippe.*
- Exemple de « positif » : *Merci! Je suis content que vous ayez aimé ces photos. Lisez le livre, c'est vraiment bien.*

Méthode

Pré-traitement

- Enlèvement de stopwords :

  Les mots non significatifs figurants dans un texte, ce sont souvent les prépositions, les articles, les pronoms.

- Mise des caractères en minuscule :

  Diminution du nombre de mot à traiter et l'économie de la mémoire.

Méthode

Caractéristique utilisée

- N-gramme de mot :

  Un n-gramme est une sous-séquence de *n* élément(s) construite à partir d'une séquence donnée.

  Les tokens dans le sac sans ordre, pour les traiter mieux avec le bon sens, nous avons besoin de N-gramme.

Méthode

Extraction des caractéristiques

- Sac de mots (Bag of words) :

  Représentation numériquement de nos données contextuelles aux modèles de l'apprentissage automatique, mais il ne prend pas en compte l'ordre et la structure des mots dans le corpus.

- Moyens différents :

  CountVectorizer : la méthode plus utilisée pour la conversion d'une collection du texte en une matrice de nombres de token.

  TfidfVectorizer : convertir en matrice de TF-IDF, il donne une importance plus grande à certains mots, la méthode de pondération.

Méthode

Classifieurs utilisés

- **Perceptron** (Frank Rosenblatt, 1975)

  C’est un algorithme d’apprentissage supervisé, il s'agit d'un neurone formel muni d'une règle d'apprentissage qui permet de déterminer automatiquement les poids synaptiques de manière à séparer un problème d'apprentissage supervisé[1].

- **Naive Bayes** (Thomas Bayes, 1763)

  C’est un type de classification bayésienne probabiliste simple basée sur le théorème de Bayes avec une forte indépendance des hypothèses. Un classifieur bayésien naïf suppose que l'existence d'une caractéristique pour une classe, est indépendante de l'existence d'autres caractéristiques[2].

- **Logistic Regression** (Joseph Berkson)

  C’est un modèle de régression binomiale, il s'agit de modéliser au mieux un modèle mathématique simple à des observations réelles nombreuses[3].

---

[1] https://fr.wikipedia.org/wiki/Perceptron#Règle_d'apprentissage_du_perceptron_(loi_de_Widrow-Hoff)
[2] https://fr.wikipedia.org/wiki/Classification_naïve_bayésienne#Classification_et_catégorisation_de_documents
[3] https://fr.wikipedia.org/wiki/Régression_logistique

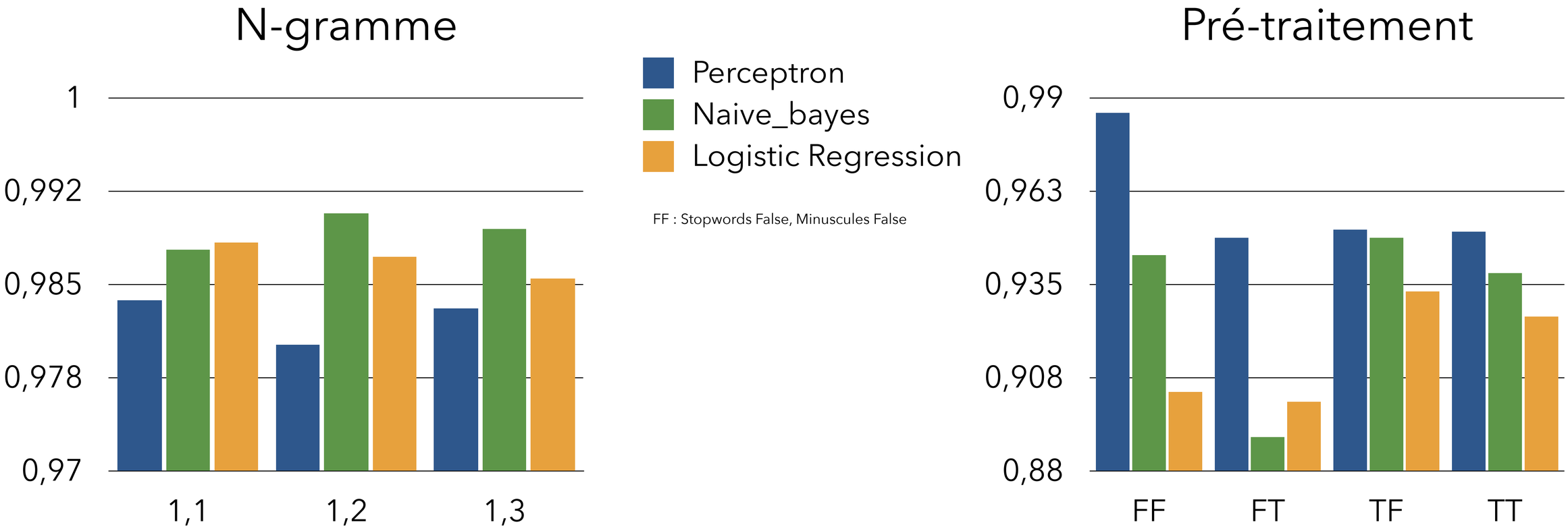
Résultat

Etapes

1. Choisir le n-gramme
2. Comparer les pré-traitements
3. Obtenir le résultat

Résultat

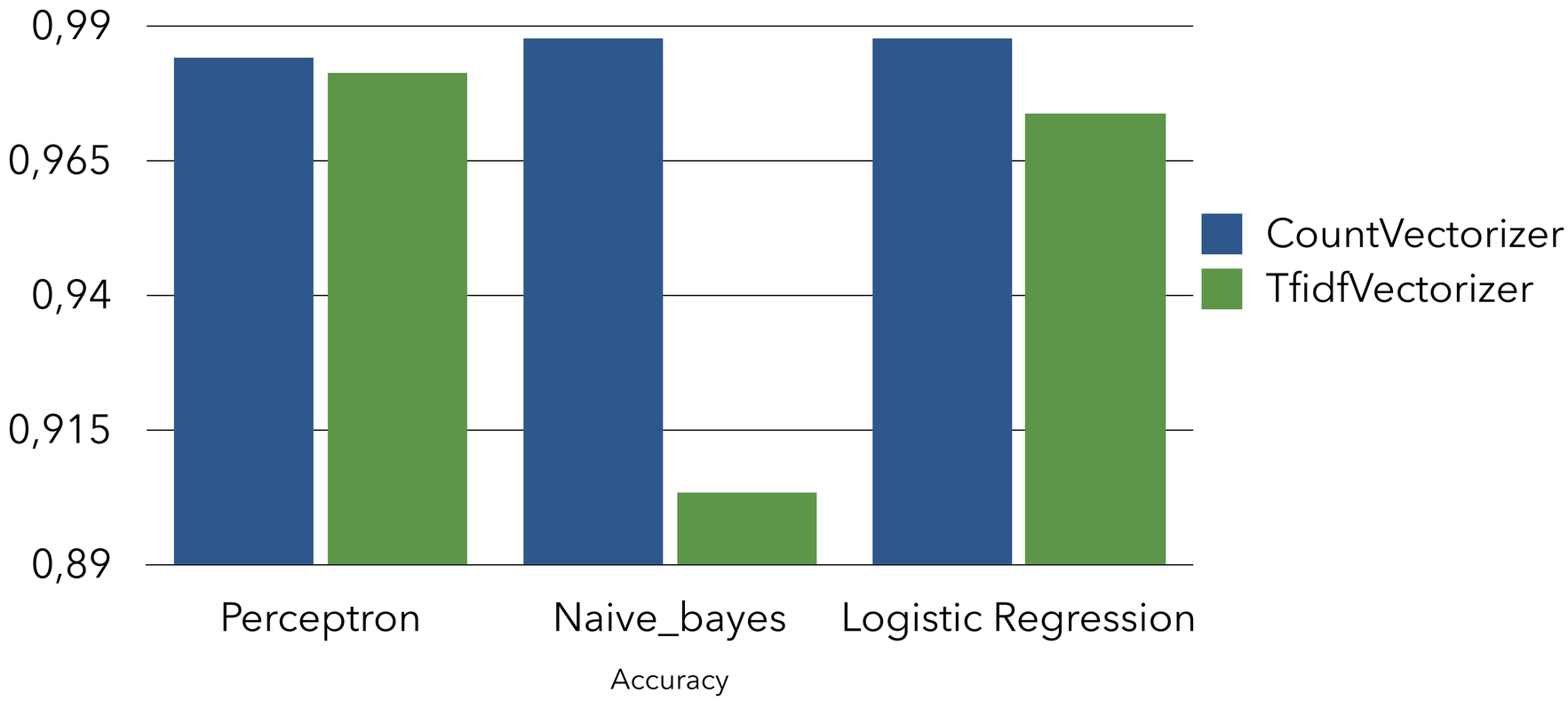
« Détection de SPAM »



Meilleure condition : Unigramme, Stopwords True, Minuscules False

Résultat

« Détection de SPAM »



Résultat

« Détection de SPAM »

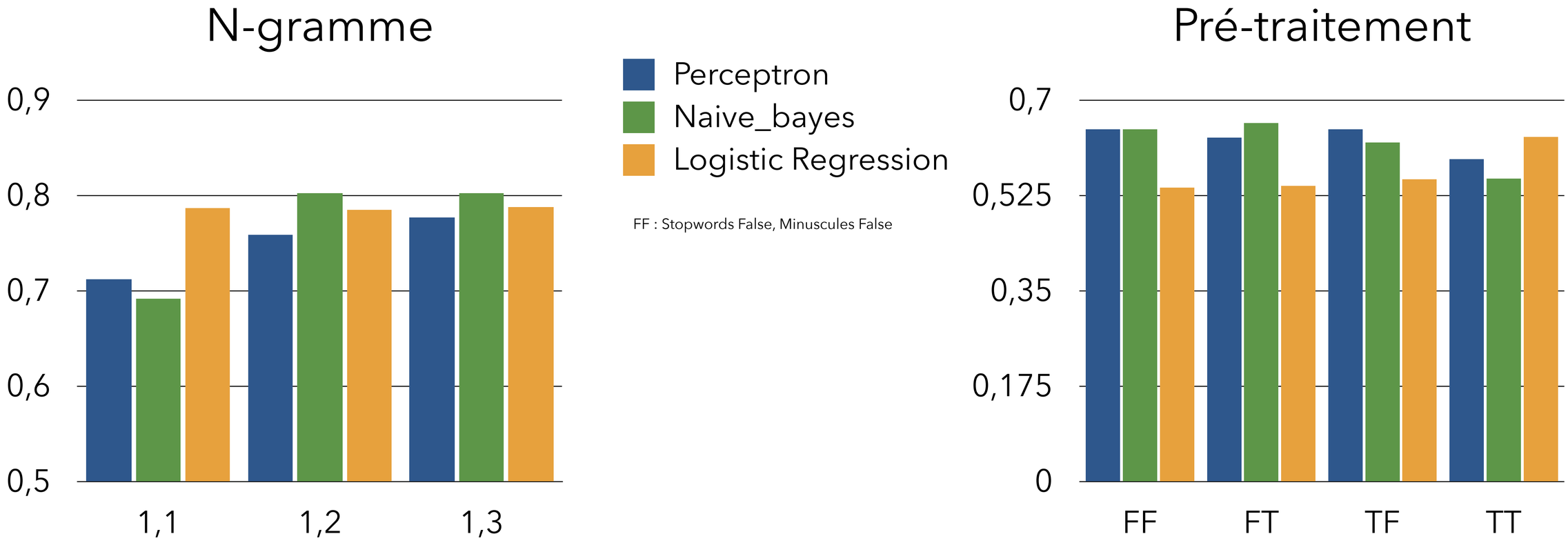
CountVectorizer						
		precision	recall	f1-score	confusion matrix	
					ham	spam
Perceptron	ham	0.99	0.98	0.99	1294	7
	spam	0.95	0.98	0.97	20	398
naive_bayes	han	1	0.99	0.99	1298	5
	spam	0.96	0.99	0.97	16	400
Logistic Regression	han	0.99	0.99	0.99	1304	11
	spam	0.98	0.97	0.97	10	394
TfidfVectorizer						
		precision	recall	f1-score	confusion matrix	
					ham	spam
Perceptron	ham	0.99	0.99	0.99	1301	19
	spam	0.97	0.95	0.96	13	386
naive_bayes	han	0.89	1	0.94	1313	165
	spam	1	0.59	0.74	1	240
Logistic Regression	han	0.97	1	0.98	1313	44
	spam	1	0.89	0.94	1	361

Tableau 1. Résultat de « spamham »

- Marche bien
- 1 n’est pas 100%
- Naive bayes avec TfidfVectorizer moins bon
- Total : 1314 ham, 405 spam

Résultat

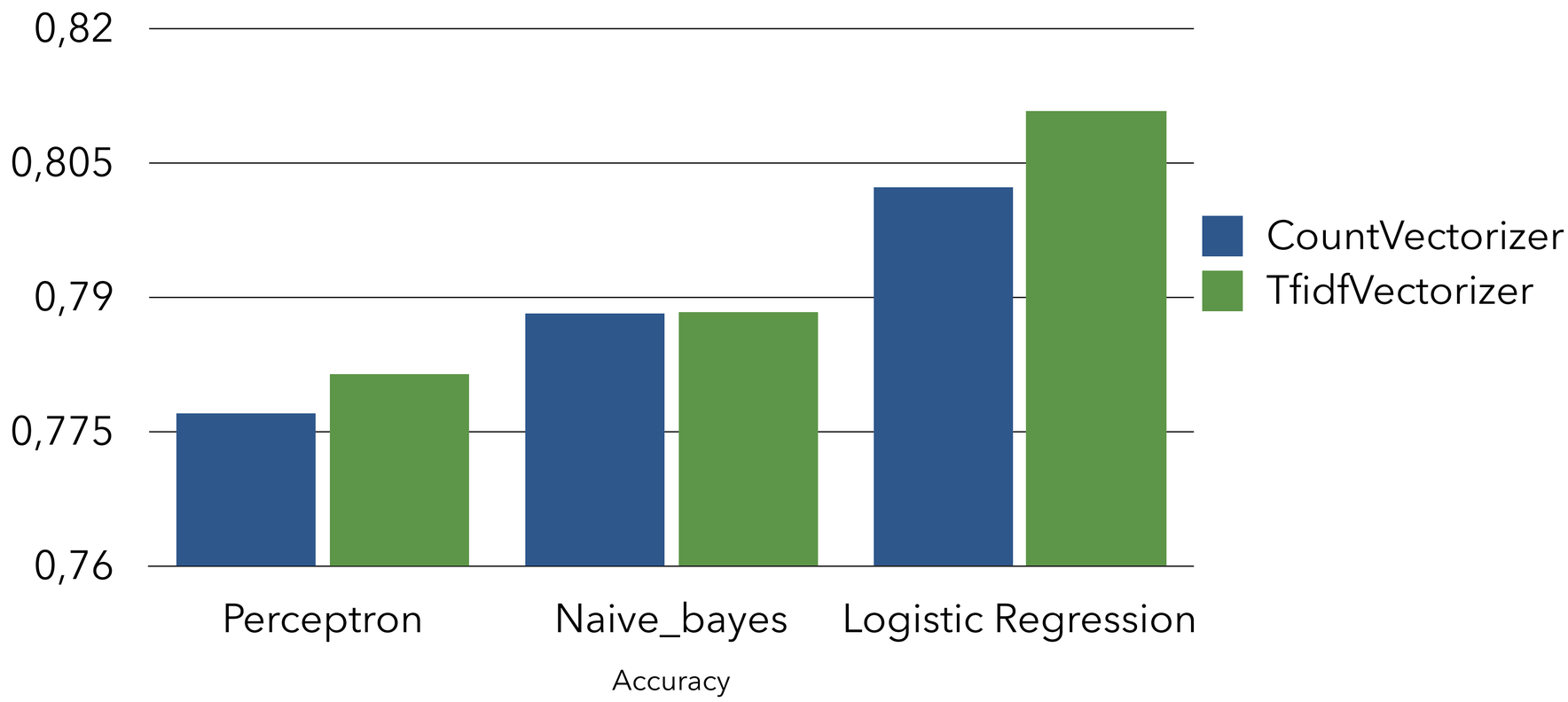
« Détection de polarité dans les tweets français »



Meilleure condition : Unigramme, bigramme et trigramme, Stopwords False Minuscules True

Résultat

« Détection de polarité dans les tweets français »



Résultat

« Détection de polarité dans les tweets français »

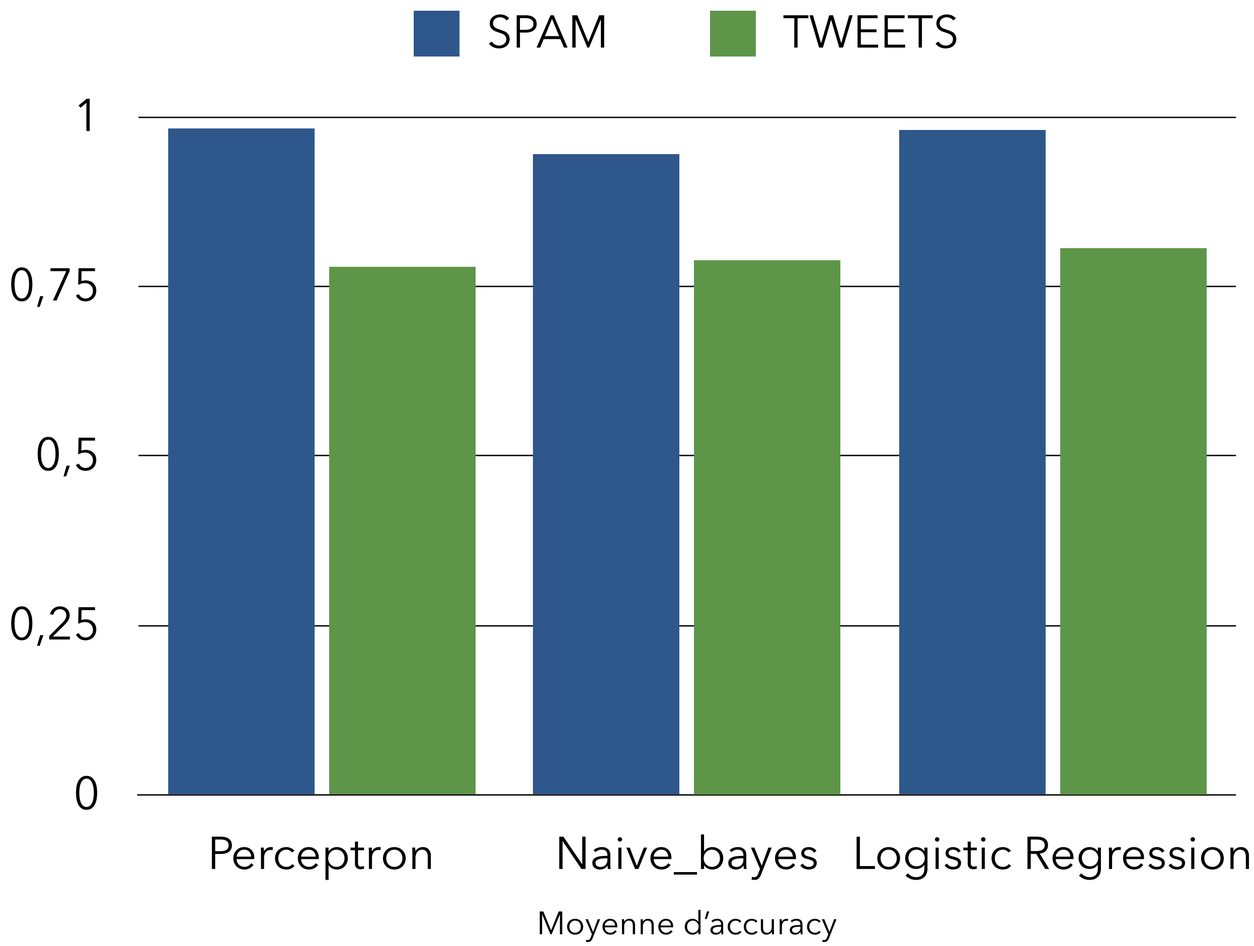
CountVectorizer						
		precision	recall	f1-score	confusion matrix	
					negatif	positif
Perceptron	negatif	0.79	0.75	0.77	174112	44991
	positif	0.76	0.8	0.78	57102	181813
naive_bayes	negatif	0.77	0.83	0.8	191438	57187
	positif	0.81	0.75	0.78	39776	169617
Logistic Regression	negatif	0.81	0.8	0.8	184399	43716
	positif	0.8	0.81	0.8	46815.	183088
TfidfVectorizer						
		precision	recall	f1-score	confusion matrix	
					negatif	positif
Perceptron	negatif	0.8	0.76	0.78	174609	43455
	positif	0.76	0.81	0.79	56605	183349
naive_bayes	negatif	0.76	0.86	0.8	198635	64317
	positif	0.83	0.72	0.77	32579	162487
Logistic Regression	negatif	0.82	0.8	0.81	185041	40499
	positif	0.8	0.82	0.81	46173	186305

Tableau 2. Résultat de « tweets »

- Marche moins bien
- TfidfVectorizer
- Logistic Regression
- Total : 231214 négatif, 226804 positif

Discussion

Résultat



Moyenne d'accuracy

- Classifieurs marchent bien
- SPAM > TWEETS
- Logistic Regression

Discussion

Amélioration

- Augmenter la pureté du sac de mots :

  Enlèvement maximum des éléments dont nous n'avons pas besoin.

  Exemple : *Ça va être un looooooonnnnnnnggggggggg nuit au travail.*

- Pour d'autres langues :

  Langues en lettres latines (italien, espagnol…) : cette méthode pourrait marcher.

  Langues en caractère (chinois, coréen…) : pas suffisante. Puisque les mots de ces langues ne sont pas séparés par l'espace, la méthode split ne fonctionne plus. Il est nécessaire d'importer la méthode spécifique pour découper le token (Jieba pour le chinois).

Bibliographie

- Rémy Kessler, Juan Manuel Torres-Moreno, Marc El-Bèze. (s. d.). Classification automatique de courriers électroniques par des méthodes mixtes d’ apprentissage. juanmanuel.torres.free.fr. Consulté le 29 décembre 2020, à l’adresse http://juanmanuel.torres.free.fr/downloads/RSTI_ISI_kessler_torres_elbeze.pdf
- S. Seth and S. Biswas, "Multimodal Spam Classification Using Deep Learning Techniques," 2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Jaipur, 2017, pp. 346-349, doi: 10.1109/SITIS.2017.91.
- M. Kanakaraj and R. M. R. Guddeti, "NLP based sentiment analysis on Twitter data using ensemble classifiers," 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN), Chennai, 2015, pp. 1-5, doi: 10.1109/ICSCN.2015.7219856.
- Benamara, Farah and Grouin, Cyril and Karoui, Jihen and Moriceau, Véronique and Robba, Isabelle Analyse d'opinion et langage figuratif dans des tweets : présentation et résultats du Défi Fouille de Textes DEFT2017. (2017) In: Atelier TALN 2017 : Défi Fouille de Textes (DEFT 2017), 26 June 2017 - 26 June 2017 (Orléans, France).
- Aurélien Massiot, Léa Naccache. (2020, 5 mai). NLP : une classification multilabels simple, efficace et interprétable. www.blog.octo.com. https://blog.octo.com/nlp-une-classification-multilabels-simple-efficace-et-interpretable/