

Analyse à grande échelle de sentiments sur Twitter

Dehbi ilyas
POMALEGNI Augustin Primous
Prince

Table of Contents

01

Introduction

...

02

Présentation des
technologies utilisées

...

03

Expérimentations

...

04

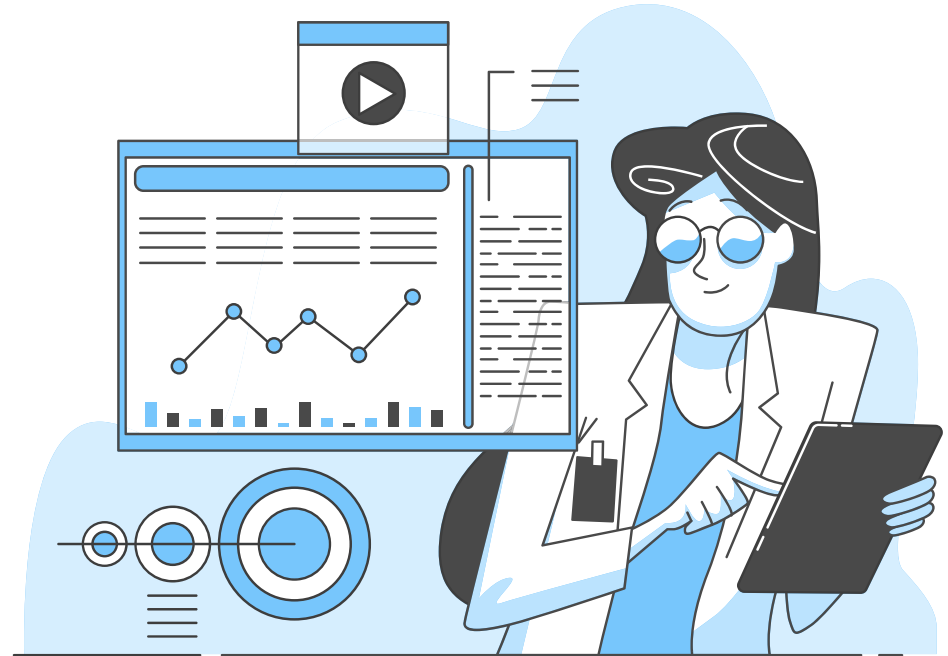
Mesure et analyse

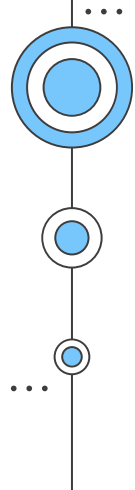
...

05

Comparaison et Conclusion

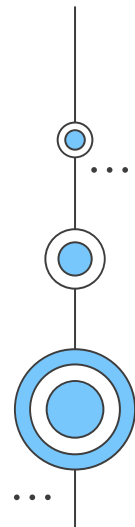
...



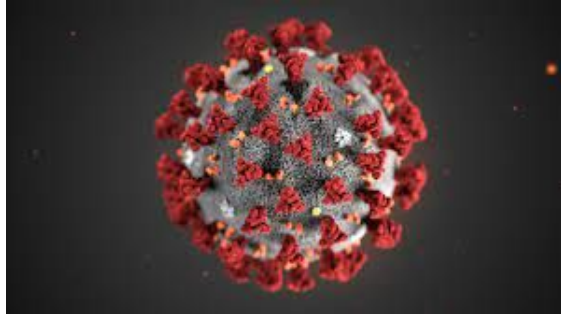


01

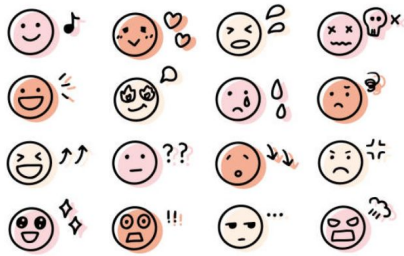
Contexte



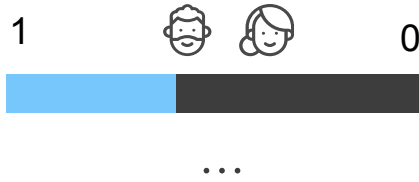
Contexte

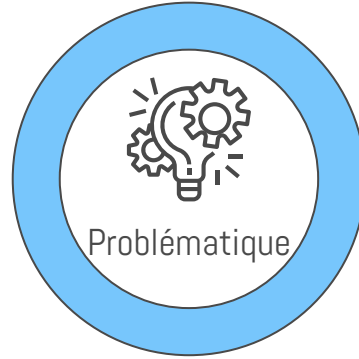


Depuis Décembre 2019



Sentiment barre





Comment faire de l'analyse de sentiments des données sur le Covid-19 provenant de Twitter?

...



02

Technologie
utilisée





Technologies utilisés



01 **PySpark**

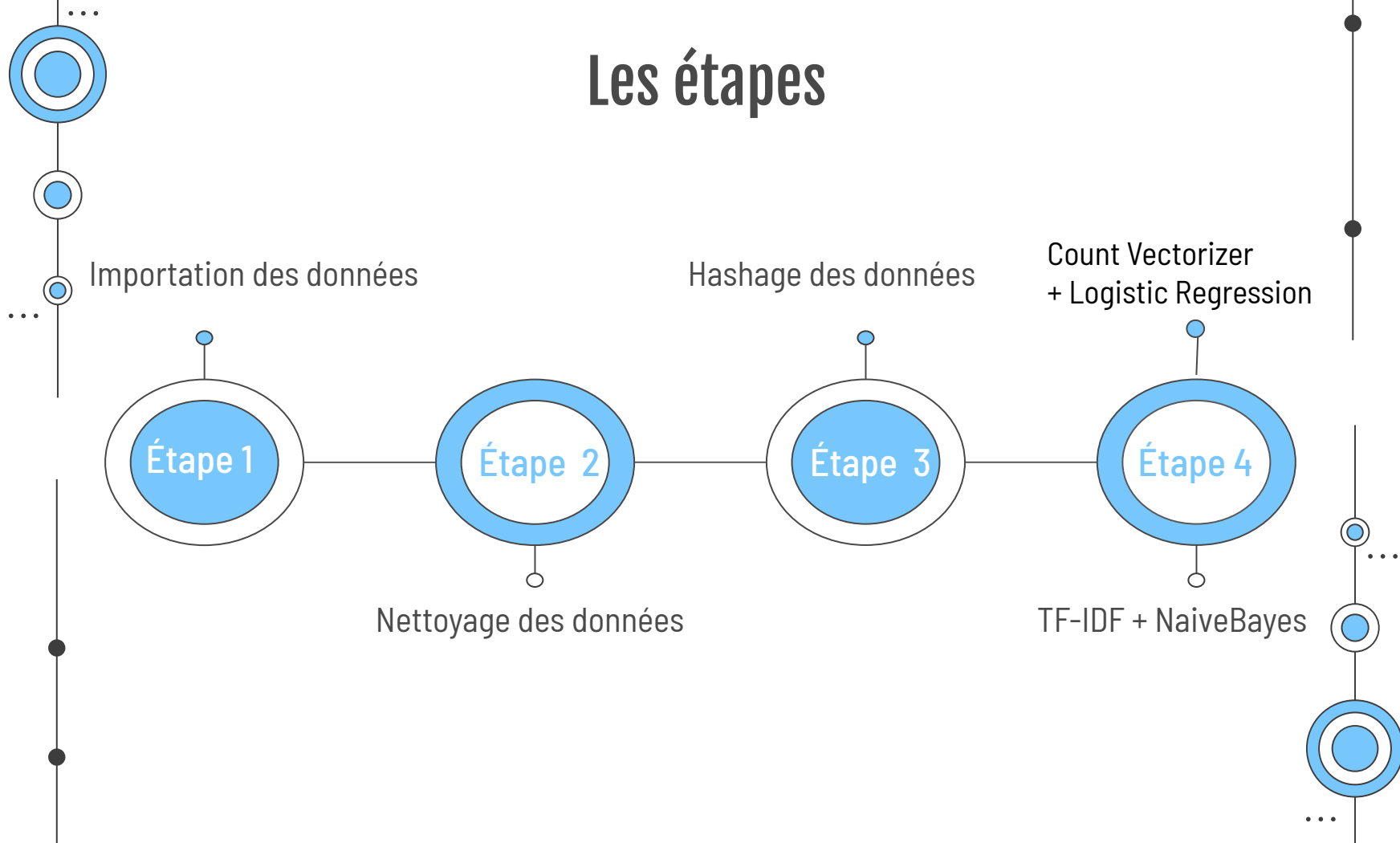
02



The image features the word "PySpark" in a large, bold, black font, centered within a light blue, irregularly shaped cloud-like background. Surrounding this central element are several circular nodes, each consisting of a blue outer ring and a white inner circle. These nodes are connected by thin black lines, forming a network diagram. In the top right, a path of three nodes leads to a larger, more prominent node. In the bottom left, a path of four nodes leads to a larger, more prominent node. Ellipses (...) are used to indicate that the network continues beyond the visible nodes.

PySpark

Les étapes



TF-IDF

$$TF(M_i, D_j) = \frac{\text{nombre de fois que le mot } M_i \text{ apparait dans le document } D_j}{\text{nombre de mots dans le document } D_j}$$

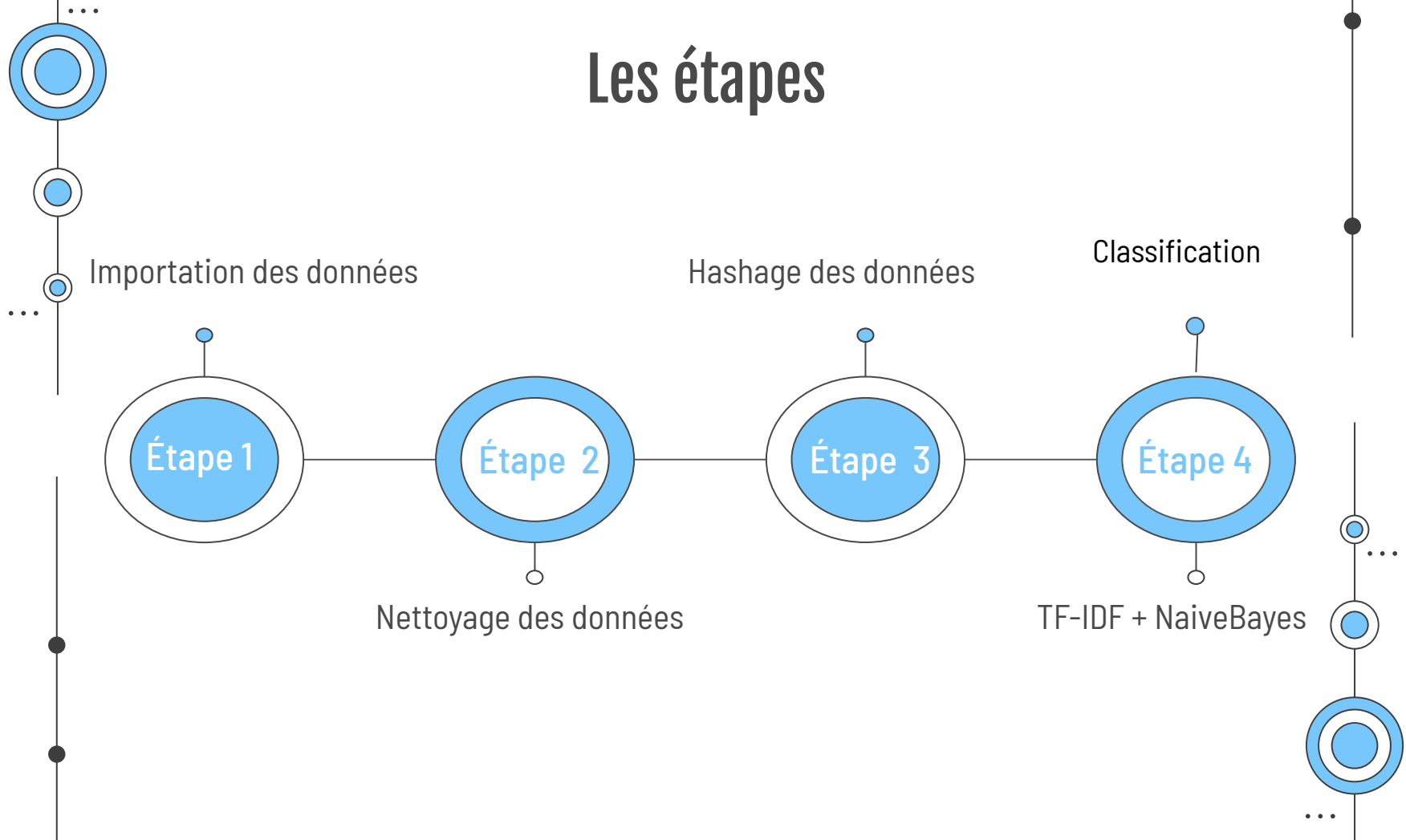
$$IDF(M_i) = \ln\left(\frac{\text{Le nombre total des documents}}{\text{Le nombre de documents qui contiennent } M_i}\right)$$

$$TF - IDF(M_i, D_j) = TF(M_i, D_j) \times IDF(M_i)$$

The image features the NLTK logo, which consists of the letters "NLTK" in a bold, black, sans-serif font. The text is centered within a large, light blue, irregular blob shape. Surrounding this central element are several circular nodes, each composed of three concentric circles (a small blue inner circle, a white middle ring, and a blue outer ring). These nodes are connected by thin black lines, forming a network diagram. One node is located at the top left, another at the top right, and a cluster of three nodes is at the bottom. Ellipses (...) are placed near the top right and bottom left nodes, indicating that the network continues beyond the visible elements.

NLTK

Les étapes





03

Expérimentation





Expérimentation



01

Présentation des données

- Méthodes d'acquisition
- Exemples de données

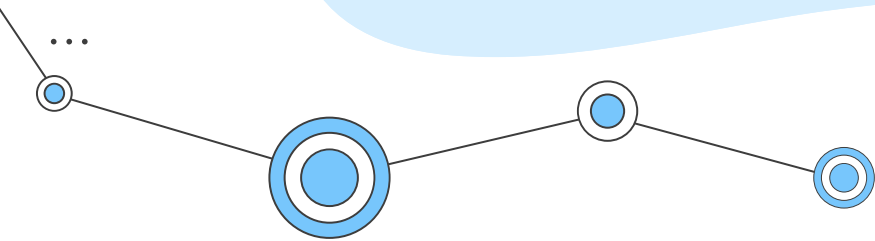
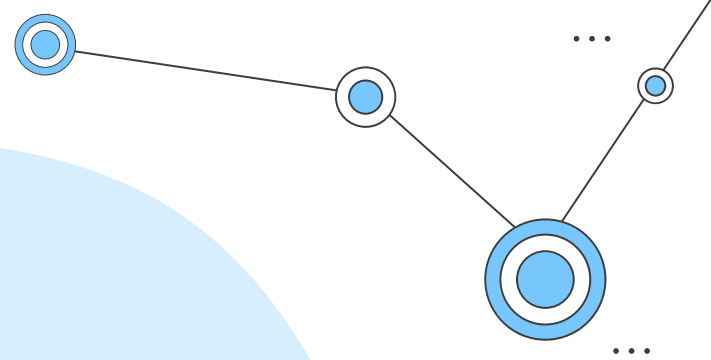
02

Expérimentation

Environnement utilisé



Données





192 000 000

D'utilisateurs au quotidien en fin 2020

504 000 000

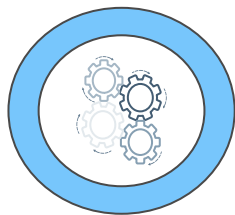
De tweets par jour soit 5900/s

XXX 000 000

De sujets différents

Twitter





Comment charger ces
tweets concernant le
covid dans un projet?



Extraction de données sur twitter notre projet

Asynchrone

Utilisation des exemples de données qu'on charge en CSV

Synchrone

Librairies ou API d'extraction

- Twitterscrapper
- Twint
- GetOldTweets3
- [Tweepy + API Twitter](#)

Exemples des tweets

Infos Covid19 @CovidBot1 · 1 h
Il y a actuellement 142662310 personnes infectées partout dans le monde. 121298906 personnes ont été soignées et 3041802 personnes sont décédées depuis le début du **#covid19**

Salia DIABATE @SaliaDIABATE5 · 15 s
Après le VIH, les gaz à effet de serre et la **COVID-19**, un bien vital est en train de nous échapper : L'AIR, de par sa qualité de plus en plus médiocre. Si les nations du monde ne unissent pas pour UN TRAITÉ, l'air respiré, quel qu'en soit le lieu planétaire, deviendra un poison.

La Commission de la santé ment... · 1 h
#Budget2021 Un fil!
Dans l'ensemble, nous sommes ravis de voir des investissements d'environ un milliard de dollars pour la **#SantéMentale**, en particulier compte tenu des impacts sur la santé mentale de la **#COVID19**.

[Afficher cette discussion](#)

Maât @Flora30255705 · 1 h
Et pendant ce temps là à l'hôpital, on continu de recruter des chargés de mission "performance hospitalière" ... Il y aura bientôt plus de personnel pour mesurer l'efficacité des services publics que pour les faire fonctionner... **#COVID19** **#servicepublic**

Chargé de mission performance hospitalière et Ressources Humaines (H/R) (Emploi ouvert aux titulaires et/ou aux contractuels)

- Réf 2021-578219
- Régions : Auvergne-Rhône-Alpes
- Mission proposée : CF PJ

Mia 🌸 **Ingridentement Votre** ❤️ @Ingr... · 1 h
Mon best qui me dit qu'il est arrivé à Mada en urgences car son père est dans le coma a cause du **Covid 19** 😞

Karina Gould @karinagould · 1 h
Canada government official
Le 🇨🇦 continuera d'être un chef de file en appuyant les pays en développement à répondre aux besoins humanitaires dans le 🌍, tout en répondant également à la crise **COVID19**.

Notre reprise économique comprendra également des investissements pour les plus vulnérables. **#Budget2021**

Annie Koutrakis @AnnieKoutrakis · 28 m
Le **#Budget2021** trace une voie pour sortir de la pandémie du **#COVID19**. Il prévoit des investissements pour créer des emplois et faire croître l'économie tout en jetant les bases d'une prospérité canadienne à long terme. **#VotreBudget** budget.gc.ca/2021/home-accu...

André MASCARDI @aldivers1 · 42 m
#COVID19 ça fait + d'un an qu'Olivier **#veran** nous ment et dit n'importe quoi : "La situation s'améliore depuis 5 jours" (le 19/04/21).
Jugez sur les 3 dernier jours NB MORTS:
189(J-2) 140(J-1) 447(J) +258, NB en **#reanimation** : 5877(J-2) 5893(J-1) 5970(J) +93 0/10 en Mathématiques

user_id	user_id_str	username	name	day	hour	link	retweet	likes	replies	retweets	quote_url
714990338649095	714990338649095	SticLucas	CherchiUitN	1	11	https://twitter.com/SticLucas	FALSE	1	0	0	
1210351232331796	1210351232331796	izaviale	izaviale	1	11	https://twitter.com/izaviale	FALSE	0	0	0	
1214315619031476	1214315619031476	Confita_FR	Confita	1	11	https://twitter.com/Confita	FALSE	249	11	132	
362050772	362050772	Edlaugren10	qMBSSI 〰️ 〰️ 〰️ 〰️	1	11	https://twitter.com/Edlaugren10	FALSE	2	0	1	
378817170	378817170	EdinWenke	La Sene Padona	1	11	https://twitter.com/EdinWenke	FALSE	0	0	0	
2162947538	2162947538	jeloqui83	INDEPENDANT	1	11	https://twitter.com/jeloqui83	FALSE	0	0	0	
57597804	57597804	bolanielle	alain bolanielle	1	11	https://twitter.com/bolanielle	FALSE	8	1	4	
85862615	85862615	SportCATeam	Sport24 ++	1	11	https://twitter.com/Sport24	FALSE	0	0	0	
3918165962	3918165962	senegal7com	senegal7	1	11	https://twitter.com/senegal7	FALSE	0	0	0	
140496388	140496388	legisocial	LegiSocial	1	11	https://twitter.com/legisocial	FALSE	0	0	0	

Dataset de travail créé

	Tweets	Label
0	keffkorvos cdube_sante ben moi déjà je conna...	Neutre
1	rt Instantfoot florentino perez j'ai décidé d...	Neutre
2	bigsuais jmblanquer vous savez quoi mettez les ...	Neutre
3	rt viedecarabin imaginez on n'aurait pas eu du...	Neutre
4	rt souleygk_pas étonnant pour un pays qui se ...	Négatif



Environnement d'expérimentation



Environnement d'expérimentation

Pyspark



NLTK




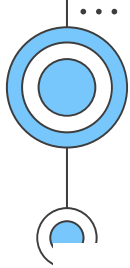
**Mais surtout sur un PC
personnel**



04

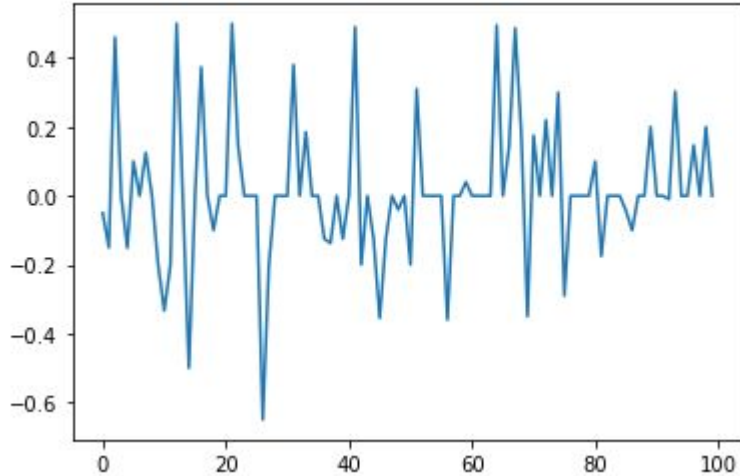
Mesure et Analyse



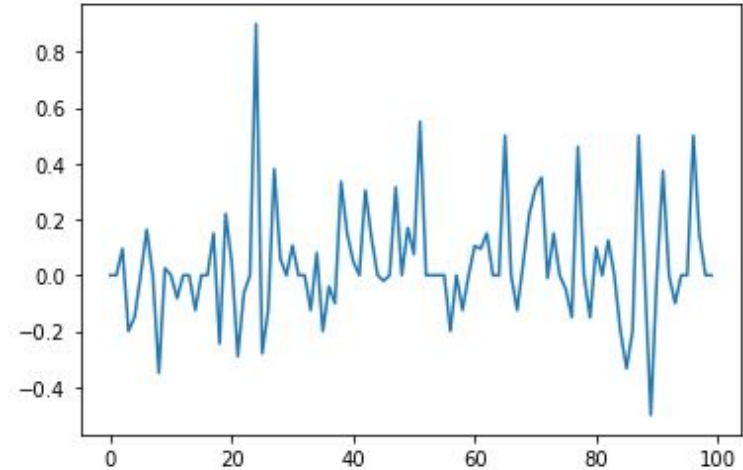


Mesure: Polarité des tweets(100 x 2)

...

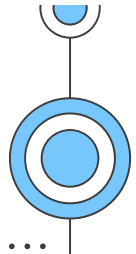


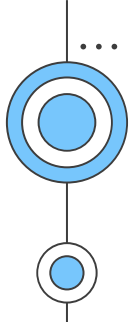
Extrait 1



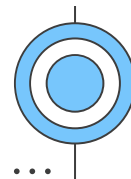
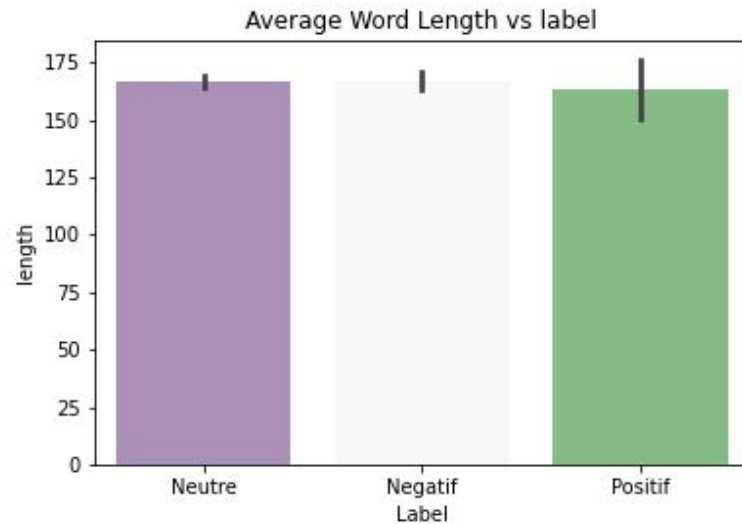
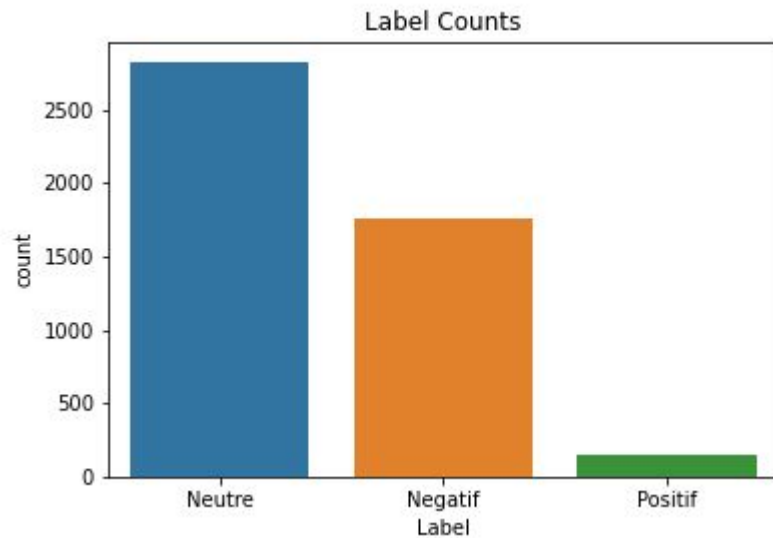
Extrait 2

..





Mesure : d'autres tendances



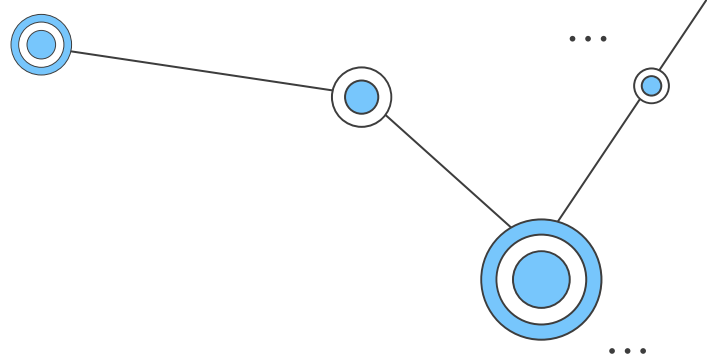
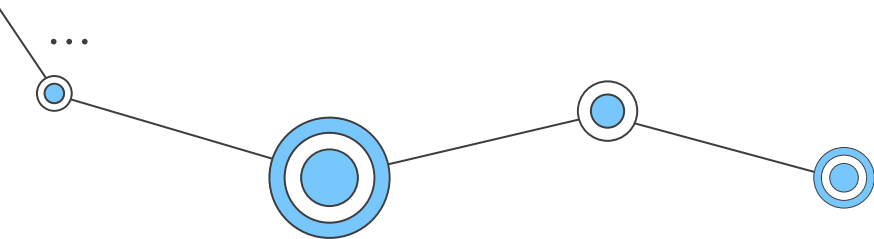


PySpark

Résultats et Analyse

$$\text{Sensitivity} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}}$$

$$\text{specificity} = \frac{\text{TrueNegative}}{\text{TrueNegative} + \text{FalsePositive}}$$



$$PPV = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}}$$

$$NPV = \frac{\text{TrueNegative}}{\text{TrueNegative} + \text{FalseNegative}}$$

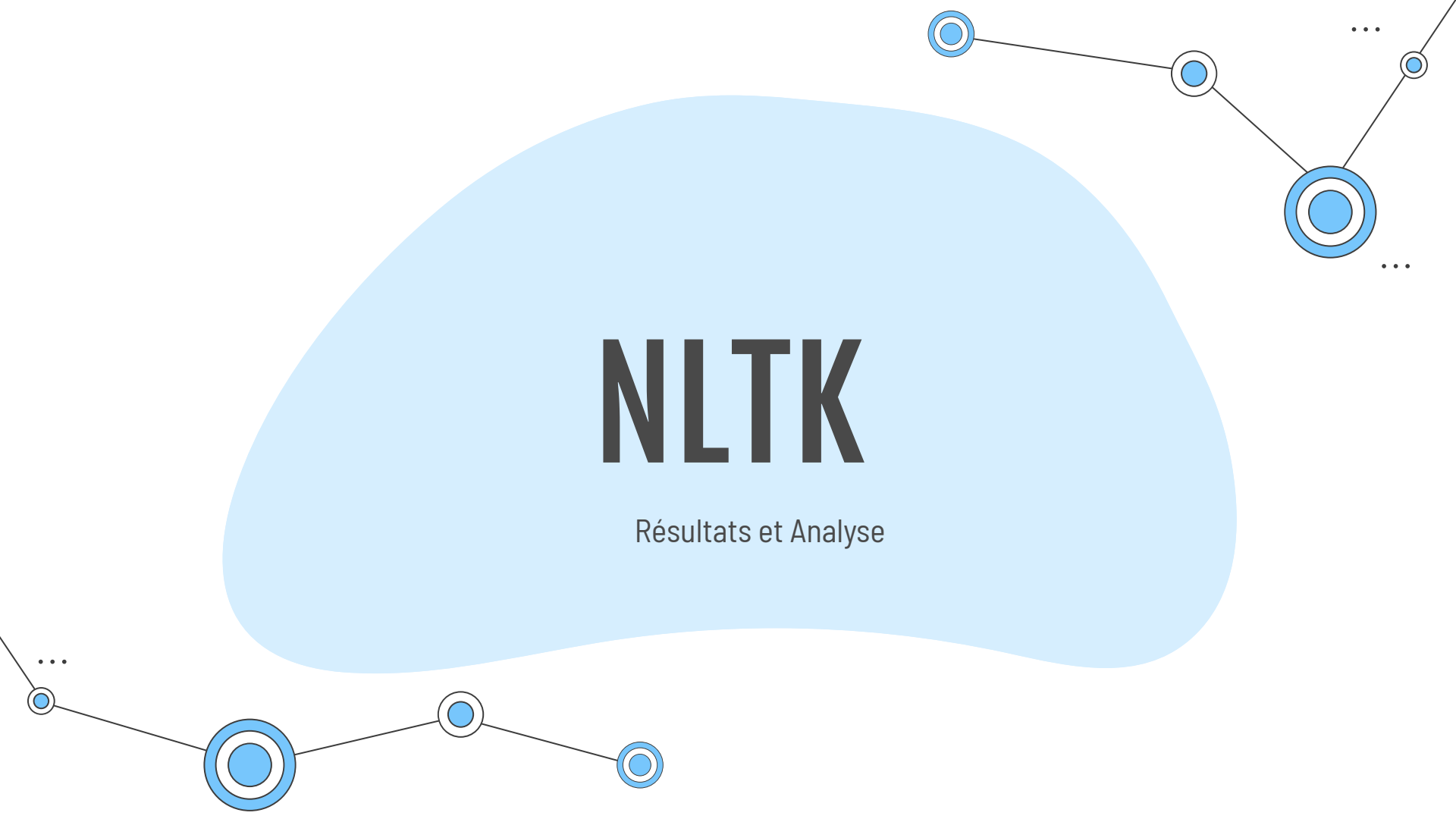
$$\text{Accuracy} = \frac{\text{sensitivity} + \text{specificity}}{2}$$

TF-IDF + NaiveBayes

- **Sensitivity** = 0.80
- **Specificity** = 0.70
- **PPV** = 0.80
- **NPV** = 0.70
- **Accuracy** = 0.75
- **Temps** de classification 92.938 secondes

NLTK

Résultats et Analyse



TF-IDF + NaiveBayes

Temps d'exécution 24.441

	precision	recall	f1-score	support
-1	0.53	0.81	0.64	90
0	0.95	0.81	0.88	384
1	0.00	0.00	0.00	0
accuracy			0.81	474
macro avg	0.49	0.54	0.51	474
weighted avg	0.87	0.81	0.83	474

Matrice de confusion

```
[[ 73  16   1]
 [ 64 312   8]
 [  0   0   0]]
```



05

Comparaison et
Conclusion












Conclusion : Comparaison des deux implémentations

Pyspark

NLTK

Main Competitors

	PySpark	NLTK
		
		
		

Sentiments des personnes sur la crise

Négatifs

Colère, incompréhension,
la fatigue

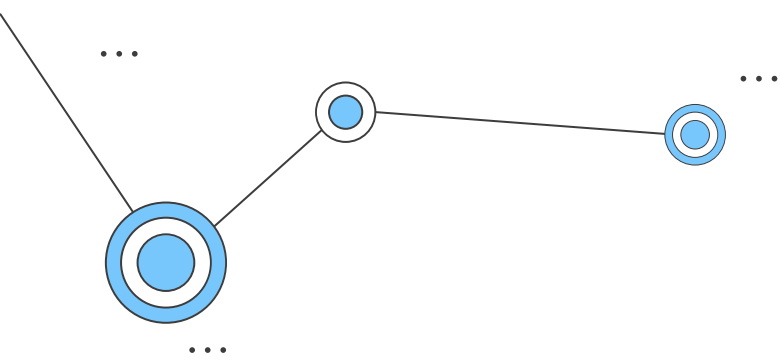
Neutre

Ceux qui s'en foutent,
ceux qui se font du profit

Positifs

Le gouvernement





Merci pour
votre attention!

