

# Prédiction de notes tripadvisor vF

## Projet d'initiation aux différents outils

Paul FAVIER

2025-01-27

# Table des matières

- 1 Introduction
- 2 Analyse exploratoire des données
- 3 Estimation par taille de commentaire
- 4 Répartition des longueurs de commentaires parmi les notes
- 5 Observation de la fréquence d'apparition des différents mots dans le corpus
- 6 Création d'un modèle d'estimation des notes
- 7 Conclusion

## Section 1

### Introduction

# Introduction

## Objectif

**Officiellement** ce projet a pour objectif d'explorer la relation qu'il existe entre un commentaire et la note que les clients laissent sur Tripadvisor.

**Officieusement** ce projet était surtout un cas pratique me permettant de reprendre en main différents outils tels que R et Rmd.

# Introduction

## Base de données

La base de données est composée de 147 579 commentaires d'hôtel avec la note associée qui proviendraient de Tripadvisor (source de la base de données sur [Kaggle](#)).

# Introduction

## Méthodologies

Deux méthodologies ont été mises en place :

- estimation par taille de commentaire ;
- estimation par fouille de texte (*text mining*).

## Section 2

# Analyse exploratoire des données

# La base de données

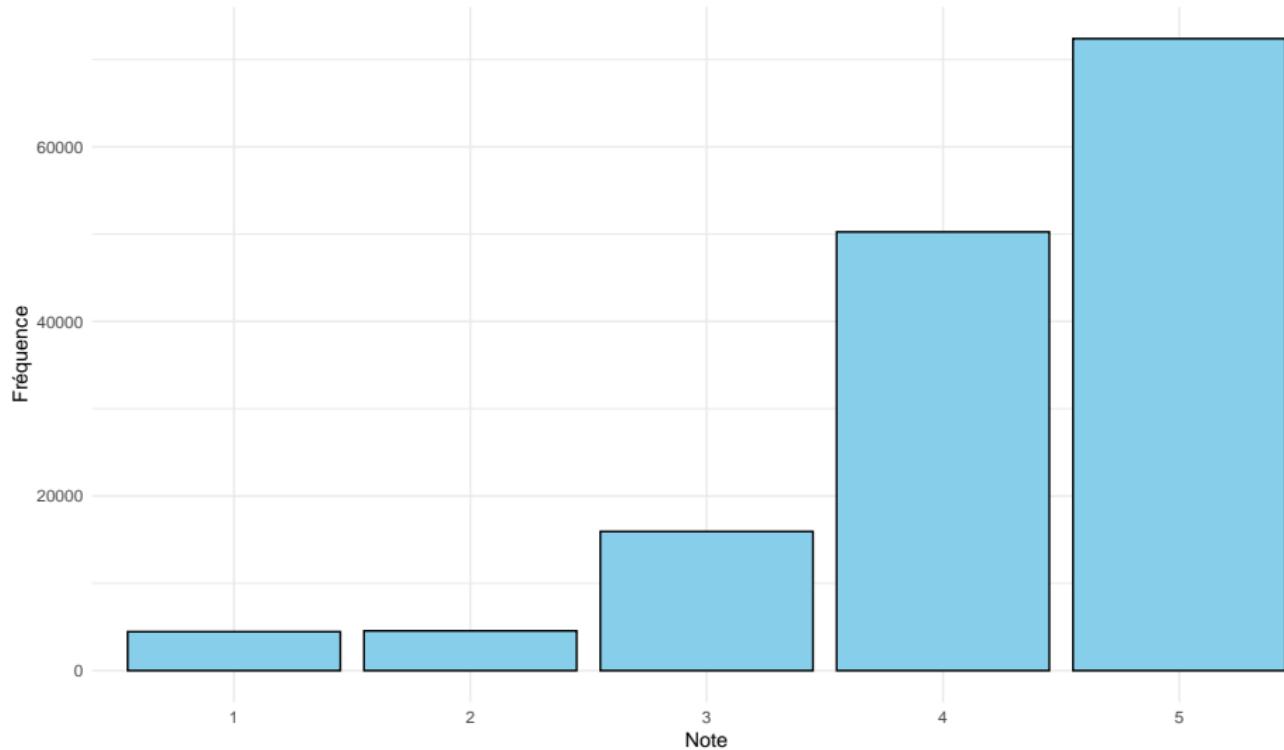
```
##      Review          Rating
##  Length:147579    Min.   :1.00
##  Class  :character 1st Qu.:4.00
##  Mode   :character Median  :4.00
##                  Mean   :4.23
##                  3rd Qu.:5.00
##                  Max.   :5.00
```

## Exemples de commentaires

```
## # A tibble: 6 x 2
##   review                               Rating
##   <chr>                                <dbl>
## 1 Totally in love with the Auro of the pla      5
## 2 I went this bar 8 days regularly with my      5
## 3 We were few friends and was a birthday c      5
## 4 Fatjar Cafe and Market is the perfect pl      5
## 5 Hey Guys, if you are craving for pizza a      5
## 6 We were looking for a special meal and w      5
```

# La répartition des notes

Distribution des notes



# La répartition des notes

```
##    Rate   Freq
## 1    1  4455
## 2    2  4552
## 3    3 15935
## 4    4 50248
## 5    5 72389
```

## Section 3

### Estimation par taille de commentaire

## Estimation par taille de commentaire

L'idée est de regarder comment est répartie la longueur des commentaires (par nombre de caractères et par nombre de mots) afin de créer quatre classes de même taille ("très petit", "petit", "grand", "très grand").

```
summary(df$character_number)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##      6.0   163.0  263.0    404.7  462.0  8192.0
```

```
summary(df$words_number)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##      1.0   29.0   47.0    72.7   84.0  1609.0
```

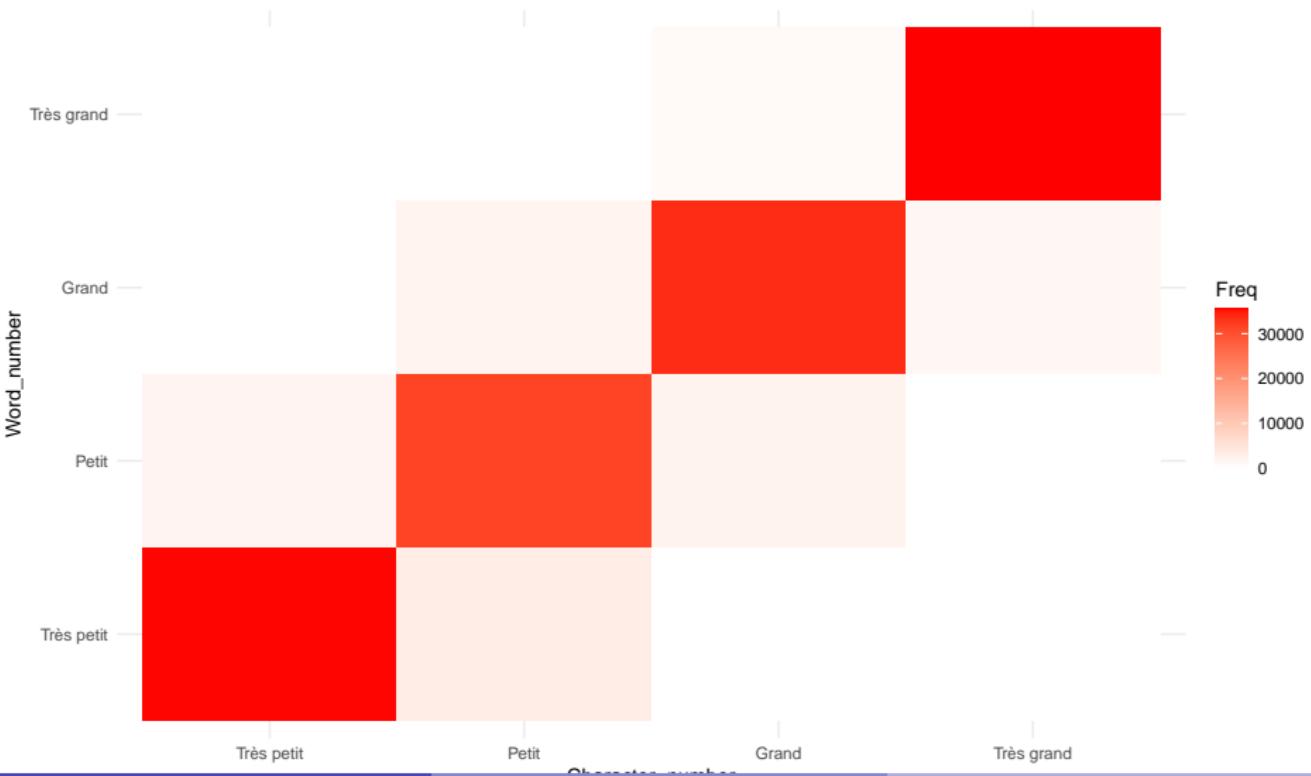
# Estimation par taille de commentaire

## Potentielle limite

J'ai décidé arbitrairement, sur la base des quartiles *infra*, quelles seraient les bornes des classes. Un autre échantillon aurait probablement donné des catégories bornées par des nombres de caractères et de mots différents.

# Vérifier la cohérence

Après avoir construit des catégories de taille de commentaires selon deux critères, je regarde si ces indicateurs convergent.



## Vérifier la cohérence

On constate dans le tableau *infra* que la grande majorité des commentaires emploient des mots d'une taille moyenne comprise entre 5 et 7 caractères.

|    |    |     |       |       |      |     |     |    |    |    |  |
|----|----|-----|-------|-------|------|-----|-----|----|----|----|--|
| ## |    |     |       |       |      |     |     |    |    |    |  |
| ## | 2  | 4   | 5     | 6     | 7    | 8   | 9   | 10 | 11 | 12 |  |
| ## | 1  | 245 | 62401 | 77937 | 6101 | 673 | 129 | 37 | 24 | 8  |  |
| ## | 16 | 18  | 19    | 22    | 26   | 29  | 39  | 41 | 49 | 53 |  |
| ## | 2  | 2   | 1     | 1     | 1    | 2   | 1   | 1  | 1  | 1  |  |

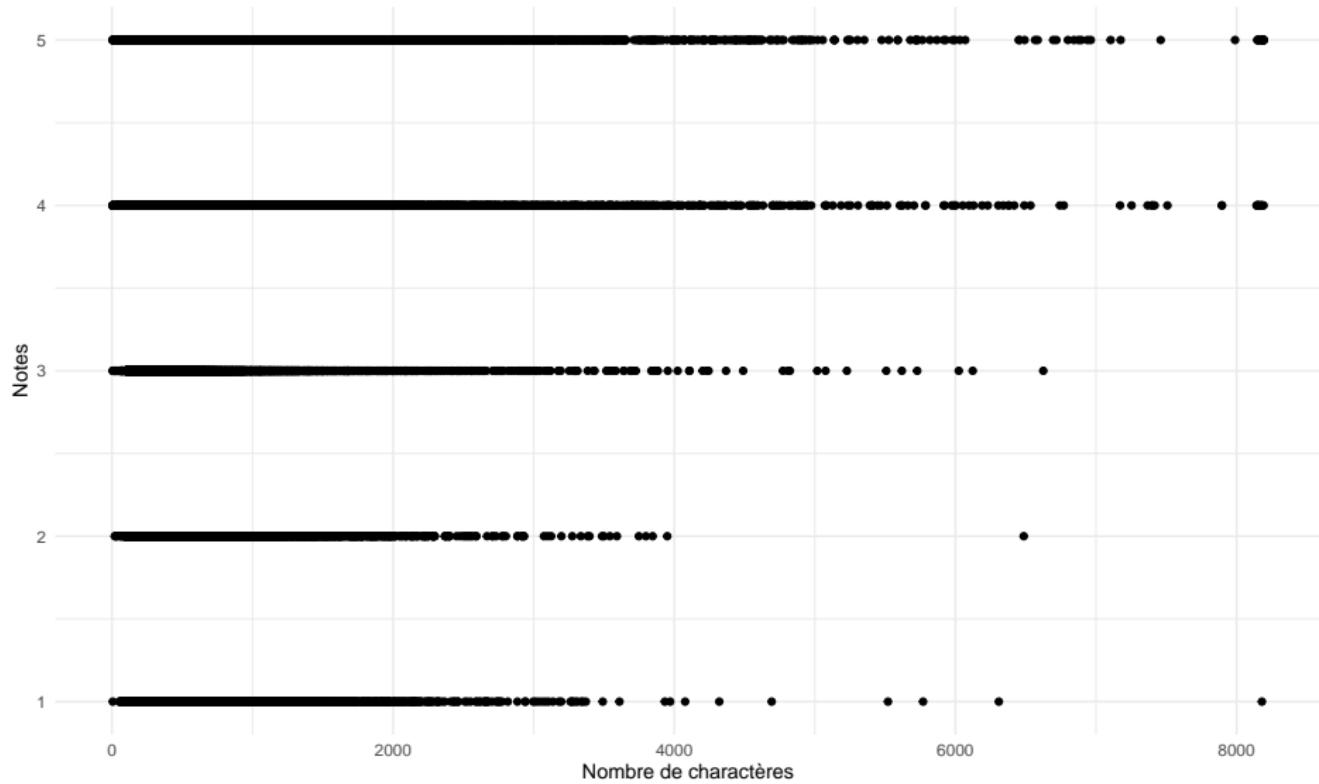
### Potentielle limite

Il s'agit d'une taille moyenne de mot par commentaire faite un peu "brutalement" (=nb de caractères / nombre de mots). Il pourrait être intéressant de regarder la répartition de la taille des mots au sein de chaque commentaire et ainsi pouvoir observer une éventuelle corrélation entre l'utilisation de mots longs et une bonne note par exemple.

## Section 4

# Répartition des longueurs de commentaires parmi les notes

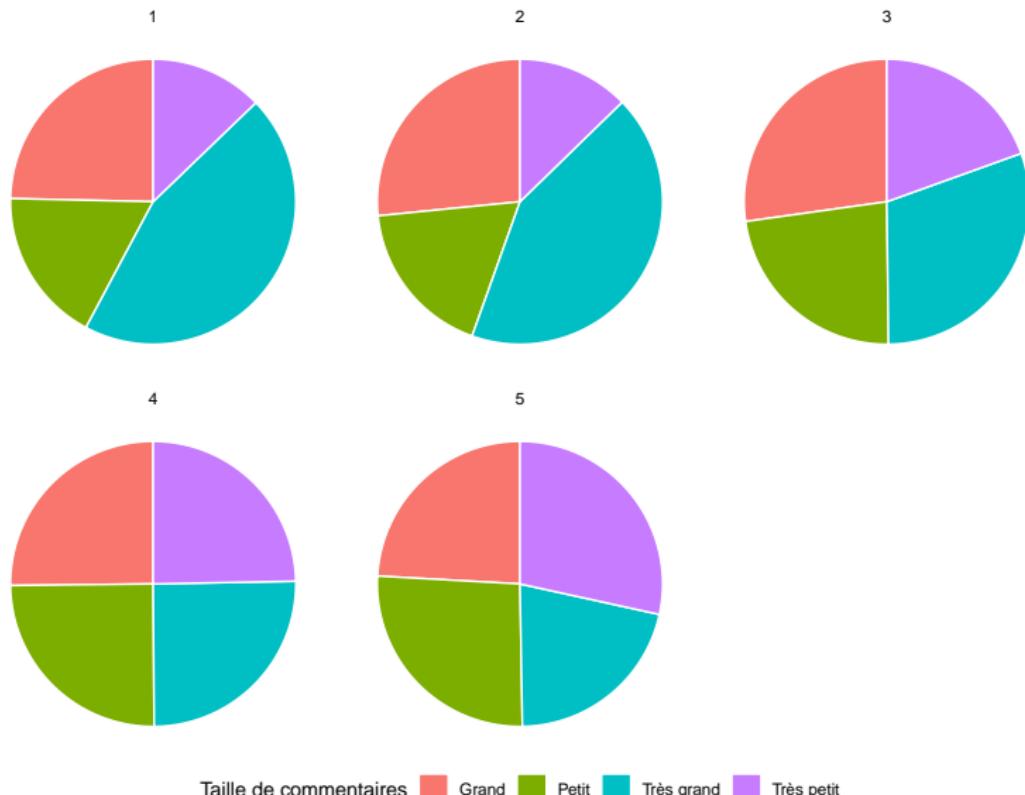
# Répartition des longueurs de commentaires parmi les notes



# Pour rappel

```
##    Rate   Freq
## 1     1 4455
## 2     2 4552
## 3     3 15935
## 4     4 50248
## 5     5 72389
```

# Proportion des longueurs de commentaires parmi les notes



# Test statistique

## Chi 2

Il faut vérifier si cette répartition est lié à une interdépendance des variables.  
Pour ce faire, un test du chi 2 est réalisé.

```
##  
## Pearson's Chi-squared test  
##  
## data: contingence_table_mtrx  
## X-squared = 26.72, df = 12, p-value = 0.008476
```

# Conclusion

À partir de ces données et selon la méthode présentée précédemment, dans la mesure où la valeur calculée du Chi-2 est supérieure à la valeur du Chi-2 théorique pour 12 degrés de liberté au seuil de 0,05 ( $26,72 > 21,026$ ), l'hypothèse H0 d'indépendance entre la variable la taille d'un commentaire et la note attribuée peut être rejetée. Autrement dit, **il existe une corrélation entre la taille d'un commentaire et la note attribuée.**

## Orientation

En l'espèce, ce travail travail n'explore pas plus en détail la corrélation qu'il existerait entre la taille d'un commentaire et la note qui lui est attribuée. Cela est toutefois un sujet qu'il peut être intéressant d'investiguer.

## Section 5

# Observation de la fréquence d'apparition des différents mots dans le corpus

## Observation de la fréquence d'apparition des différents mots dans le corpus

L'idée est de s'intéresser à la fréquence d'apparition des mots qui constituent chaque commentaire, dans chaque note attribuée.

L'intuition est que des mots positifs devraient être sureprésentés dans les bonnes notes (et inversement) ainsi l'apparition de ces mots dans un commentaire devraient donner des indices sur la note qui a été attribuée à ce commentaire.

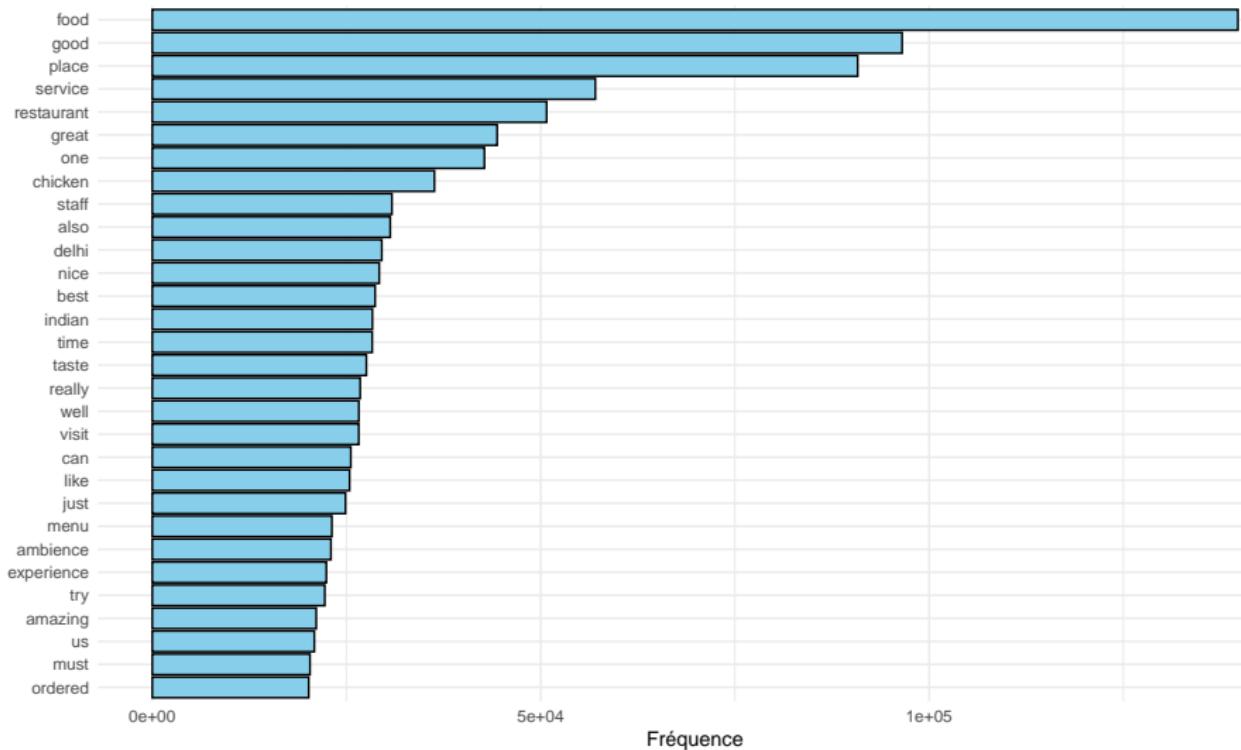
# Observation de la fréquence d'apparition des différents mots dans le corpus

Je commence par :

- mettre en forme toute la df (uniformisation de la casse, remplacement de la ponctuation et des chiffres par des espaces, suppression des espaces multiples) ;
- découper ma df par mot (je la tokenise), ce qui donne une df de 10 712 979 observations et deux variables : "Rating" et "word" ;
- supprimer les mots trop courants qui n'apportent pas de sens particulier (stop words) tels que : "me", "you", "he", "will" etc.. (il s'agit d'une liste de 175 mots que intégrée dans la fonction `get_stopwords()` du package `tidytext`) et les mots d'un seul caractère.

# Décompte des mots les plus fréquents dans le corpus

Top 30 des mots les plus fréquents dans le corpus

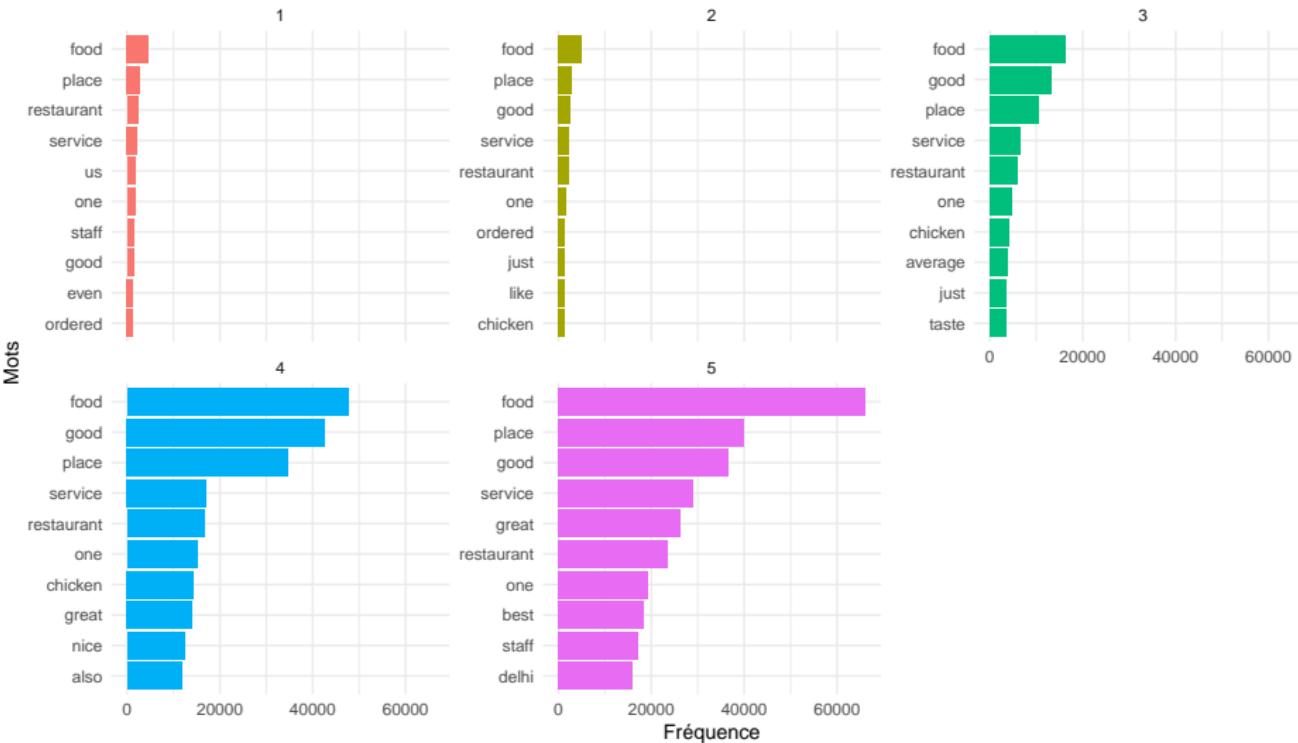


## Décompte des mots les plus fréquents dans le corpus

```
## # A tibble: 30 x 2
##   word          n
##   <chr>     <int>
## 1 food      139747
## 2 good      96524
## 3 place     90806
## 4 service    57036
## 5 restaurant 50768
## 6 great      44392
## 7 one        42757
## 8 chicken    36320
## 9 staff       30850
## 10 also      30632
## # i 20 more rows
```

# Décompte des mots les plus fréquents par note

Top 10 des mots les plus fréquents par note



## Term Frequency - Inverse Document Frequency

Les mêmes mots reviennent indépendamment de la note attribuée, c'est pourquoi il faut parvenir à pondérer les mots propres à chaque note afin de faire ressortir des tendances.

Pour ce faire on utilise la méthode du **TF-IDF**.

# Term Frequency - Inverse Document Frequency

## Term Frequency (TF)}

$$\text{TF}(t, d) = \frac{\text{Nombre d'occurrences de } t \text{ dans } d}{\text{Nombre total de mots dans } d}$$

## Inverse Document Frequency (IDF)

$$\text{IDF}(t, D) = \log \left( \frac{\text{Nombre total de documents dans } D}{\text{Nombre de documents contenant } t} \right)$$

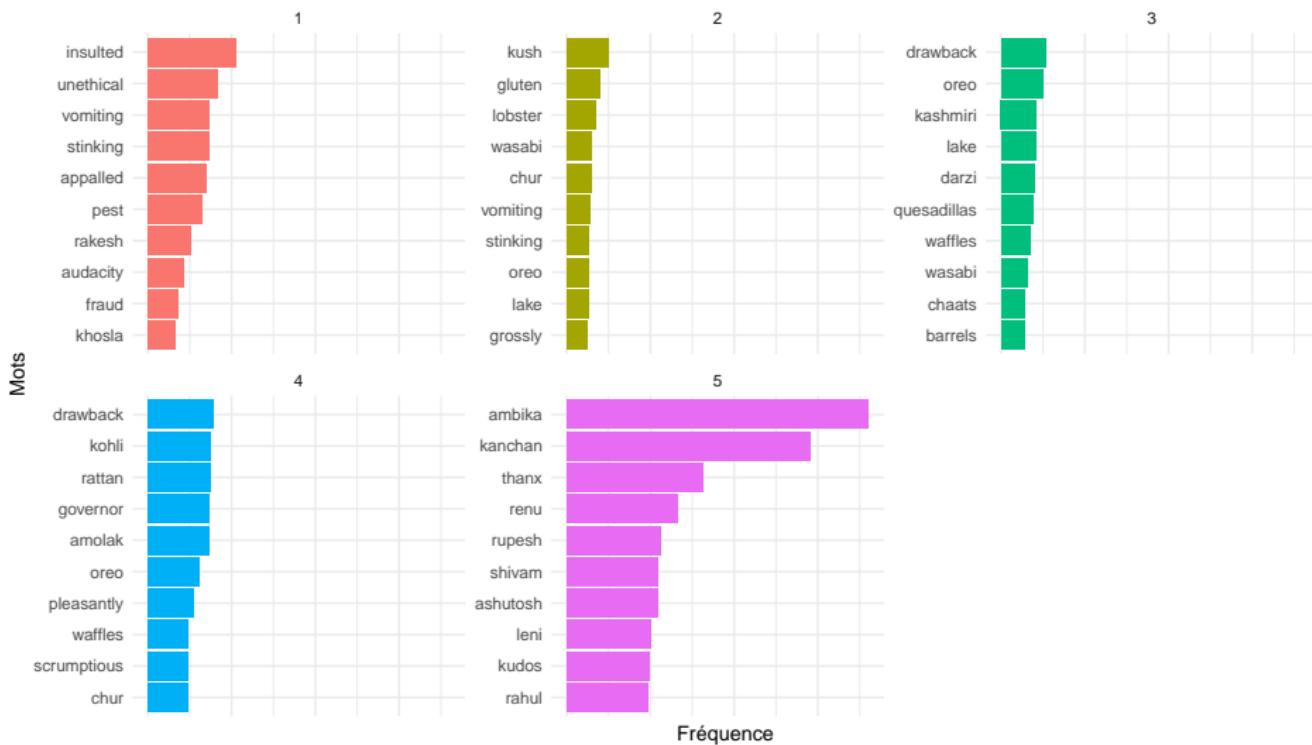
## TF-IDF

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

Où  $t$  est le terme,  $d$  le document et  $D$  le corpus de documents.

# Décompte des mots les plus fréquents par note (après TF-IDF)

Top 10 des mots les plus fréquents par note (TF-IDF)



## Décompte des mots les plus fréquents par note (après TF-IDF)

On constate que les mots qui reviennent le plus souvent dans les bonnes notes (5/5) sont des noms propres. En réalité il s'agit du personnel des bars-restaurants qui sont souvent cités dans ces commentaires.

On constate des mots comme : “insulted”, “unethical”, “vomiting” parmi les mauvaises notes (1/5).

### Orientation

Pourrait être intéressant de faire de la reconnaissance d'entités nommées (NER) pour regarder dans notre corpus, quels bar-restaurants reçoivent des bonnes ou des mauvaises notes

## Section 6

# Création d'un modèle d'estimation des notes

# Création d'un modèle d'estimation des notes

L'idée est d'entrainer un modèle simple sur une partie de la df (80%) afin d'observer s'il parvient à estimer les notes attribuées aux commentaires restants (20%) sur la seule base des mots qui les composent.

Pour ce faire je vais :

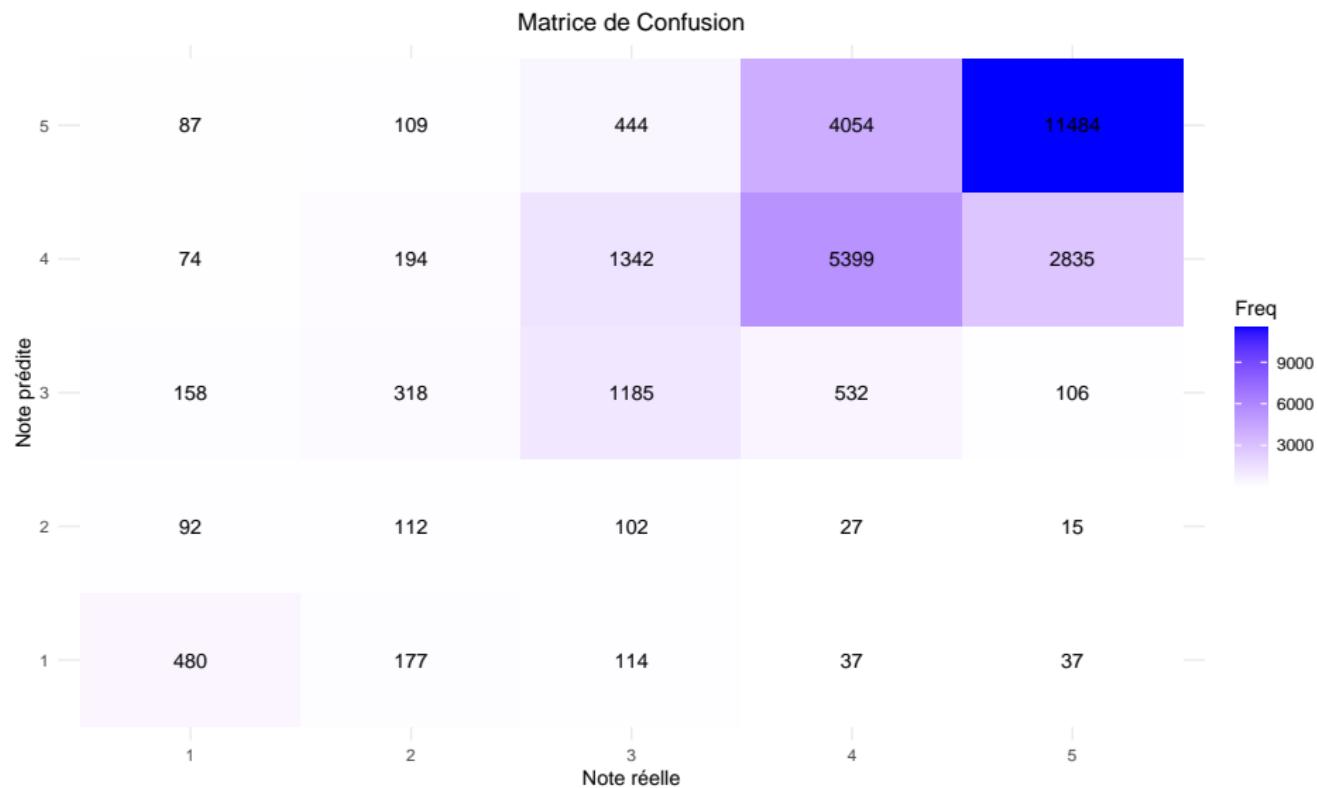
- diviser aléatoirement les commentaires en un ensemble d'entraînement et en un ensemble de test ;
- préparer les données avec la fonction `recipe()` du package `textrecipes` (on limite notamment la taille des commentaires à 1000 mots) ;
- configurer le contrôle d'entraînement (définir le nombre de sous-ensemble de la base d'entraînement pour améliorer la fiabilité de l'entraînement, la méthode de vérification de l'entraînement etc..) ;
- lancer l'entraînement ;
- évaluer le modèle.

# Création d'un modèle d'estimation des notes

## Attention

Bien que l'on se soit intéressé à la fréquence d'apparition de chaque mot du corpus dans la partie précédente, l'entraînement du modèle se fait non pas par mot mais par commentaire.

# Résultats du modèle



## Notions importantes d'évaluation d'un modèle

Comme vu dans la slide précédente, une situation paradoxale peut apparaître lorsque l'on cherche à évaluer un modèle de classification. En effet, le modèle semble être bon lorsqu'il prédit une note or pour certaines notes le modèle semble se tromper plus souvent qu'il n'a raison (par exemple pour les notes 2 et 3).

C'est pourquoi lorsque l'on évalue un modèle de classification il est indispensable de différencier la ***précision*** et le ***recall***.

# Notions importantes d'évaluation d'un modèle

## Précision

Mesure la proportion de vrais positifs parmi les prédictions positives. Un modèle avec une précision élevée fait peu de faux positifs, mais peut manquer de nombreux vrais positifs.

## Recall

Mesure la proportion de vrais positifs parmi les réels positifs. Un modèle avec un recall élevé détecte la plupart des vrais positifs, mais peut faire beaucoup de faux positifs

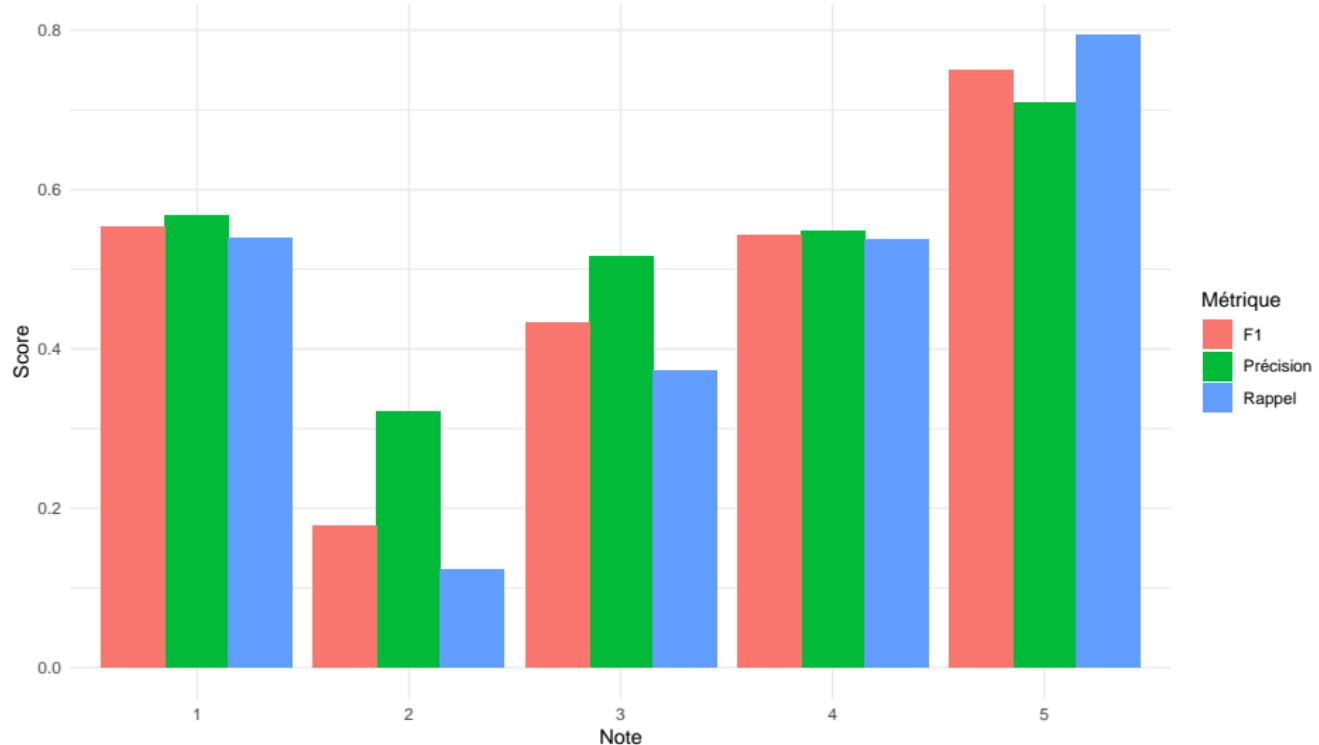
Pour palier à ce problème on utilise régulièrement le score F1.

## F1

Moyenne harmonique de la *Précision* et du *Recall*.

# Résultats du modèle

Métriques de performance par classe



# Conclusion

Le modèle est plus performant pour la prédiction des notes 1, 4 et 5. À noter que pour la note 1 il a quand même eu un échantillon très réduit pour s'entraîner.

J'ai entraîné le modèle selon une méthode relativement simple qui est la régression logistique classique. L'objectif est seulement de chercher une relation linéaire entre les mots et la probabilité de chaque note.

## Orientation

Si à l'avenir je suis amené à réentraîner des modèles il est nécessaire de mieux maîtriser les formules derrière les différentes méthodes. Il serait intéressant (et faisable à moindre coût) d'entraîner des modèles selon d'autres méthodologies (modèles bayésiens, k-plus proches voisins etc..) mais je n'ai pas réussi à le faire avec mon PC en raison de la RAM limitée.

## Section 7

# Conclusion

# Conclusion

Cette petite mise en pratique m'a permis de me refaire la main sur R, de découvrir Rmd, d'apprendre ce qu'est le TF-IDF et de manipuler les critères d'évaluation d'un modèle de classification.

Il m'a également permis de naviguer sur Hugging face et d'en apprendre un peu sur les différentes méthodologies d'estimation.

Ce travail manque toutefois cruellement d'un support théorique, d'une problématique, d'un terrain mieux documenté etc.. pour pouvoir réellement être intéressant du point de vue de la recherche.