Twitter
Topic classification.

BEKKAR Zakaria - CHRIMNI Walid

Sommaire.

Problématique & But du projet.

- **Depuis 2 ans** : une grande quantités de tweets sur le covid
  - réflexion
  - opinion
  - idée
  - actualités
  - sentiments
  - news

- **But du projet** : tweet modelling en fonction de leur topic
  - **input** : tweet
  - **output** : topic du tweet parmi 15 topics

- Méthodologie pour déterminer les topics sera expliquée plus tard

Source de données & modèles utilisés.

- **Tweets récupérés en direct sur Twitter avec :**
    - **Kafka**
    - **Tweepy**
    - **API Twitter**

- **Chaque 15 secondes, 100 tweets datant des 15 dernières secondes sont récupérés**

- **Tweets anglais contenant le mot "covid" sont récupérés**

Source de données & modèles utilisés.

- **Tweets transformés en vecteur :**
  - **de 400 composantes**
  - **via un modèle word2vec pré-entrainé sur 400M tweets anglais**
  - **avec nettoyage, suppression des stops words et lemmatisation.**

- Exemple de nettoyage :
  - **Tweet non-nettoyé** : RT @WakeAwake1: I don't know of a single unvaccinated person who got myocarditis from Covid. Do you? 🤔
  - **Tweet nettoyé** : i know single unvaccinated person got myocarditis covid do

Data analysis.

Online word cloud (actualisés toutes les 5 secondes)
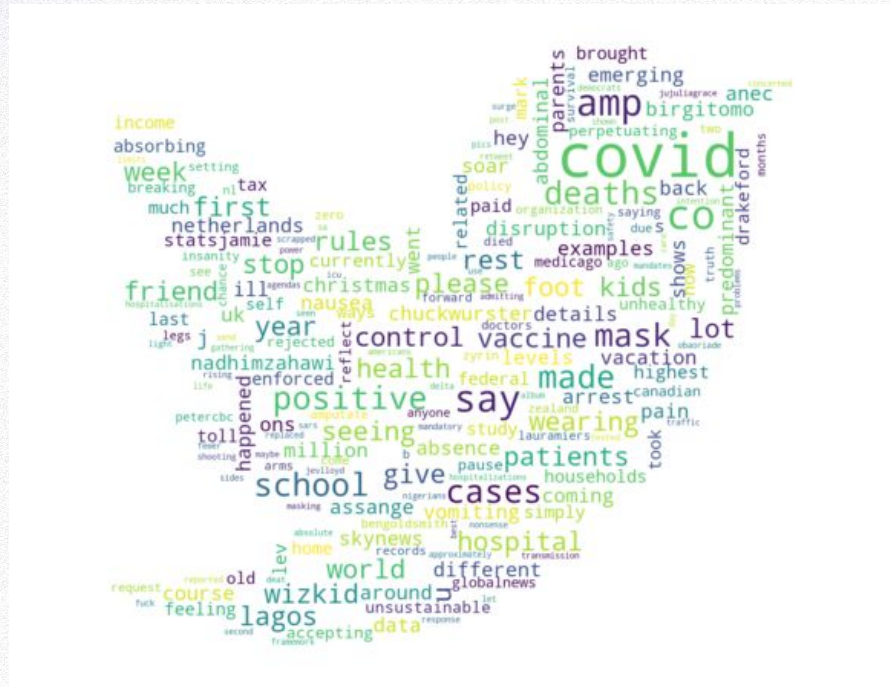
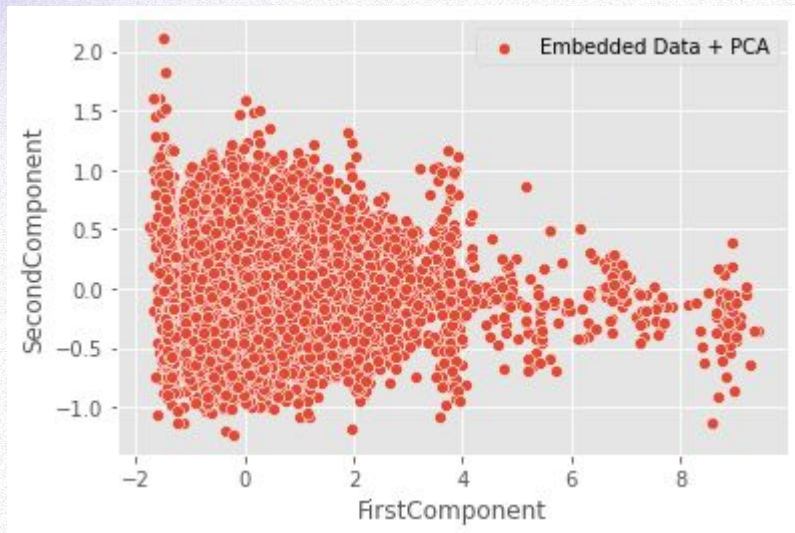# Méthodologie suivie.

30 000 tweets et leur embeddings récupérés (sur une période d'environ 2h)

- **Ces tweets vont nous servir à :**
  - faire une **PCA en 2D**
  - faire une **PCA en 3D**
  - créer les **clusters** qui vont définir les **topics** (**MiniBatckKMeans** et **partial_fit** method)

Ensuite, on fera de l'online classification sur des tweets avec ces clusters

PCA 2D & 3D.

Création des clusters.

- **Utilisation de KMeans pour créer 15 clusters différents**

- **Chacun des 30 000 tweets est associés à un cluster**

- **Pour chaque cluster, les 10 mots avec le score TF-IDF le plus élevé sont récupérés**

- **Le dernier de ces 10 mots est désignés comme topic du cluster auquel il est assignés**

		0	1	Class
4	years	0.013299		0
2	sacked	0.061503		1
9	feel	0.018014		2
3	nhs	0.021411		3
4	doug	0.541858		4
6	ivermectin	0.469384		5
8	bennyjohnson	0.018433		6
1	myocarditis	0.192534		7
5	ordered	0.502942		8
5	deaths	0.017067		9
3	mask	0.019129		10
7	skynews	0.211171		11
3	nausea	0.077289		12
8	vaccines	0.023123		13
9	footpath	0.000000		14

Création des clusters.
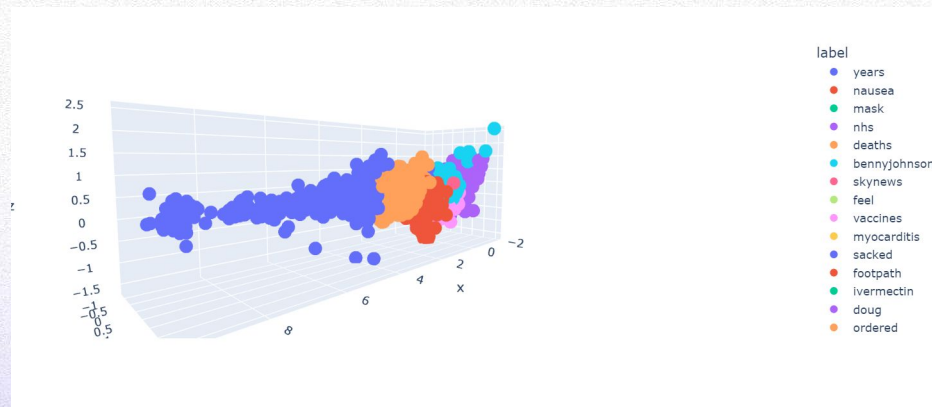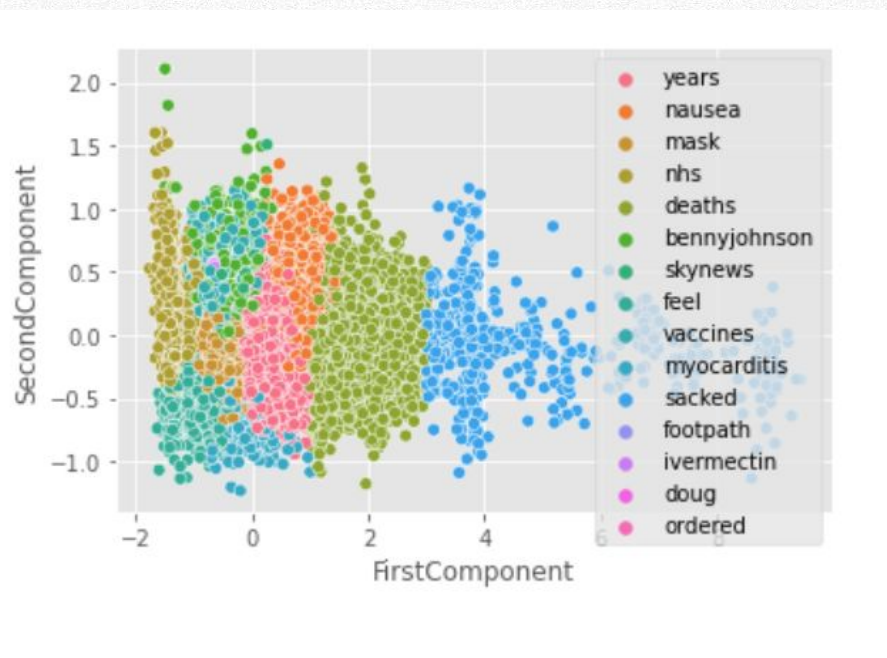
Online topic classification.

- **Toutes les 15 secondes :**
  - **100 tweets** sont récupérés
  - **Classifiés dans l'un des 15 topics cités précédemment**
  - ajoutés à un dataframe de prédiction qui contient le **tweet et sa prédiction**

1		tweet	prediction
2	0	RT @L0gg0l: U.S COVID DEATH TOLL RISES ABOVE 1 MILLION -- @WORLDOMETER DATA	years
3	1	RT @MAGAJew2: My friend was on the Covid frontlines since March 2020. She was fired by Eric Adams for refusing the jal	mask
4	2	RT @RishiSaunam: Edmonton woman who died of vaccine-induced blood clot was turned away from ER - F Big Pharma"Th	mask
5	3	@Vote_CNP If Trump had done exactly the same things as Biden on covid, every democrat in America would be outraged. E	feel
6	4	RT @ClayTravis: In 2020 more people under 65 in the United States died of alcohol related deaths than died with covid. http	nausea
7	5	@ThePoorMan95 @RealDoctorT @theresamtequil1 @BrandonTrump_ @TetraQammon @Mountain_Buddy @prynne_hyste	sacked
8	6	RT @BNODesk: ZOE COVID Study estimates daily coronavirus cases in the UK are at a record-high, but official figures are	vaccines
9	7	RT @SeeLuluRun: My whole timeline right now is people sharing they've tested positive for COVID. Post-pandemic times m	years
10	8	First, COVID symptoms resembled my regular seasonal allergies and now they resemble my daily IBS symptoms. I can truly	vaccines
11	9	RT @dlbernson: Evidence grows that vaccines lower the risk of getting long COVID https://t.co/j9TPDAJf7K	nausea
12	10	RT @DrJBhattacharya: This is an explosive interview between @danwootton and Sky News & ITV former executive Ma	years
13	11	RT @MajorDMalpas: Thank goodness Nicola Sturgeon and Devi Sirdar completely eradicated Covid from Scotland that time	years
14	12	RT @BarbLarochelle: Kenney calls members of his own party bugs, kooky, fringe element, extreme, hateful, intolerant, bigo	mask
15	13	RT @briallysia: my surgery was pushed back due to covid , so im again asking if you could donate or spread this to help me	years
16	14	Significant Increase In COVID Cases – As New Round Of Vaccinations Begin #DumfriesandGallowayhttps://t.co/99sQV3VW	years
17	15	I'm crafting the perfect post that will end the covid-19 pandemic	years
18	16	RT @MuseWendi: are you uninsured but need a covid shot/booster? please make sure you take care of it asap bc the govt	years
19	17	RT @DrFrancesRyan: I see the news weekly covid cases in the U.K. have shot up by 1 million and I think of all the clinically	myocarditis
20	18	RT @pblest: There's not gonna be a better take today than "invading other countries is a possible symptom of long COVID"	years
21	19	RT @ProfMangen: « A country that so readily forgets its dead is surely prone to also forgetting the lessons of the all-too-rec	mask
22	20	In covid times Kejriwal should have advertised in YouTube instead of spending 293 crore on advertisement on newspaper an	sacked
23	21	RT @SaffronHeart02: " Not just making Fun of Kashmir Genocide..He Has also Crossed All The Limits of Looting People! Fc	nhs

Time for the demo!