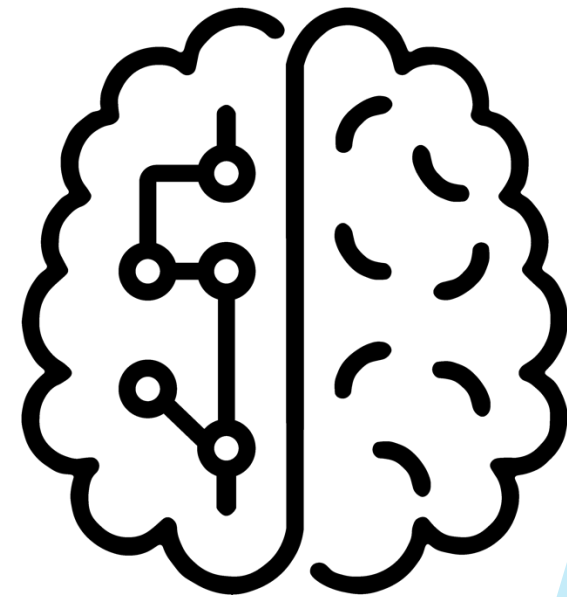
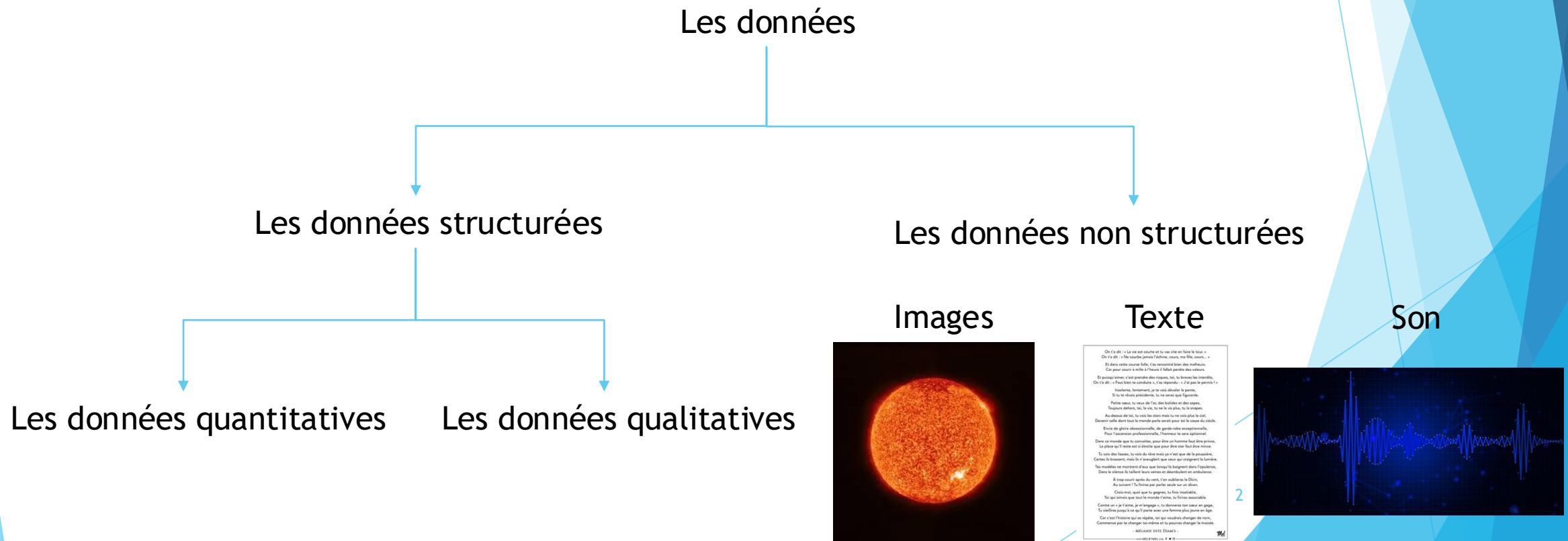


# Bag of words

AIForYou - Morgan Gautherot



# Les catégories de données



# Le traitement naturel du langage

## Texte

On t'a dit : La vie est courte et tu vas vite en faire le tour. »  
On t'a dit : « Ne courre jamais l'échine, cours, ma fille, cours. »  
Et dans cette course folle, t'as rencontré bien des malheurs.  
Car pour courir à mille à l'heure il fallait perdre des valeurs.

Et puisqu'aimer, c'est prendre des risques, toi, tu braves les interdits,  
On t'a dit : « Faut bien te conduire », t'as répondu : « J'ai pas le permis ! »

Insolente, lentement, je te vois dévaler la pente,  
Si tu te révais présidente, tu ne seras que figurante.

Petite sœur, tu veux de l'or, des bolides et des sapes,  
Toujours dehors, toi, la vie, tu ne la vis plus, tu la snapes.

Au-dessus de toi, tu vois les stars mais tu ne vois plus le ciel,  
Devenir celle dont tout le monde parle serait pour toi le casse du siècle.

Envie de gloire obsessionnelle, de garde-robe exceptionnelle,  
Pour l'ascension professionnelle, l'honneur te sera optionnel.

Dans ce monde que tu convoites, pour être un homme faut être prince,  
La place qu'il reste est si étroite que pour être star faut être mince.

Certes tes fiascos, tu vois du rêve mais ça n'est que de la poussière,  
Certes ils brassent, mais ils n'avenglent que ceux qui craignent la lumière.

Tes modèles ne montrent d'eux que lorsqu'ils baignent dans l'opulence,  
Dans le silence ils taillent leurs veines et déambulent en ambulance.

À trop courir après du vent, t'en oulerais le Divin,  
Au suivant ! Tu finiras par parler seule sur un divan.

Crois-moi, quoi que tu gagnes, tu finis insatiable,  
Toi qui aimais que tout le monde t'aime, tu finiras accessible.

Contre un « je t'aime, je m'engage », tu donneras ton cœur en gage,  
Tu vieilliras jusqu'à ce qu'il parte avec une femme plus jeune en âge.

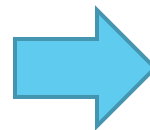
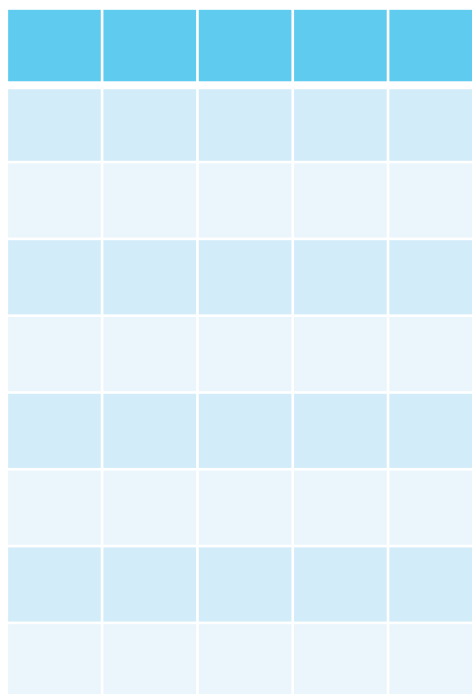
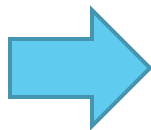
Car c'est l'histoire qui se répète, toi qui voudrais changer de nom,  
Commence par te changer toi-même et tu pourras changer le monde.

- MÉLANIE DITE DIAM'S -

— [www.MELBYMEL.com](http://www.MELBYMEL.com)   



## Données exploitables



## Modèles



# Le vecteur de vocabulaire

$$V = \begin{bmatrix} arachide \\ \vdots \\ bain \\ \vdots \\ voiture \\ \vdots \\ zumba \end{bmatrix}$$

# D'une phrase à un sac de mots

Doc 1

Le thé est bouillant



Traitements du texte

Doc 1

[ 'thé', 'être', 'bouillant' ]



Transformation en bag of words

$V = [\textit{arachide}, \dots, \textit{bain}, \textit{bouillant}, \dots, \textit{\`etre}, \dots, \textit{th\'e}, \dots, \textit{voiture}, \dots, \textit{zumba}]$

$D_1 = [ \quad 0, \dots, \quad 0, \quad 1, \dots, \quad 1, \dots, \quad 1, \dots, \quad 0, \dots, \quad 0 ]_5$

# D'un corpus à un dataframe

Doc 1



Tu me donnes envie de faire table rase du passé, et de t'aider à faire table rase du tien. Tu me donnes envie d'être pour toi ce que je n'ai jamais su être pour quelqu'un d'autre. Tu me donnes envie de t'offrir mon cœur, mes ambitions, mon amour sur un plateau d'argent. Tu me donnes envie de t'aimer à en faire trembler le monde entier. Tu me donnes envie de parler de sentiments et de bonheur tout en conjuguant mes verbes au présent. Tu me donnes envie de devenir quelqu'un qui t'apportera fierté et douceur. Tu me donnes envie de décrocher la Lune, de croire en l'impossible. Tu me donnes envie de vivre, bordel. Parce que tu es l'étincelle au fond de mes yeux, mon sourire terriblement sincère. Parce que je ne suis plus seule désormais, tu es là. Et c'est si précieux.

$= D_1$

Doc 3

006

## L'AMITIÉ

L'amitié est un repère, une ligne d'horizon. Une lumière et une porte secrète vers les émotions... Elle nous aide par temps durs, elle fait rire, elle rassure. Elle ne juge pas, elle accepte, elle est là... Tantôt près, tantôt loin, mais toujours à distance de coeurs, notre amitié c'est du bonheur. Notre amitié, c'est une plante que je cultive avec grands soins, le trésor de mon petit jardin. Une pierre précieuse qui m'éclaire dans le noir, et quelqu'un qui connaît toute mon histoire... J'avance sur mon chemin, tu avances sur le tien. L'horizon devant nous est infini, tout comme notre amitié, merci.

$= D_3$

Doc 2

## Le supermarché

Sophie fait les courses au supermarché avec son nouveau Clément. Après avoir garé sa voiture sur le parking, elle a pris un chariot dans lequel Clément s'est installé. Tous deux pénètrent maintenant dans le magasin. Les clients sont nombreux à cause des promotions et des soldes. Sophie sort sa liste des courses pour ne rien oublier. Elle regarde les marques et compare les prix. Elle choisit des produits frais, des conserves et du poisson surgelé qu'elle met dans son chariot. Clément est turbulent, il s'agite sur son siège, il ne cesse de réclamer des bonbons, des biscuits, du chocolat, des jouets. Sophie le gronde gentiment puis elle regarde sa montre, il est déjà seize heures. Elle doit se hâter de rentrer car sa sœur elle est invitée avec Jean à dîner chez la maman de Clément, sa sœur Lucie. Sophie termine les courses, elle se dirige vers la caisse où elle attend calmement son tour. A la caisse, elle place ses achats sur le tapis roulant puis elle règle sa note avec sa carte bancaire. A présent, elle regagne le parking, cherche quelques instants sa voiture. Elle la repère au fond de l'allée puis se dirige vers son véhicule. Elle ouvre le coffre et y met les courses puis elle installe Clément dans son siège pour enfant. Elle part retrouver Jean et se préparer pour la soirée.

$= D_2$

Doc 4

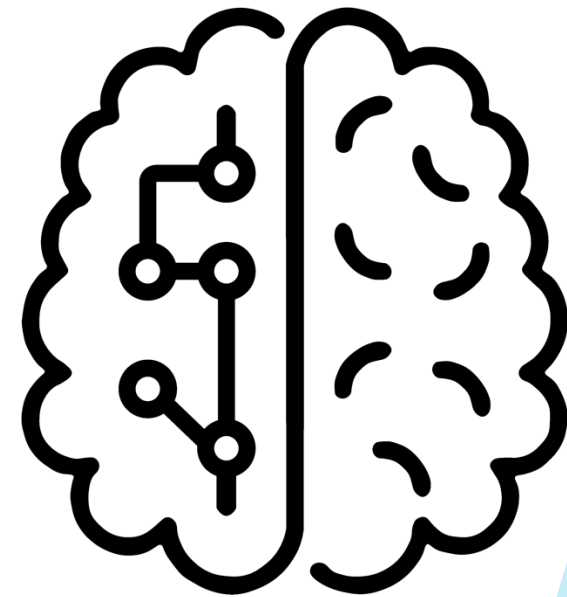
On s'a dit : « La vie est courte et tu vas vite en faire la tour »  
On s'a dit : « Ne sois pas triste/lâche, mais, meille, courage »  
Et dans cette course folle, t'as rencontré bien des malheurs  
Car pour courir à mille à l'heure t'fallait perdre des valeurs  
Et même parfois, c'est possible des risques, toi, tu braves les tempêtes  
On s'a dit : « Faut bien se conduire », t'as répondu : « J'ai pas le permis ! »  
Insolente, lentement, je te vais dévoiler la partie  
Si tu te dirais présidente, tu ne serais que l'égérie  
Petite sœur, tu veux de l'or, des milliards et des sapins  
Tropes d'effort, toi, tu vois, tu ne te vois plus le ciel  
Au-dessus de toi, tu vois les stars mais tu ne vois plus le ciel  
Devenir celle dont tout le monde parle pour toi le caser du siècle  
Bonne des glorie obsolescence, de grande classe obsolescence  
Pour l'accession professionnelle, l'homme ne sera optionnel  
Dans ce monde que tu connais, pour dire un homme faut être prince  
La chose qu'il t'enseigne et il dit que pour dire être prince  
Tu vois des lasses, tu vois du rêve mais ça n'est que de la poussière  
Certes la beauté, mais tu te rends compte que ça n'est que la poussière  
Tes modèles ne montrent d'ailleurs que lorsqu'ils lèguent dans l'apathie  
Dans le silence de taillat leurs vaines et débilitent en ambulance  
A trop courir après du vent, t'en maitrises la direction  
Au secours ! Tu feras pas partie, mais tu en diras  
Crisse moi, quoi que tu gagnes, tu fais inévitable  
Toi qui crèves que tout le monde t'aime, tu feras inévitable  
Contre un « je t'aime, je m'engage », tu donnes ton cœur en gage  
Tu vas faire jusqu'à ce qu'il parte avec une femme plus jeune en âge  
Car c'est l'histoire qui se répète, toi qui vas devoir changer de route  
Commence par te changer toi-même et tu pourras changer le monde  
- MÉLANIE DITE DINAÏS -

$= D_4$

V	$D_1$	$D_2$	$D_3$	$D_4$	...	$D_n$
$mot_1$	0	1	0	0	...	0
$mot_2$	0	0	0	1	...	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$mot_m$	0	1	0	0	0	0

# Le TF-IDF

AIForYou - Morgan Gautherot



# Définition du TF-IDF

TF = Term Frequency

IDF = Inverse Document Frequency



# Term Frequency

$$tf_{w,d} = \frac{n_{w,d}}{\sum_k n_{k,d}}$$

$d$  est un document de notre jeu de données

$w$  est un mot de notre document

$n_{w,d}$  est le nombre d'occurrences du mot  $w$  du document  $d$

# Inverse Document Frequency

$$idf_w = \log\left(\frac{|D|}{|\{d_i: t_w \in d_i\}|}\right)$$

$D$  désigne tous les documents de notre jeu de données

$w$  est un mot de notre document

$f(w, D)$  est le nombre de document  $D$  contenant le mot  $w$

# Calcul du TF-IDF

$$tfidf_{w,d} = tf_{w,d} * idf_w$$

$$tfidf_{w,d} = \frac{\text{fréquence du terme dans le document}}{\text{fréquence du terme dans le corpus}}$$

# Vecteur de vocabulaire

$$V = \begin{bmatrix} arachide \\ \vdots \\ bain \\ \vdots \\ voiture \\ \vdots \\ zumba \end{bmatrix}$$

# D'une phrase à un vecteur

Doc 1

Le thé est bouillant



Traitements du texte

Doc 1

[ 'thé', 'être', 'bouillant' ]



Transformation en vecteur

$V = [arachide, \dots, bain, bouillant, \dots, être, \dots, thé, \dots, voiture, \dots, zumba]$

$D_1 = [ \quad 0, \dots, \quad 0, \frac{1}{nb_{bouillant}}, \dots, \frac{1}{nb_{être}}, \dots, \frac{1}{nb_{thé}}, \dots, \quad 0, \dots, \quad 0 ]$

# D'un corpus à une matrice



Doc 1

Tu me donnes envie de faire table rase du passé, et de t'aider à faire table rase du tien. Tu me donnes envie d'être pour toi ce que je n'ai jamais su être pour quelqu'un d'autre. Tu me donnes envie de t'offrir mon cœur, mes ambitions, mon amour sur un plateau d'argent. Tu me donnes envie de t'aimer à en faire trembler le monde entier. Tu me donnes envie de parler de sentiments et de bonheur tout en conjuguant mes verbes au présent. Tu me donnes envie de devenir quelqu'un qui t'apportera fierté et douceur. Tu me donnes envie de décrocher la Lune, de croire en l'impossible. Tu me donnes envie de vivre, bordel. Parce que tu es l'étincelle au fond de mes yeux, mon sourire terriblement sincère. Parce que je ne suis plus seule désormais, tu es là. Et c'est si précieux.

$= D_1$

Doc 3

006  
L'AMITIÉ  
L'amitié est un repère,  
une ligne d'horizon. Une lumière  
et une porte secrète vers les émotions...  
Elle nous aide par temps durs, elle fait rire,  
elle rassure. Elle ne juge pas, elle accepte,  
elle est là... Tantôt près, tantôt loin,  
mais toujours à distance de coeurs, notre  
amitié c'est du bonheur. Notre amitié, c'est  
une plante que je cultive avec grands soins,  
le trésor de mon petit jardin. Une pierre  
précieuse qui m'éclaire dans le noir, et  
quelqu'un qui connaît toute mon histoire...  
J'avance sur mon chemin, tu avances sur  
le tien. L'horizon devant nous est infini,  
tout comme notre amitié, merci.

$= D_3$

Doc 2

Sophie fait les courses au supermarché avec son neveu Clément. Après avoir garé sa voiture sur le parking, elle a pris un chariot dans lequel Clément s'est installé. Tous deux pénètrent maintenant dans le magasin. Les clients sont nombreux à cause des promotions et des soldes. Sophie sort sa liste des courses pour ne rien oublier. Elle regarde les marques et compare les prix. Elle choisit des produits frais, des conserves et du poisson surgelé qu'elle met dans son chariot. Clément est turbulent, il s'agite sur son siège, il ne cesse de réclamer des bonbons, des biscuits, du chocolat, des jouets. Sophie le gronde gentiment puis elle regarde sa montre, il est déjà seize heures. Elle doit se hâter de rentrer car sa mère est invitée avec Jean à dîner chez la maman de Clément, sa sœur Lucie. Sophie termine les courses, elle se dirige vers la caisse où elle attend calmement son tour. A la caisse, elle place ses achats sur le tapis roulant puis elle règle sa note avec sa carte bancaire. A présent, elle repasse le parking, cherche quelques instants sa voiture. Elle la repère au fond de l'allée puis se dirige vers son véhicule. Elle ouvre le coffre et y met les courses puis elle installe Clément dans son siège pour enfant. Elle part retrouver Jean et se préparer pour la soirée.

$= D_2$

Doc 4

On t'a dit : « La vie est courte et tu vas vite en faire le tour »  
On t'a dit : « Ne sois pas l'heureux, mais l'heureux »  
Et dans cette course folle, t'as rencontré bien des malheurs  
Car pour courir à mille à l'heure t'as dû perdre des valeurs  
Et même alors, t'as perdu des rêves, toi, tu braves les tempêtes  
On t'a dit : « Faut bien se conduire », t'as répondu : « J'ai pas le permis ! »  
Insolent, tant mieux, je te vais diviser la pente,  
Si tu te dirais président, tu ne serais que l'opinion.  
Petite sœur, tu veux de l'or, des milliards et des sapins,  
Trop vite défilé, toi, tu vois, tu ne te vois plus le ciel.  
Au-dessus de toi, tu vois les stars mais tu ne vois plus le ciel.  
Devenir celle dont tout le monde parle pour toi le caser du siècle.  
Bonne de gloire déconcertante, de grande classe incontestable.  
Pour l'ascension professionnelle, l'homme ne sera optionnel.  
Dans ce monde que tu connais, pour être un homme faut être prince,  
La reine qu'il t'aime est si délicate que pour dire vite faut être reine.  
Tu vois des lasses, tu vois du rêve mais ça n'est que de la poussière,  
Certes la poussière, mais la poussière que nous qui créons la lumière.  
Tes modèles ne montrent d'ailleurs que lorsqu'ils laissent l'apathie.  
Dans le silence de taillat leurs vaines et déshabillent en ambulance.  
A trop vouloir signer du vent, t'en as fait la pluie.  
Au secours ! Tu feras pas partie, mais tu en es sûr.  
Crisse moi, tout que tu gagnes, tu fais l'invincible.  
Toi qui crèves que tout le monde t'aime, tu feras l'invincible.  
Contre un « je t'aime, je m'engage », tu donneras ton cœur en gage,  
Tu vas faire jusqu'à ce qu'il parte avec une femme plus jeune en âge.  
Car c'est l'histoire qui se répète, toi qui veux être d'âge de nous.  
Commence par te changer toi-même et tu pourras changer le monde.

$= D_4$

V	D <sub>1</sub>	D <sub>2</sub>	D <sub>3</sub>	D <sub>4</sub>	...	D <sub>n</sub>
mot <sub>1</sub>	0	$\frac{3}{mot_1}$	0	0	...	0
mot <sub>2</sub>	0	0	0	$\frac{1}{mot_2}$	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
mot <sub>m</sub>	0	$\frac{8}{mot_m}$	0	0	0	0

# Notebook 2 - Implémentation du TF-IDF

Objectifs :

- ▶ Implémenter le bag of word pour un corpus de texte
- ▶ Implémenter le TF-IDF pour un corpus de texte