# PROJET AMAZON : ANALYSE DE SENTIMENTS

## Soutenance d'Introduction au Machine Learning

Mohamed NIANG - Hoang Dung NGUYEN - Yao GNONSOU
**Enseignante** : Pr. Agathe GUILLOUX

Master 2 Data Science : Université Paris-Saclay

23 octobre 2019

# PLAN

# Introduction

Dans ce projet, nous fessons face à un problème de classification supervisé sur des commentaires récoltés sur amazon.

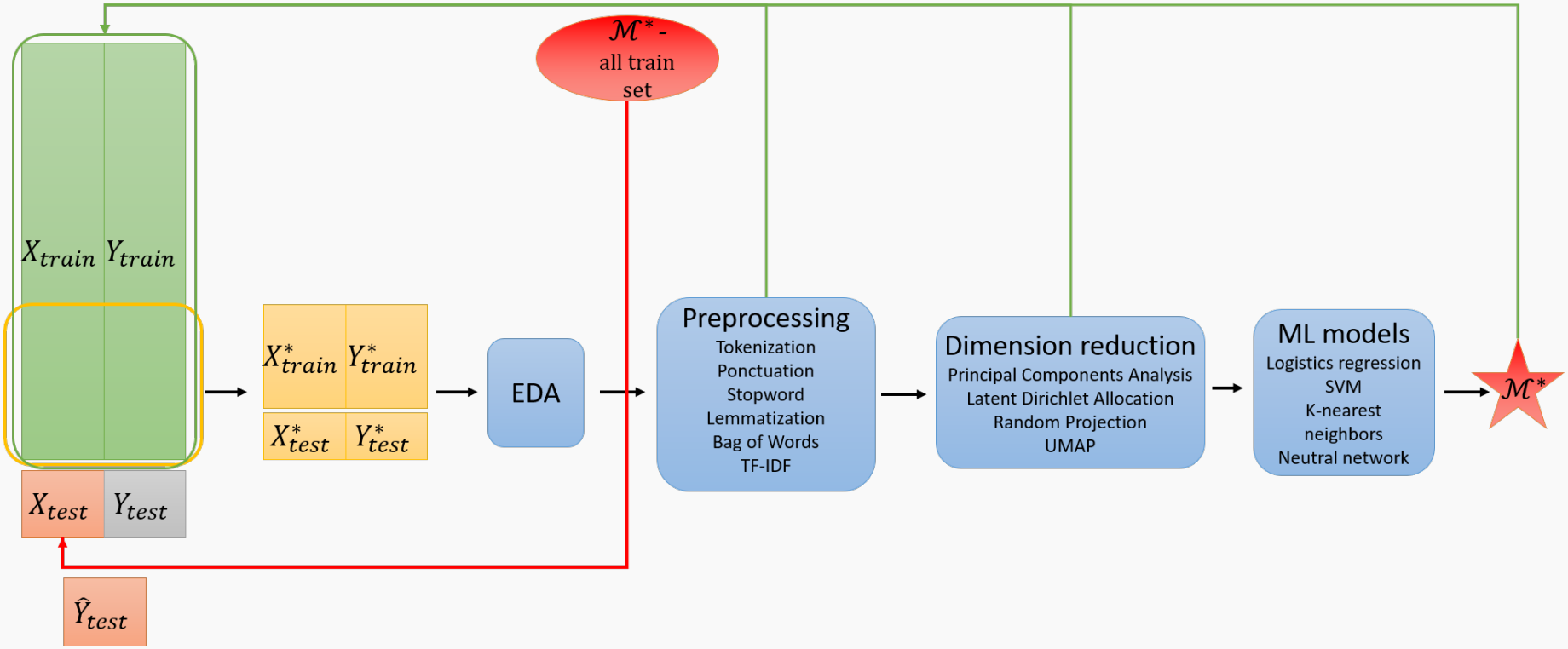
Le but ici est de traiter puis d'implémenter différent modèles afin de déterminer le meilleur modèle pour faire de la prédiction sur de nouveaux commentaires.

Les principaux modèles utilisés dans ce projet sont :

- Logistic Regression
- Support vector machine
- K-nearest neighbors
- Neural network

# Méthodologie Suivie

Notre plan de travail est résumé par le schéma ci-après :



# Description des données

```
df_train.describe(include='all')
```

| | label | text |
|---|---|---|
| count | 3600000 | 3600000 |
| unique | 2 | 3600000 |
| top | __label__1 | the hands do not glow in dark: the quality of ... |
| freq | 1800000 | 1 |

```
df_train['text'][2]
```

```
'Amazing!: This soundtrack is my favorite music of all time, hand
s down. The intense sadness of "Prisoners of Fate" (which means a
ll the more if you\'ve played the game) and the hope in "A Distan
```

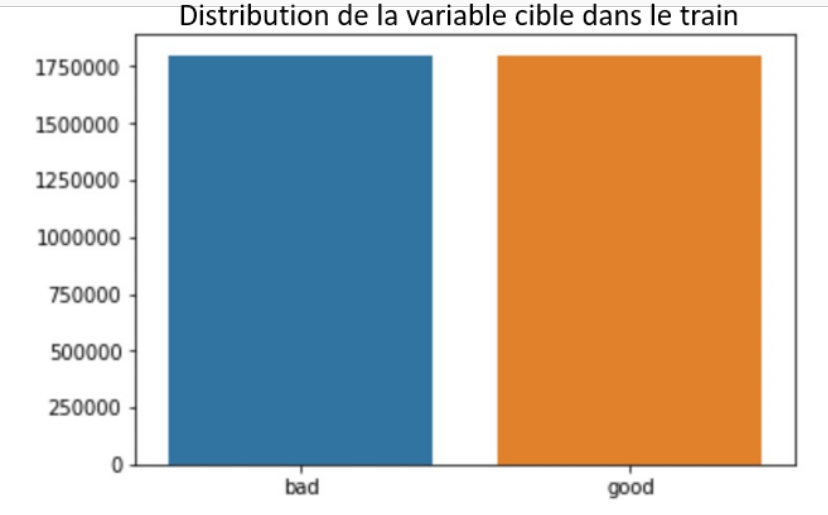
```
df_train['text'][10]
```

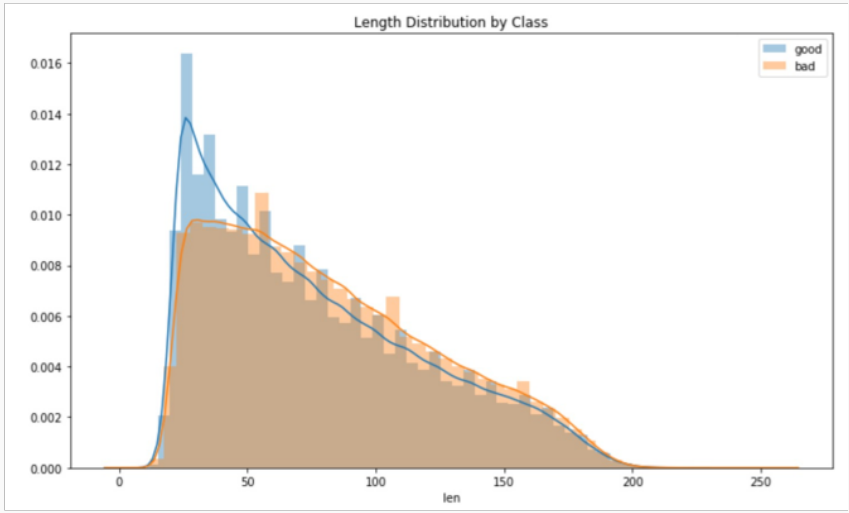
```
"The Worst!: A complete waste of time. Typographical errors, poor
grammar, and a totally pathetic plot add up to absolutely nothin
g. I'm embarrassed for this author and very disappointed I actual
ly paid for this book."
```

```
df_train.head(20)
```

| | label | text |
|---|---|---|
| 0 | __label__2 | Stuning even for the non-gamer: This sound tra... |
| 1 | __label__2 | The best soundtrack ever to anything.: I'm rea... |
| 2 | __label__2 | Amazing!: This soundtrack is my favorite music... |
| 3 | __label__2 | Excellent Soundtrack: I truly like this soundt... |
| 4 | __label__2 | Remember, Pull Your Jaw Off The Floor After He... |
| 5 | __label__2 | an absolute masterpiece: I am quite sure any o... |
| 6 | __label__1 | Buyer beware: This is a self-published book, a... |
| 7 | __label__2 | Glorious story: I loved Whisper of the wicked ... |
| 8 | __label__2 | A FIVE STAR BOOK: I just finished reading Whis... |
| 9 | __label__2 | Whispers of the Wicked Saints: This was a easy... |
| 10 | __label__1 | The Worst!: A complete waste of time. Typograp... |
| 11 | __label__2 | Great book: This was a great book,I just could... |
| 12 | __label__2 | Great Read: I thought this book was brilliant,... |

# Description des données




## Remarque

- Une proportion équilibrée entre les deux labels
- Les distributions de taille par label sont légèrement différentes

# Description des données

Exemple du texte qui est marqué "good" :

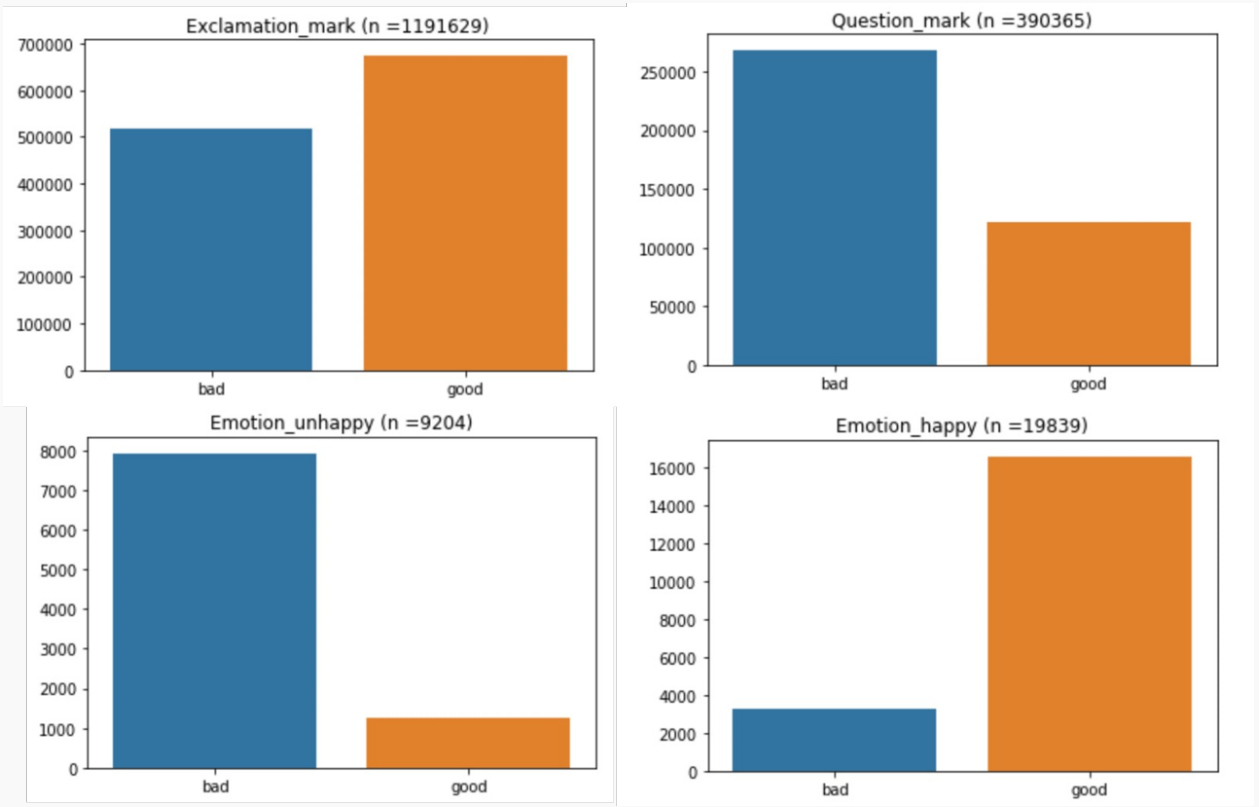
Great desk and a great buy ! Thanks Amazon :) : We bought this desk for my seven year old daughter to keep her lap top on and to do homework. It was not to disappoint...

Exemple du texte qui est marqué "bad" :

Awful beyond belief ! : I feel I have to write to keep others from wasting their money. This book seems to have been written by a 7th grader with poor grammatical skills for her age ! As another reviewer points out, there is a misspelling on the cover, and I believe there is at least one per chapter...

# Description des données

On observe maintenant la distribution de quelques ponctuations spéciales :




En conclure, on conservera ces ponctuations lors de la tokenzation.

# Analyse de l'intensité des sentiments

## Avec l'algorithme VADER (Valence Aware Dictionary and Sentiment Reasoner)

VADER est un modèle utilisé pour l'analyse des sentiments textuels qui est sensible à la polarité (positive/négative) et à l'intensité (force) des émotions. Introduite en 2014, l'analyse du sentiment textuel VADER utilise une approche centrée sur l'humain, combinant l'analyse qualitative et la validation empirique en utilisant des évaluateurs humains et la sagacité des autres.

This presentation is really good !

| Compound | Negative | Neutral | Positive |
|---|---|---|---|
| 0.54 | 0.0 | 0.534 | 0.466 |

This presentation is awful!!!

| Compound | Negative | Neutral | Positive |
|---|---|---|---|
| -0.5962 | 0.564 | 0.436 | 0.0 |

This presentation is not good, but it satisfies me

| Compound | Negative | Neutral | Positive |
|---|---|---|---|
| 0.4583 | 0.137 | 0.564 | 0.298 |

This presentation is normal

| Compound | Negative | Neutral | Positive |
|---|---|---|---|
| 0.0 | 0.0 | 1.0 | 0.0 |

Nous voyons donc la première prédiction obtenue par cette algorithme

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| bad | 0.85 | 0.49 | 0.62 | 5097 |
| good | 0.63 | 0.91 | 0.75 | 4903 |
| accuracy | | | 0.70 | 10000 |
| macro avg | 0.74 | 0.70 | 0.69 | 10000 |

# Détection de langues

On vient ici d'identifier des commentaires qui sont rédigés en autres langages.

| Language | Frequency |
|---|---|
| en | 9970 |
| es | 16 |
| fr | 8 |
| de | 3 |
| it | 1 |
| cy | 1 |
| id | 1 |

| | text | label_new | len | lang |
|---|---|---|---|---|
| 881 | Good read..... ..... ..... | good | 35 | cy |
| 1249 | Il grande ritorno!: E' dai tempi del tour di "... | good | 153 | it |
| 1259 | La reencarnación vista por un científico: El p... | good | 34 | es |
| 1260 | Excelente Libro / Amazing book!!: Este libro h... | good | 105 | es |
| 1261 | Magnifico libro: Brian Weiss ha dejado una mag... | good | 47 | es |
| 1639 | El libro mas completo que existe para nosotras... | good | 29 | es |
| 1745 | Excelente!: Una excelente guía para todos aque... | good | 49 | es |
| 2316 | Nightwish is unique and rocks for eva: Moi to ... | good | 47 | fr |
| 2486 | Palabras de aliento para tu caminar con Dios: ... | good | 80 | es |
| 2760 | Completement nul: Fait sur commande et ennuyan... | bad | 18 | fr |
| 2903 | fabuloso: mil gracias por el producto fabuloso... | good | 22 | es |
| 2908 | Geh: Blah blah, sexy girl, blah blah, fighting... | bad | 21 | id |
| 3318 | Excelentes botas.. excelentes boots: Excelente... | good | 31 | es |
| 3694 | Why not Spanish ???: Alguien me puede decir po... | bad | 103 | es |
| 4144 | LEAKED FIRST DAY FOR MY GUEST: IT HAD A LEAK F... | bad | 27 | de |
| 4820 | La mejor película de Moore: A mi juicio, esta ... | good | 21 | es |
| 4914 | De la poudre aux yeux: J'ai acheté un Sansa Vi... | bad | 65 | fr |
| 5720 | C'est magnifique! il y a du vrai dans ce qui'l... | good | 96 | fr |
| 5875 | Erreur: "Les Triplettes de Belleville" n'a pas... | good | 25 | fr |

Cependant, en terme de proportion cela ne représente que 0,31% des commentaires (sur 10.000 obs). Pour la suite on décide de ne pas retenir ce traitement car il demande énormément de temps de calcul.

# Notre tokenizer

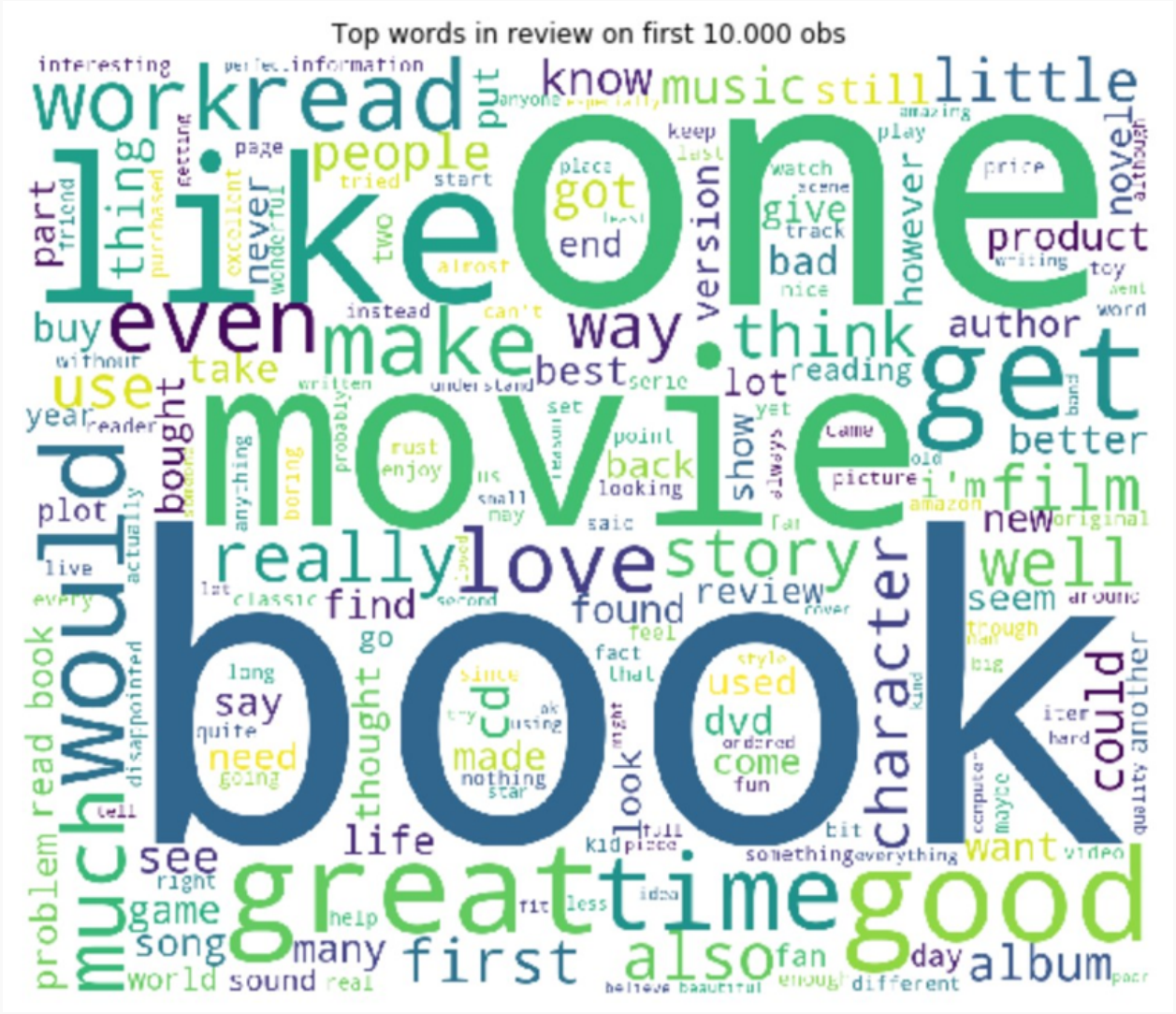
Pour cette étape, nous faisons :

- Tokenizer les text
- Enlever les ponctuations (en conservant les ponctuations spéciales)
- Enlever les stopwords
- Faire une lemmatisation avec part of speach

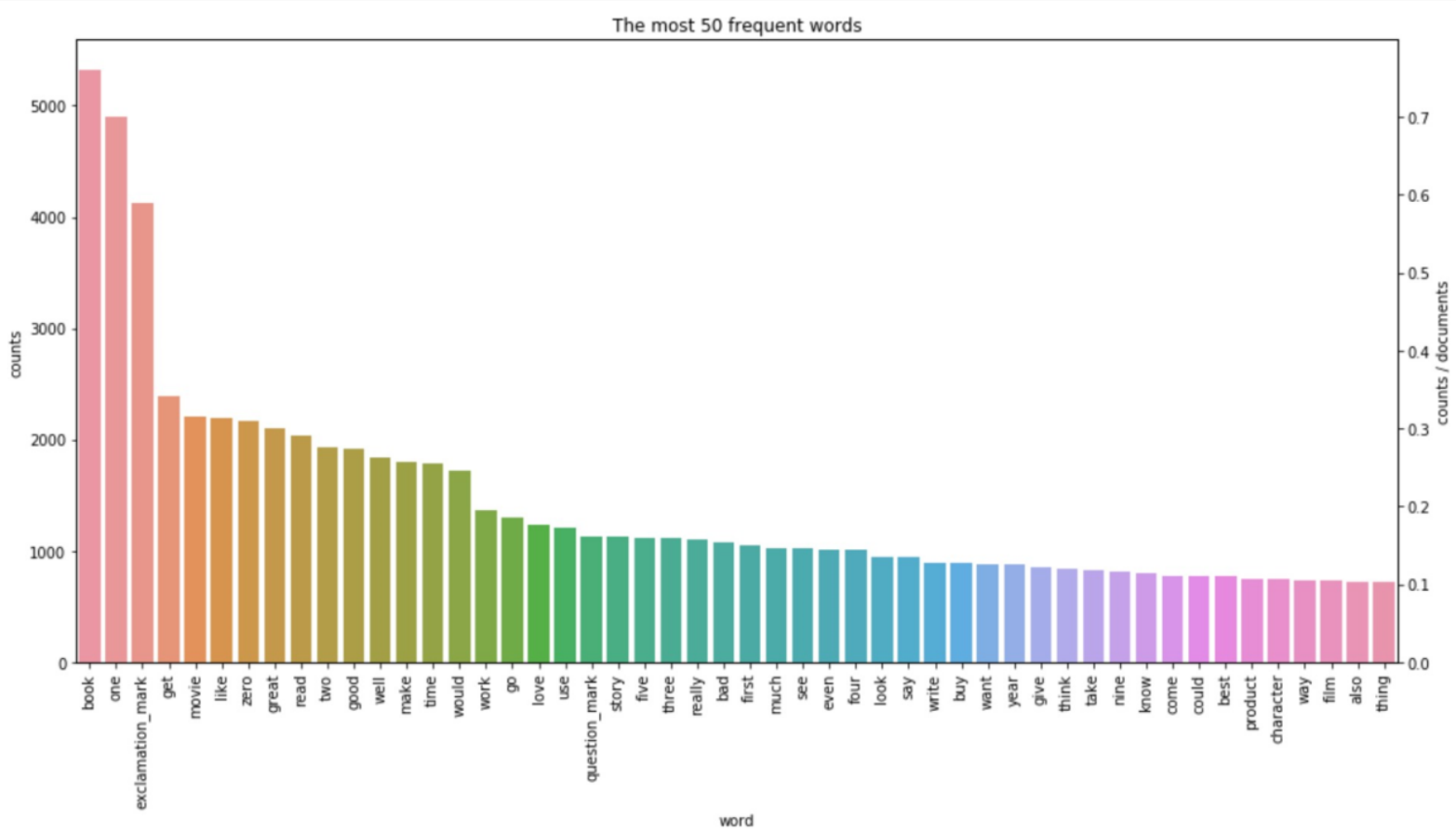
```
print('Pharse originale : \n')
print(df_train['text'][979],'\n')
print('Pharse transformée: \n')
print(' '.join(text_process(df_train['text'][979])),'\n')
print('Liste de mots (Tokenization): \n')
text_process(df_train['text'][32])
```

```
Pharse originale :

Great desk and a great buy! Thanks Amazon:): We bought this desk for my seven year old daughter to keep her lap top
on and to do homework. It was not to disappoint. The desk is perfect size for her to sit and work. Very high qualit
y and easy to put together. I would highly recommend this product to others?

Pharse transformée:

great desk great buy exclamation_mark thanks amazon emotion_happy bought desk seven year old daughter keep lap top
homework disappoint desk perfect size sit work high quality easy put together would highly recommend product others
question_mark

Liste de mots (Tokenization):


['title',
 'hollywood',
 'debacle',
 'plot',
 'ridiculous',
 'wonder',
 'even',
 'read',
```

# Fréquence d'apparition des mots



# Fréquence d'apparition des mots

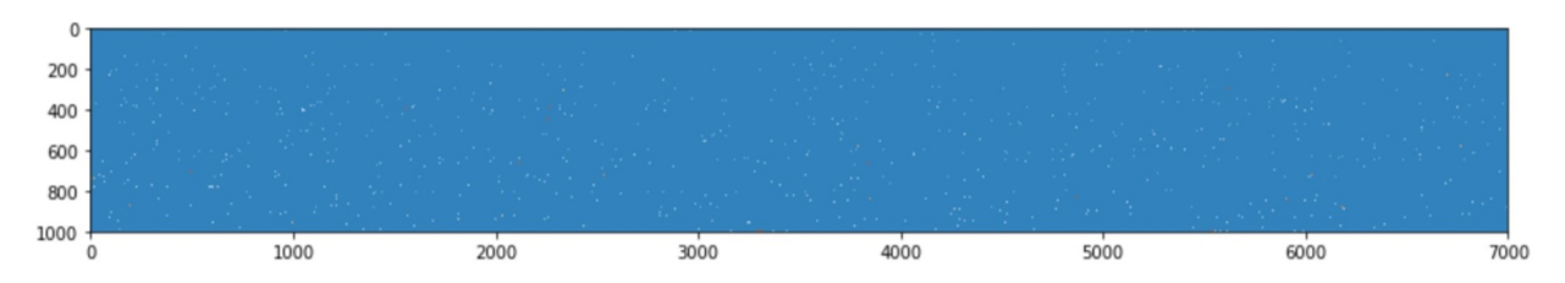



| Count | Mean | Std | Min | Quantile 25% | Quantile 50% | Quantile 75% | Max |
|---|---|---|---|---|---|---|---|
| 20611 | 14,45 | 94,36 | 1 | 1 | 2 | 5 | 5326 |

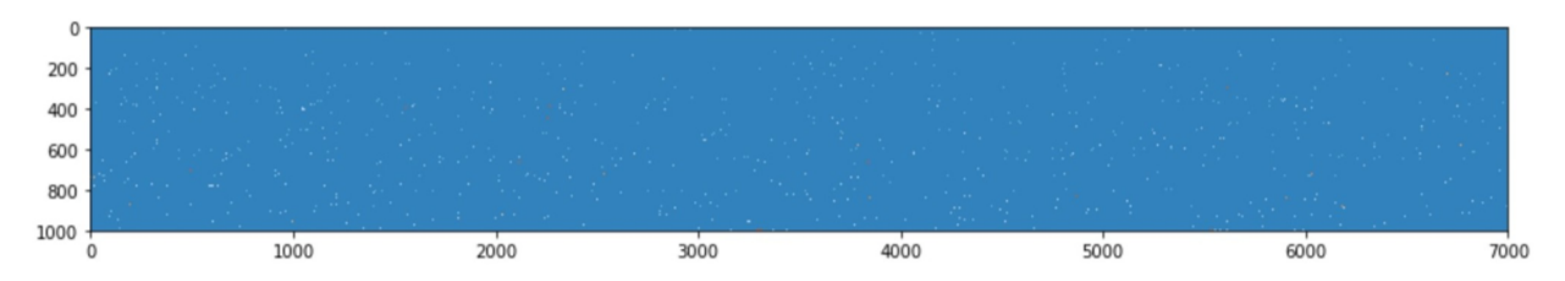
| Number of words appearing only 1 time | Number of frequencies of the 10000th word |
|---|---|
| 9730 | 2 |

# BoWs vs TF-IDF

## Le poid des mots dans la matrice de Bag of Words




## Le poid des mots dans la matrice de TF-IDF



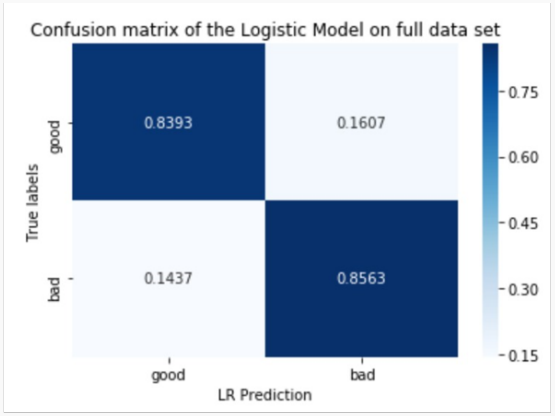
# Réduction de la dimensionnalité

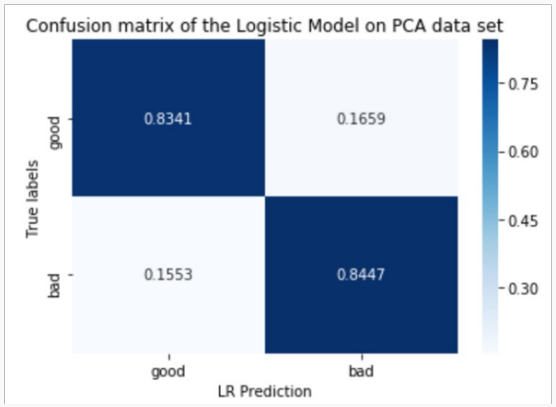
| Dimension reduction methods | Computing time (in second) | Number of new features |
|---|---|---|
| Principal component analysis (PCA) | 1 | 200 |
| Random projection (RP) | 13 | 20 |
| Latent Dirichet Allocation (LDA) | 52 | 100 |
| Uniform Manifold Approximation and Projection (UMAP) | 73 | 20 |

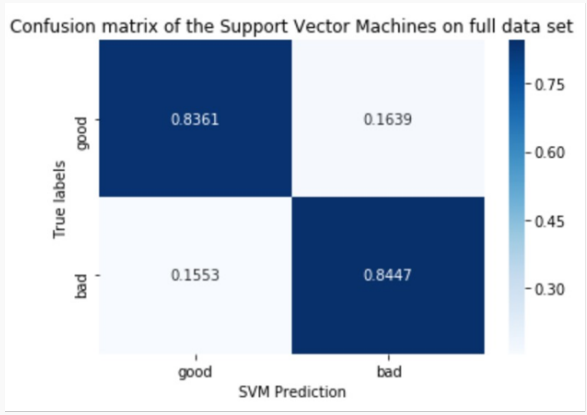
# Modèles de ML

| Machine Learning model | Reduction method | Accuracy score | Computing time (in second) | Computing time added dimension reduction time (in second) | Estimated computing time on all training set (in second) |
|---|---|---|---|---|---|
| Logistic Regression | Full matrix | 0,848 | 0,35 | 0,35 | 127 |
| Logistic Regression | PCA | 0,837 | 0,23 | 1,38 | 497 |
| Logistic Regression | LDA | 0,583 | 0,20 | 14,10 | 5 074 |
| Logistic Regression | RP | 0,664 | 0,48 | 53,38 | 19 216 |
| Logistic Regression | UMAP | 0,635 | 0,20 | 73,30 | 26 389 |
| SVM | Full matrix | 0,839 | 9,32 | 9,32 | 3 354 |
| SVM | PCA | 0,836 | 9,04 | 10,19 | 3 669 |
| SVM | LDA | 0,582 | 5,91 | 19,80 | 7 130 |
| SVM | RP | 0,664 | 10,81 | 63,71 | 22 935 |
| SVM | UMAP | 0,642 | 3,67 | 76,77 | 27 639 |
| K-nearest neighbors | Full matrix | 0,641 | 1,62 | 1,62 | 582 |
| K-nearest neighbors | PCA | 0,651 | 5,05 | 6,19 | 2 230 |
| K-nearest neighbors | LDA | 0,520 | 1,75 | 15,64 | 5 631 |
| K-nearest neighbors | RP | 0,598 | 9,20 | 62,10 | 22 356 |
| K-nearest neighbors | UMAP | 0,618 | 0,24 | 73,34 | 26 403 |
| Neural network | Full matrix | 0,840 | 3,98 | 3,98 | 1 434 |
| Neural network | PCA | 0,826 | 2,76 | 3,91 | 1 407 |
| Neural network | LDA | 0,566 | 2,65 | 16,55 | 5 957 |
| Neural network | RP | 0,668 | 3,29 | 56,19 | 20 228 |
| Neural network | UMAP | 0,654 | 2,90 | 76,00 | 27 359 |

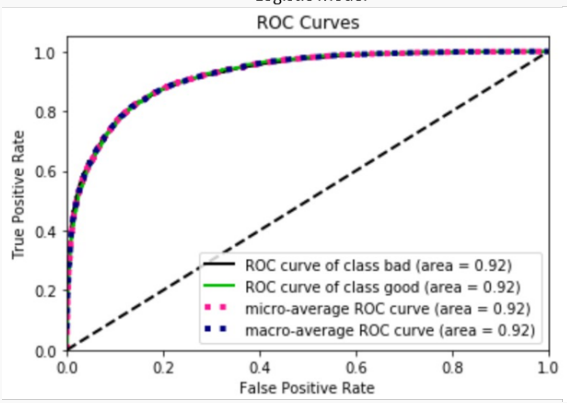
# Modèles de ML



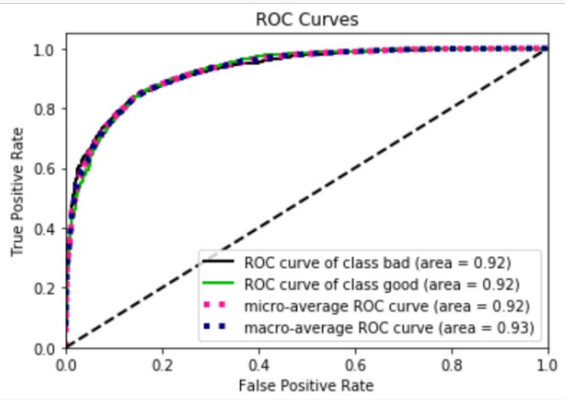



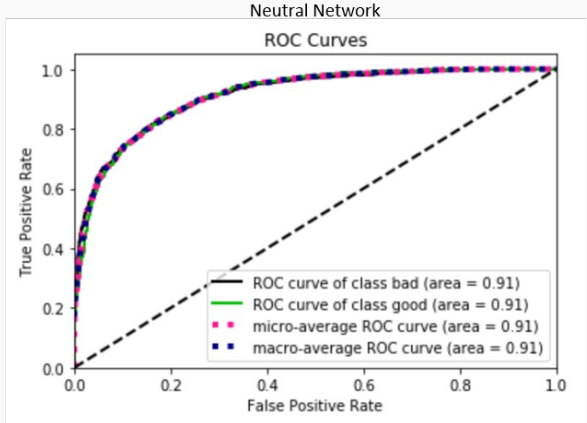

Logistic Model




SVM




Neutral Network



# Conclusion

## Perspectives

- Implémentation de la logistic regression
- Gestion du problème de big data
- Compréhension de la procédure de traitement de texte

# Conclusion

## Piste d'amélioration

- Augmentation de taille du jeu d'apprentissage
- Test de robustesse du modèle optimal sur plusieurs sous échantillon
- Test de différentes procédures de traitement de texte : N-grams, Vec2Doc, etc

# Thank you for attention !