

EVENT DETECTION IN FINANCE USING HIERARCHICAL ALGORITHMS ON NEWS AND TWEETS

Salvatore Carta, Sergio Consoli, Luca Piras, Alessandro Sebastian et Diego Reforgiato Recupera

Fait par Maeva N'GUESSAN, Roland DUTAUZIET et Lina RAGALA

Aperçu

Le papier propose une approche innovante pour détecter en temps réel les événements financiers en combinant les données des articles de presse et des microblogs. Les articles de presse fournissent une information fiable tandis que les microblogs comme Twitter, malgré leur bruit, reflètent bien les émotions et réactions du public. L'objectif est d'améliorer la détection d'événements et leur compréhension en intégrant ces deux types de sources.

Objectifs

- Optimiser la détection d'événements financiers en temps réel et analyser leur impact social.
- Améliorer la qualité de l'information en combinant les données de sources différentes.
- Développer des alertes d'événements personnalisées.

Méthodes

- Analyse lexicale.
- Clustering hiérarchique des articles de presse.
- Analyse de sentiments sur les mots des articles de presse.
- Assignation des tweets aux clusters.
- Génération d'alertes en temps réel

Plan

1

INTRODUCTION

2

DONNÉES

3

MÉTHODES

4

RÉSULTATS

1 - Élaboration du lexique

Collecte d'articles sur le S&P 500 (2 à 4 semaines avant le jour j de l'analyse)

Art. 1: "S&P 500 Rises Amid Tech Rally"

The S&P 500 gained 0.5% today, driven by strong performances in the technology sector. Investors remain optimistic about upcoming earnings reports.

Art. 2: "Energy Stocks Drag S&P 500 Lower"

The S&P 500 fell 0.7% as energy stocks declined, following a drop in oil prices. Concerns over global demand weighed on market sentiment.

Art. 3: "S&P 500 Hits Record High"

The S&P 500 reached a new all-time high, buoyed by healthcare and consumer discretionary stocks. This reflects growing confidence in economic recovery.

Art. 1: ['s', 'p', 'gain', 'today', 'driven', 'strong', 'perform', 'technolog', 'sector', 'investor', 'remain', 'optimist', 'upcom', 'earn', 'report']

Art. 2: ['s', 'p', 'fe l', 'energi', 'stock', 'declin', 'fo low', 'drop', 'oil', 'price', 'concern', 'global', 'demand', 'weigh', 'market', 'sentiment']

Art. 3: ['s', 'p', 'reach', 'new', 'a l', 'time', 'high', 'buoy', 'healthcar', 'consum', 'discretionari', 'stock', 'reflect', 'grow', 'confid', 'econom', 'recoveri']

Les valeurs correspondent à la variation du prix de l'action après la publication de l'article

	Article 1	Article 2	Article 3
discretionari	0	0 0.47	
sentiment	0 -0.21	0	
s	0.34	-0.21	0.47
fell	0 -0.21	0	
today	0.34	0	0
global	0 -0.21	0	
grow	0	0 0.47	
report	0.34	0	0
sector	0.34	0	0
weigh	0 -0.21	0	
drop	0 -0.21	0	
follow	0 -0.21	0	
declin	0 -0.21	0	
demand	0 -0.21	0	
driven	0.34	0	0
strong	0.34	0	0
p	0.34	-0.21	0.47
perform	0.34	0	0
confid	0	0 0.47	
reach	0	0 0.47	
buoy	0	0 0.47	
stock	0 -0.21	0.47	
time	0	0 0.47	

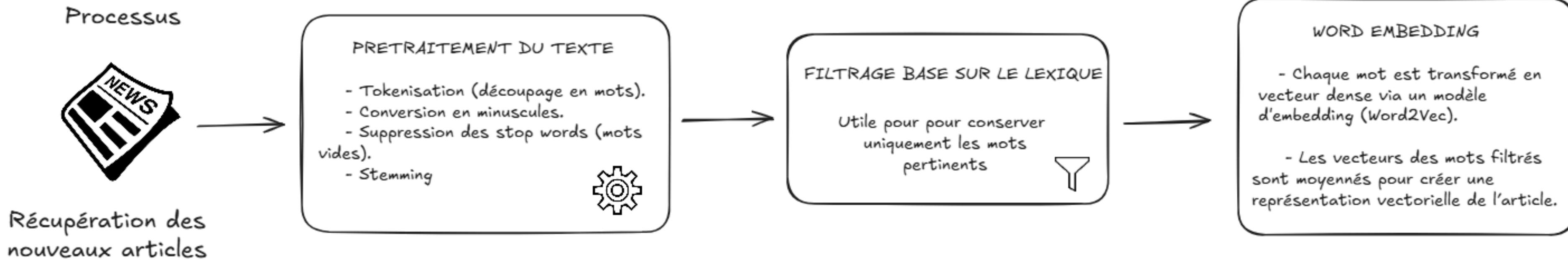
Datasets utilisés

Datasets	<u>Dow Jones DNA</u>	<u>Stocktwits</u>	<u>S&P 500 time-series</u>
Contenu principal	Articles de presse sur la finance, le business et les actualités	Tweets financiers organisés autour de cashtags (\$SPX)	Indice boursier des 500 plus grandes entreprises américaines (US)
Période/Fréquence	Immédiate à Mensuelle	Juin 2016 – Mars 2020	Quotidienne(2016 à 2020)
Volume utilisé	8403 articles	283 473 tweets	Série journalière de l'indice
Utilisation	Analyse textuelle pour identifier les événements impactant les marchés	Analyse de sentiments en temps réel et détection des événements	Corrélations entre les variations de l'indice et les événements dans le monde

2 - Features engineering

Objectif :

- Représenter les articles d'actualité dans un espace vectoriel qui capture leurs significations.
- Focus : Retenir les mots pertinents pour le domaine et ignorer ceux qui n'apportent pas d'information utile.



3 - Clustering hiérarchique agglomératif

Approche ascendante

- Chaque article commence comme un cluster individuel.
- Les clusters sont fusionnés progressivement selon un critère de similarité cosinus.

$$K = \sum \frac{(A_i \times B_i)}{(\sqrt{\sum A_i^2} \times \sqrt{\sum B_i^2})}$$

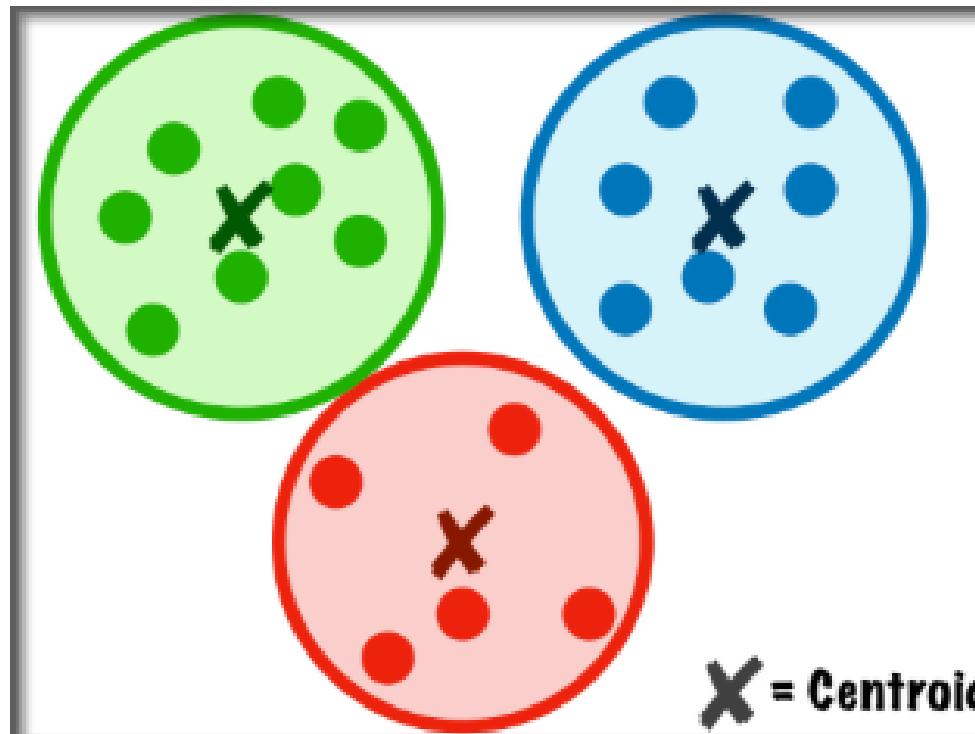


Détermination du nombre optimal de clusters

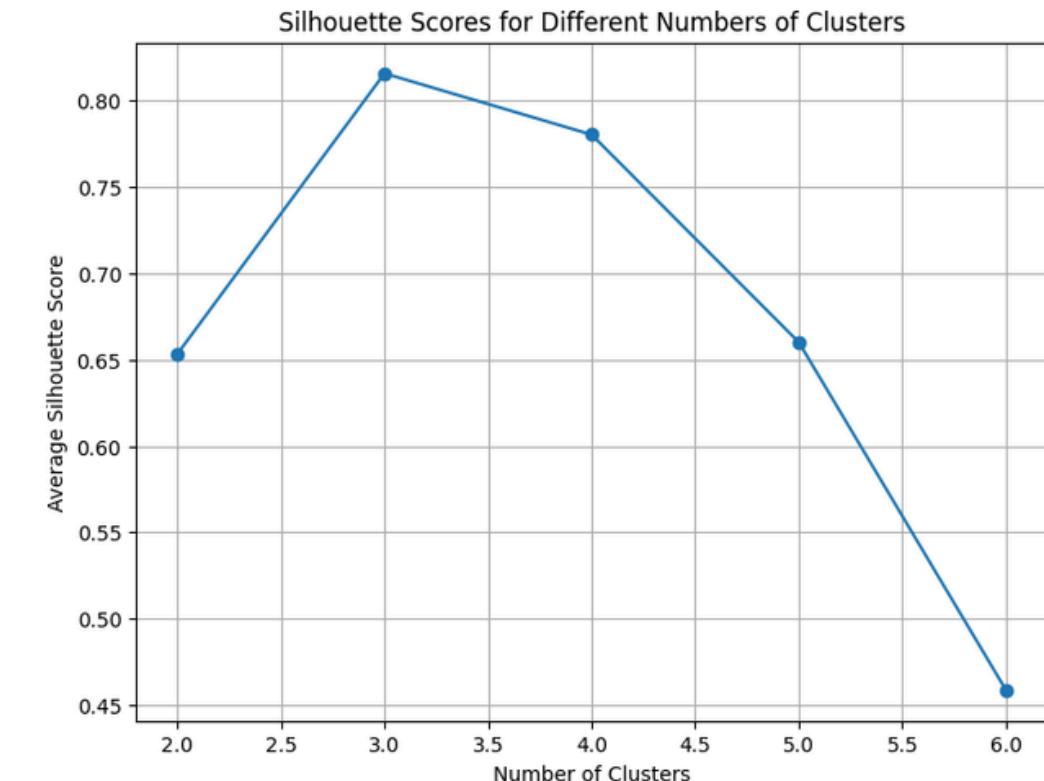
- Utilisation du Score de Silhouette
- Calcule dynamique du nombre optimal de clusters chaque jour

$$s = \frac{b - a}{\max(a, b)}$$

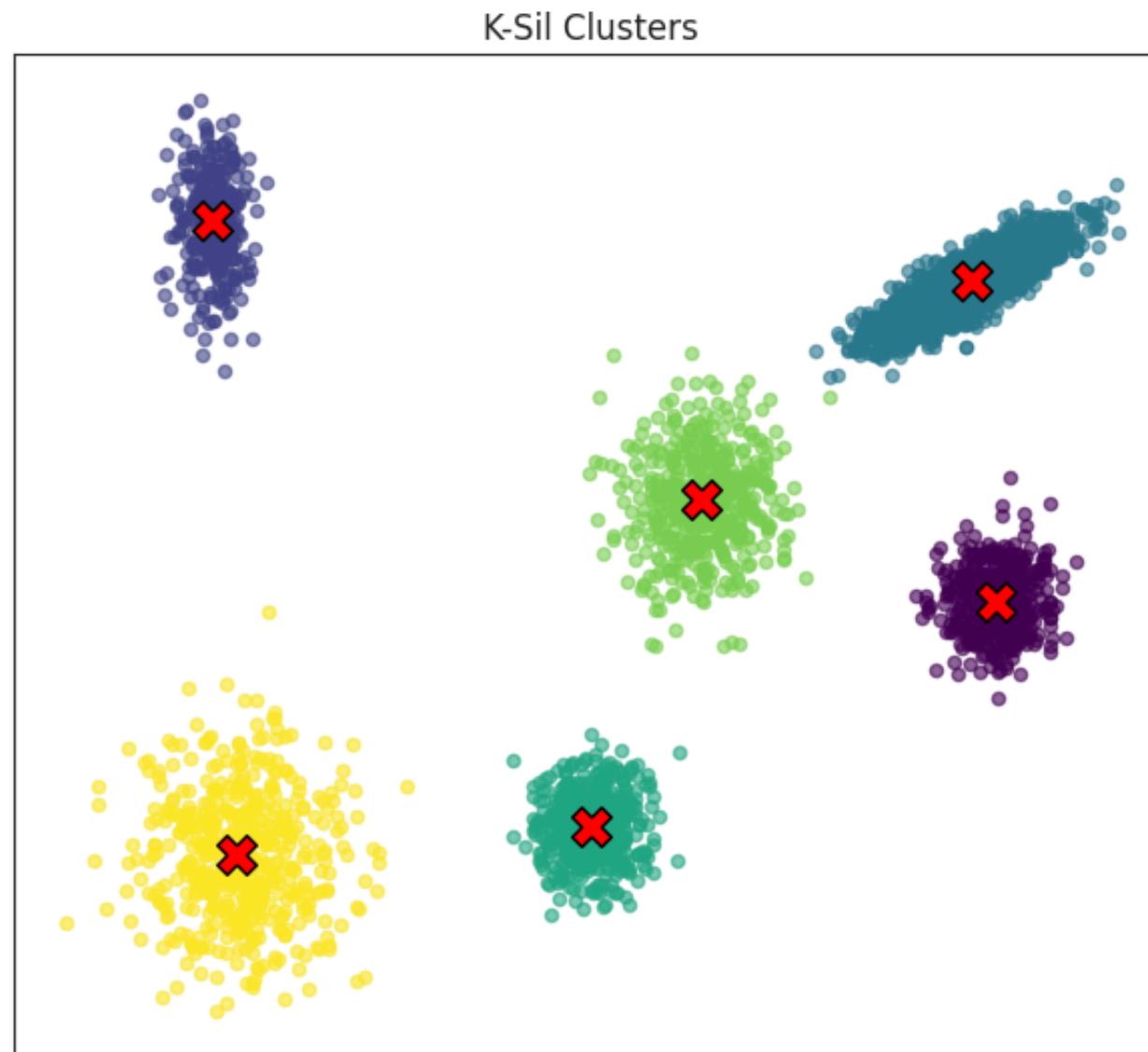
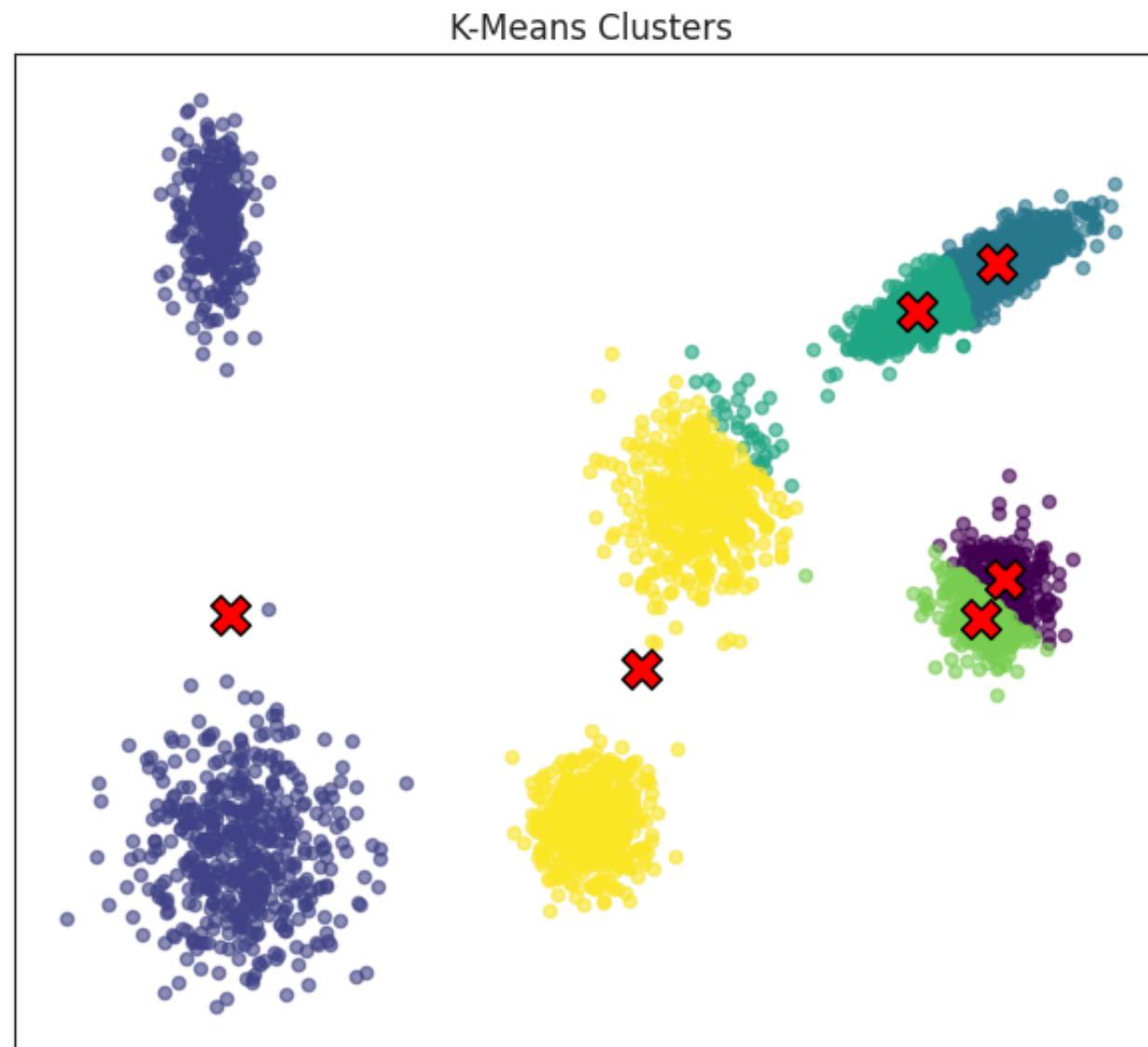
 **Minimisation de la distance moyenne entre toutes les observations des paires de clusters**



 **Maximisation de la moyenne des scores de tous les articles**

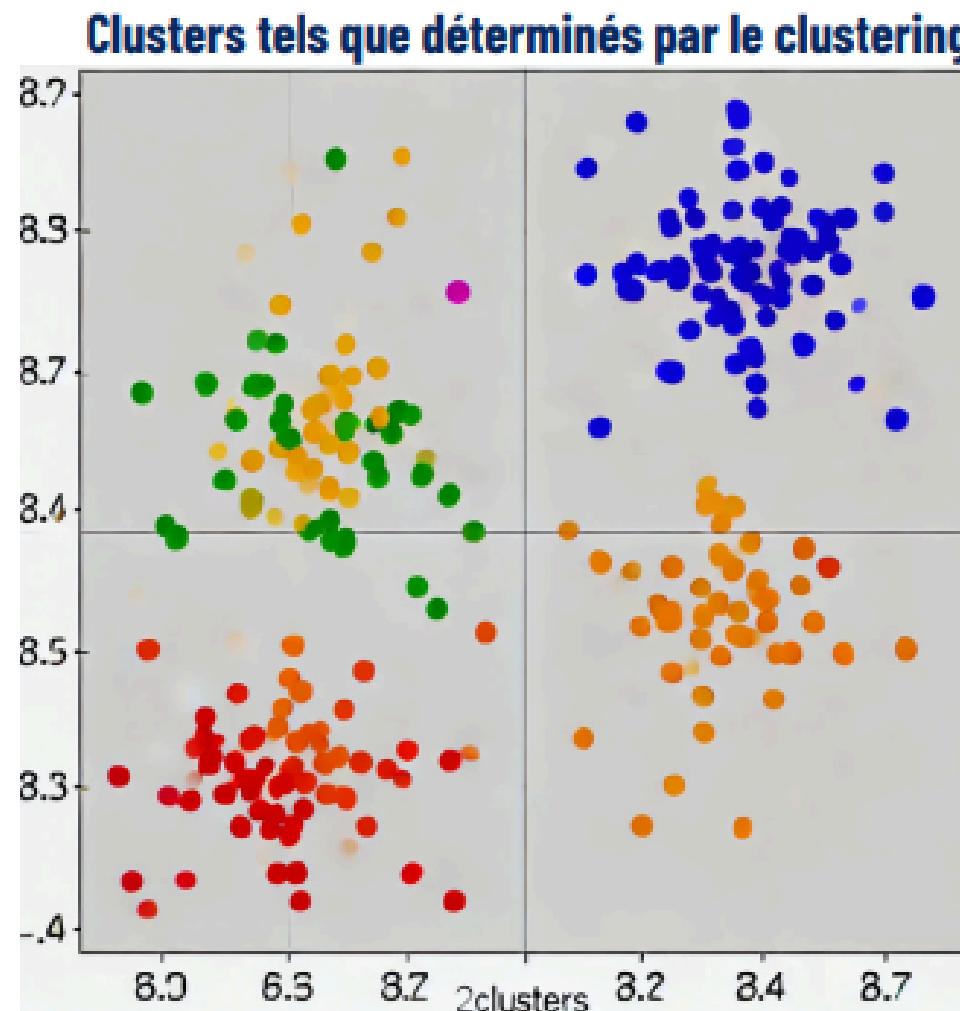


4 - Identification des outliers



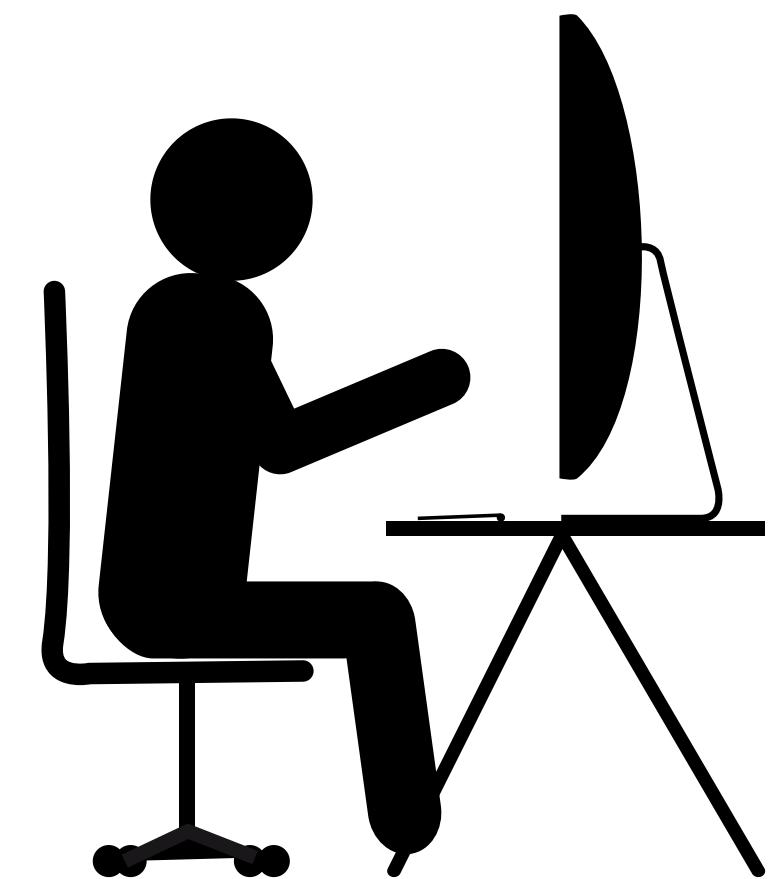
- **Le coefficient de silhouette** pour évaluer la qualité de l'appartenance au cluster
- La **distance cosinus** entre chaque document et le centroïde de son cluster
- Classement des articles selon **score de silouhette** et la distance cosinus Les documents avec les scores inférieurs au seuil de rupture sont éliminés
- Les **centroïdes** des clusters sont recalculés

5 - Extraction des mots pertinents :



Informations pertinentes à extraire des clusters :

- Titres Extraits des articles de presse
- La période couverte par le cluster
- Le pourcentage de mots positifs et négatifs
- La liste des mots pertinents



ÉVALUATION QUALITATIVE DES CLUSTERS EN TERMES DE SENS ET COHÉRENCE PAR L'UTILISATEUR

6 - Extraction des mots pertinents par cluster :

TF-IDF est une mesure qui combine la fréquence d'un mot dans un document (TF) et la rareté de ce mot dans l'ensemble du corpus (IDF) pour évaluer son importance relative.

Transformation des documents du cluster en vecteur avec la méthode TF IDF

"Banques & taux" → [banques, taux, intérêts, prêts, cryptomonnaie, Volatilité, argent]

TF : [0.25, 0.125, 0.125, 0.125, 0, 0,]

IDF : [3, 6, 6, 6]

TF-IDF : [0.75, 0.75, 0.75, 0.75]

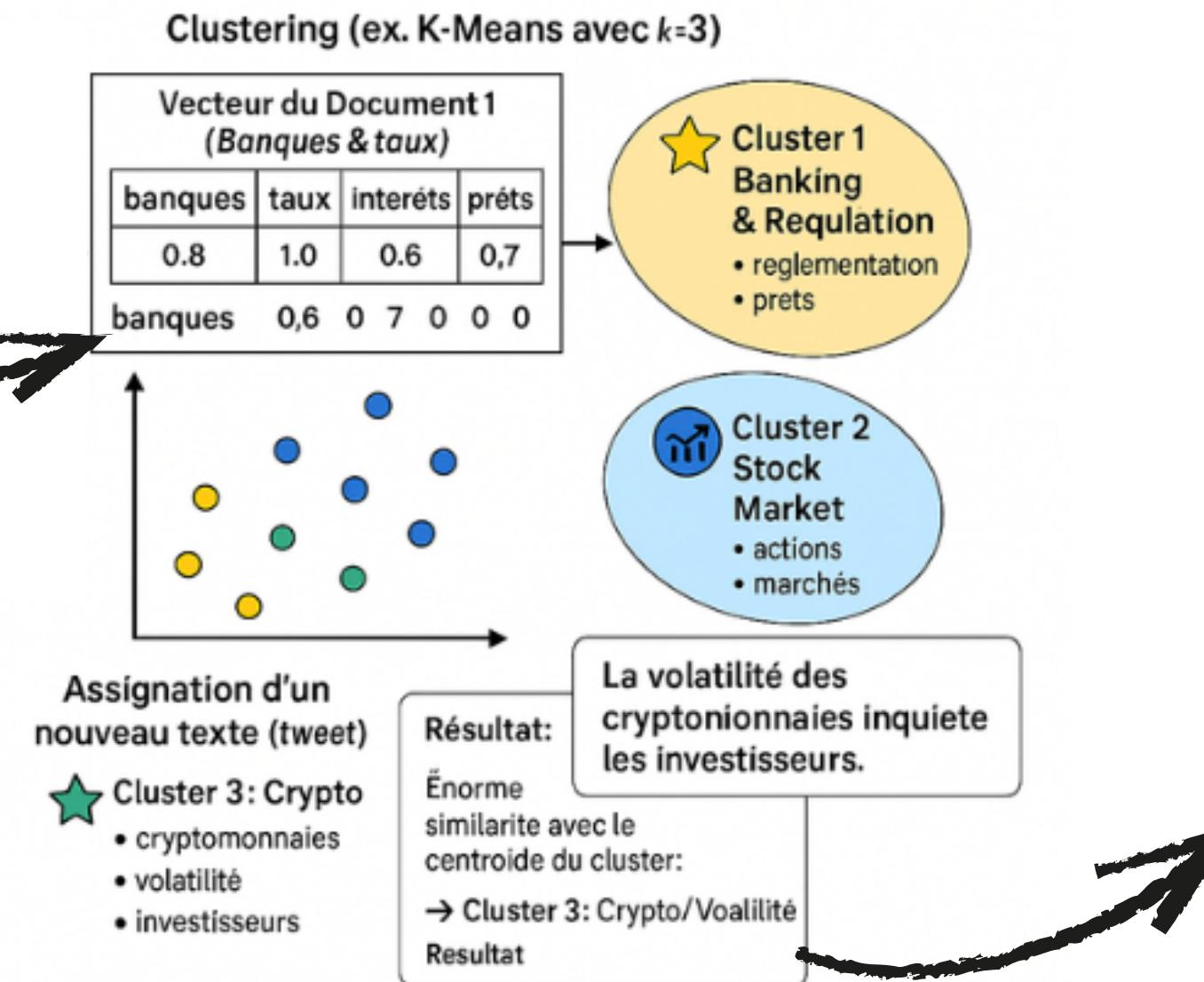
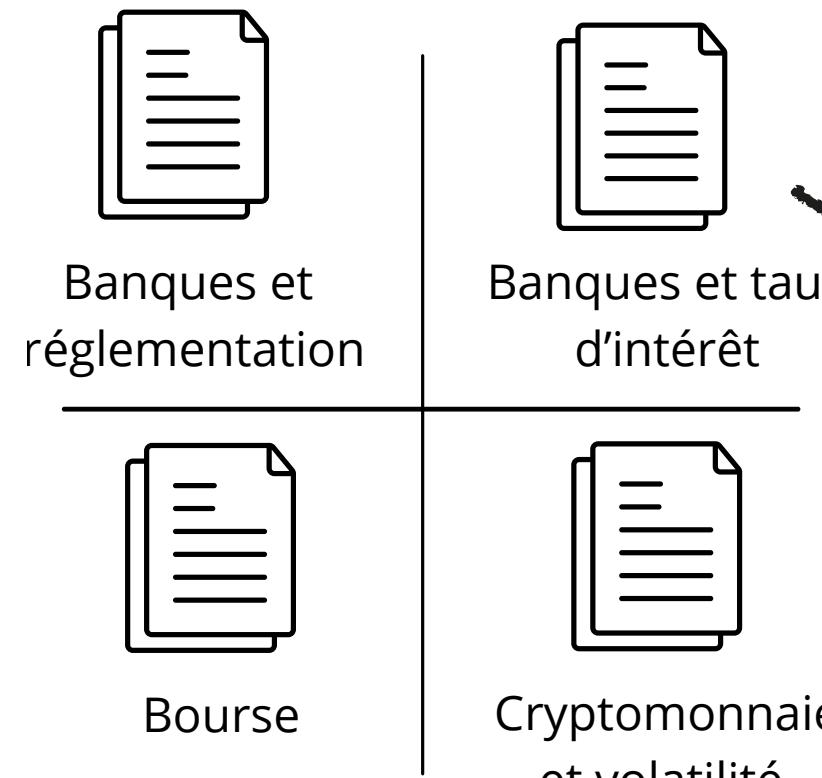
Vecteur : [0.75, 0.75, 0.75, 0.75, 0, 0, 0, 0]

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

Un mot est considéré comme pertinent pour un cluster lorsqu'il présente une moyenne TF-IDF élevée, signifiant qu'il est fréquemment utilisé dans les documents du cluster (TF élevé) tout en étant rare dans le reste des clusters (IDF élevé). Ce poids fait de ce mot un bon discriminant du cluster dans l'espace vectoriel.

7. Exemple de classement d'un tweet dans un des clusters :

Corpus de documents financiers



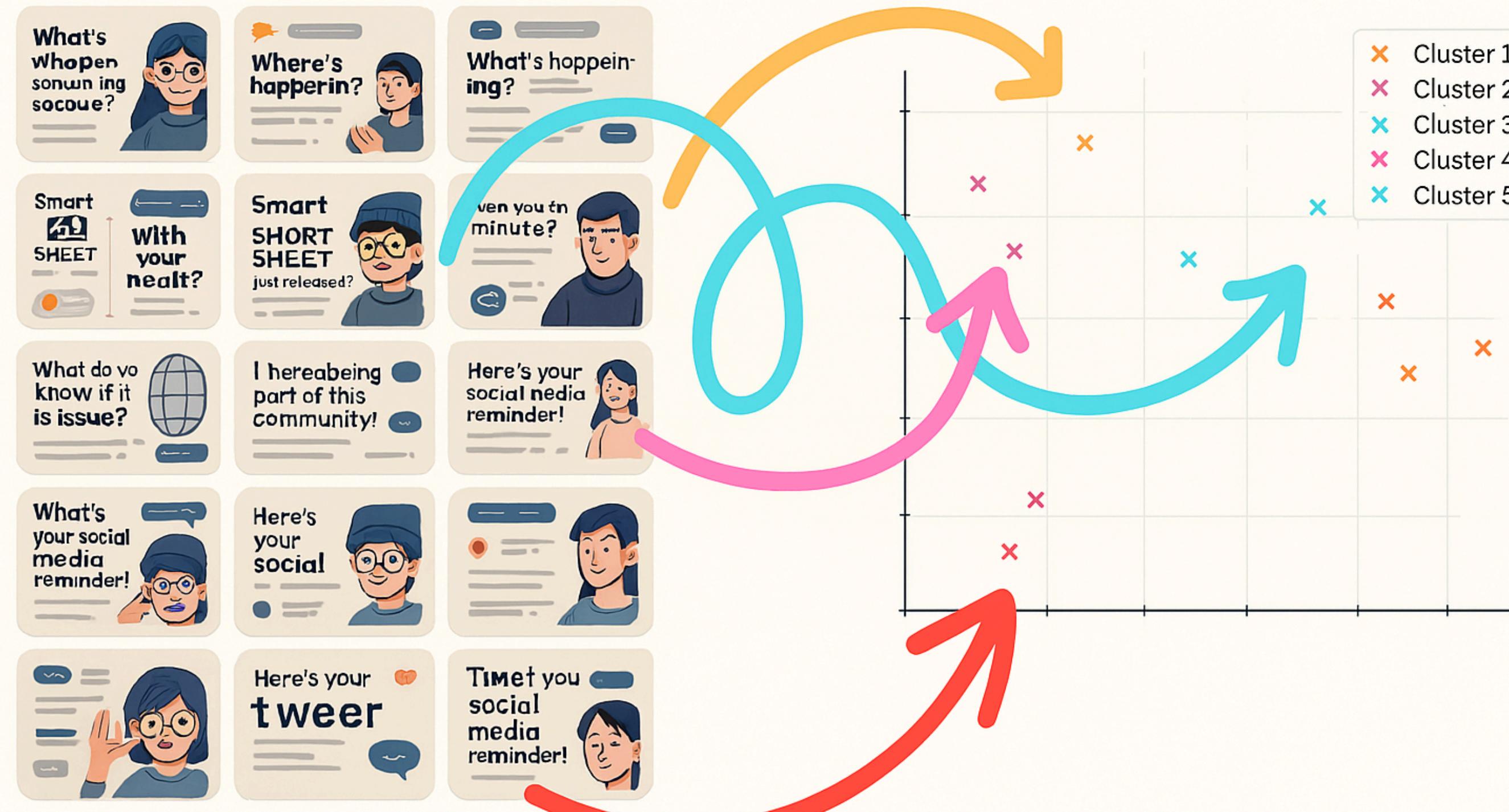
Nouveau Tweet à classer
La volatilité des cryptomonnaie inquiète les investisseurs

Vecteur TF-IDF du tweet

CRYPTOMONNAIE 0,8
VOLATILITÉ 0,9
INVESTISSEURS 0,55

L'embedding de chaque tweet est comparé aux centroïdes des clusters à l'aide de la similarité cosinus. Chaque tweet est associé au cluster le plus proche, à condition que la distance soit inférieure à un seuil prédéfini.

Correspondance entre tweets et clusters



8. Hot Events :



L'idée est de détecter des événements importants (Hot events) qui peuvent avoir un impact sur le marché.

Ces événements sont :

- Rapportés dans les actualités.
- Très discutés sur les réseaux sociaux (ex. Twitter, Stocktwits).

Pourquoi ?

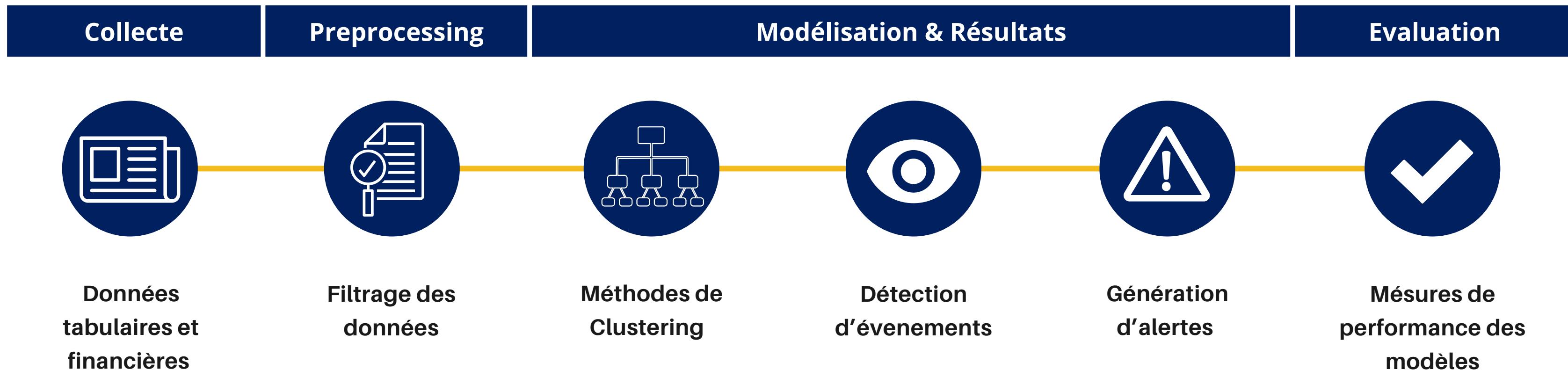
Parce qu'une forte popularité d'un sujet peut influencer les décisions des traders → d'où la nécessité de générer des alertes.

$$\text{Si } \left(\frac{\text{Nombre total de tweets assignés à des clusters}}{\text{Nombre total de tweets (assignés + non assignés)}} \times 100 \right) > \text{Seuil}$$



- On regarde la proportion de tweets qui ont été classés dans des clusters (donc liés à des thèmes identifiés).
- Si cette proportion dépasse un certain seuil (par exemple 60%), cela signifie qu'il y a un sujet dominant et l'alerte est déclenchée !

Etapes de l'expérience



Résultats - Clustering

Métriques de Clustering	Agglomerative	K-Means	K-Medians	K-Medoids
Silhouette Coefficient	0.30	0.27	0.28	0.29
Dunn Index	0.46	0.44	0.43	0.44
Nombre de clusters extraits	5.96	4.18	4.31	4.89
Overlap des mots pertinents	0.04	0.04	0.04	0.04

- Le clustering agglomératif est nettement le meilleur en termes de Silhouette, Dunn Index et de nombre de clusters.
- L'approche agglomerative fournit donc les clusters les mieux séparés et les plus cohérents sémantiquement.

Comparaisons des Silhouette Coefficient (A), Dunn Index (B), Nombre de clusters obtenus (C) et Chevauchement entre les mots pertinents des clusters (D)



Les métriques sont calculées après suppression des outliers, ce qui améliore d'environ 50% Silhouette et Dunn Index.

Résultats - Détection d'événements

Élection présidentielle américaine (2016)

clinton, trump, election, percent, october, team, victory, polls, presidential

- *World Stocks: Dollar, Asia Stocks Rise on Clinton News*
- *Stocks: Election Has Foreign Funds Wary—After Brexit surprise, some avoid investing in U.S. stocks until president determined*
- *The stock market's continual favoritism of Hillary Clinton proves that she has been bought. Corruption loves company.*
- *- Hillary Clinton Wins!*

Guerre commerciale USA-Chine (2019)

tariffs, trump, chinese, united, talks, deal, friday, imports, goods

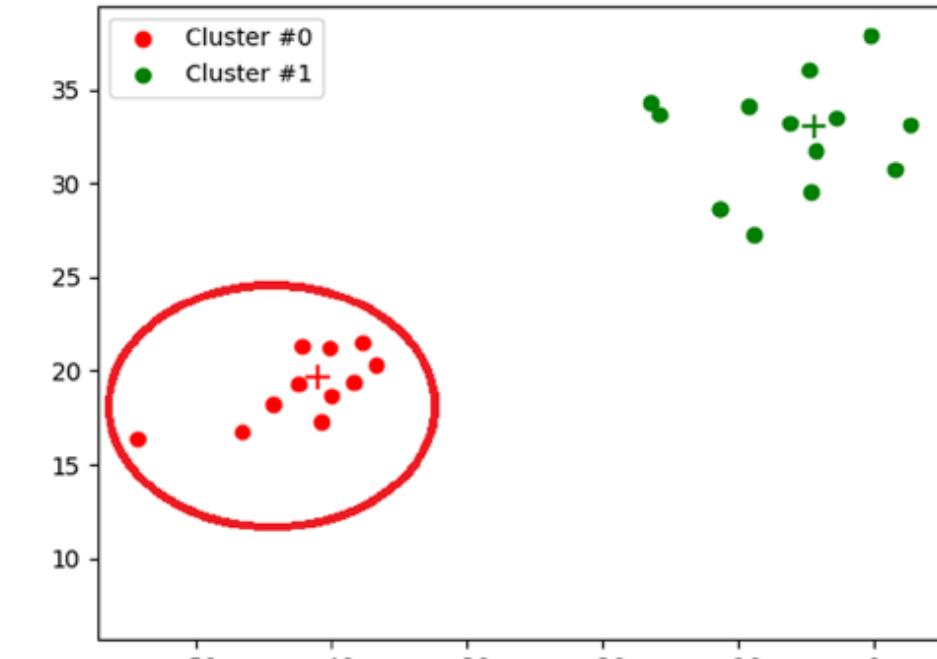
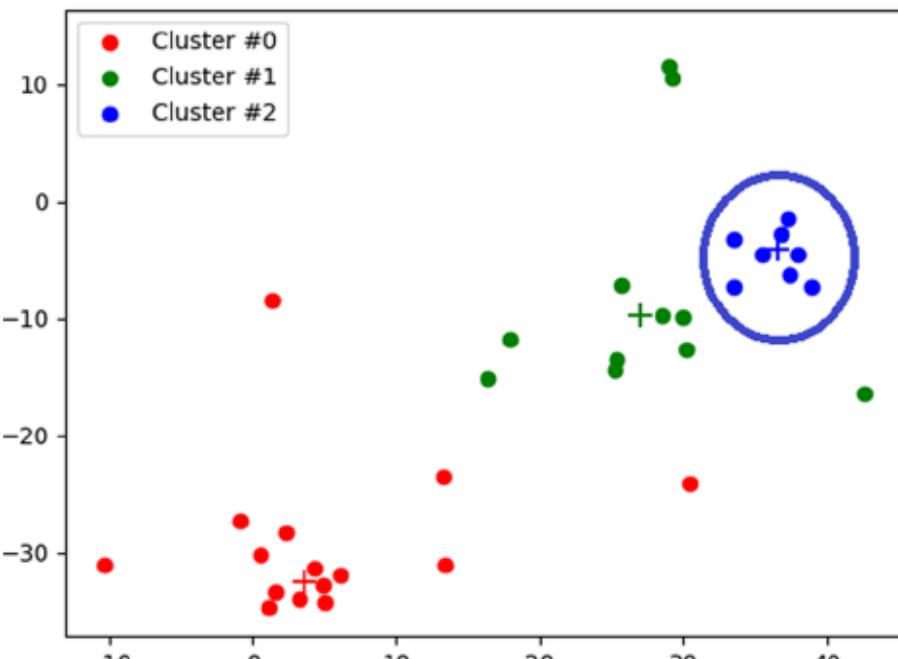
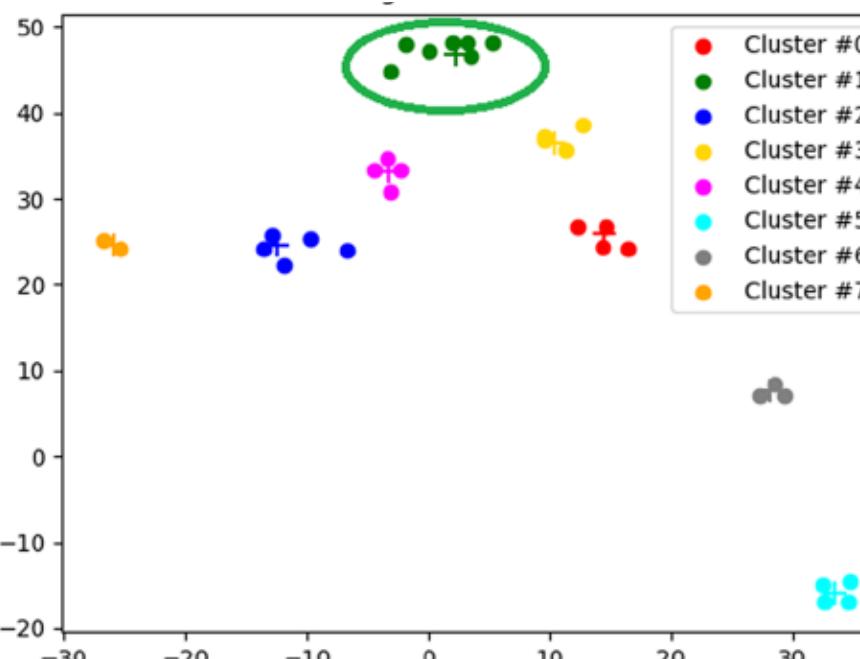
- *Fear of Tariffs Jolts Markets And Nerves*
- *U.S. Advisers Say China Is Reneging On Trade Accord*
- *Tariff increase on Chinese imports will take effect on May 10 Federal Register*
- *"Reuters: Trump's punitive tariffs will burden consumers"; yeah like it...*

Pandémie du Covid-19 (2020)

virus, outbreak, losses, impact, europe, department, boeing, hopes

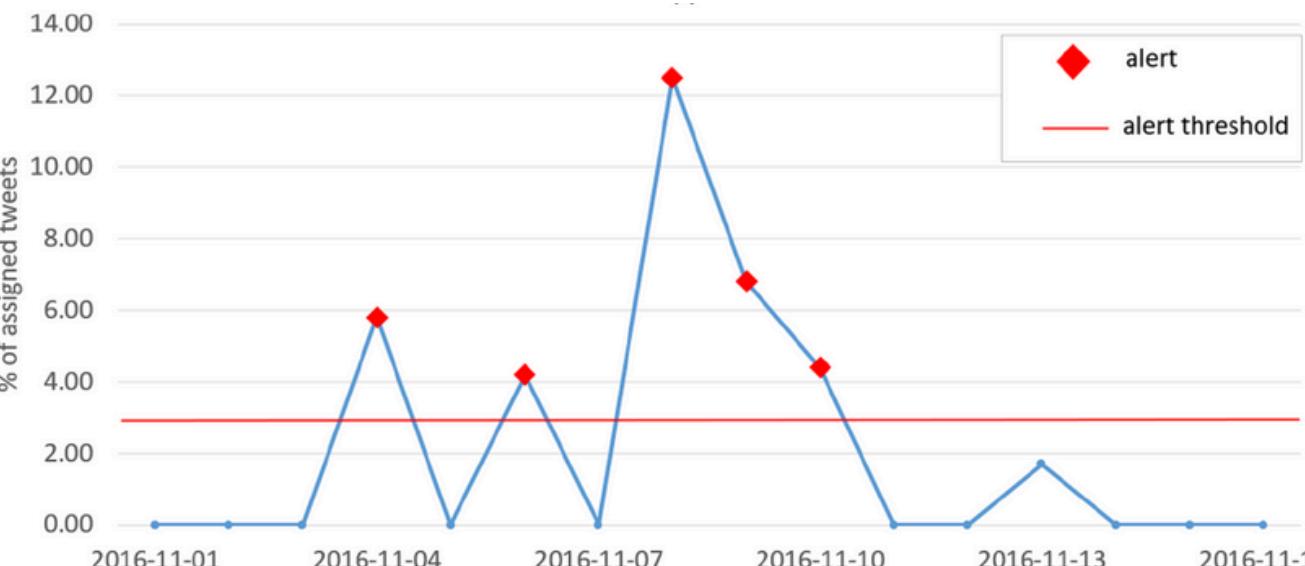
- *Markets on the slide as fears spread over virus*
- *Britons returning from China to be 'safely isolated' for 14 days*

- *Mainland Chinese, Hong Kong stocks tumble as Covid-19 death toll rises*
- *Second U.S. Covid-19 case is Chicago resident who traveled to Wuhan*

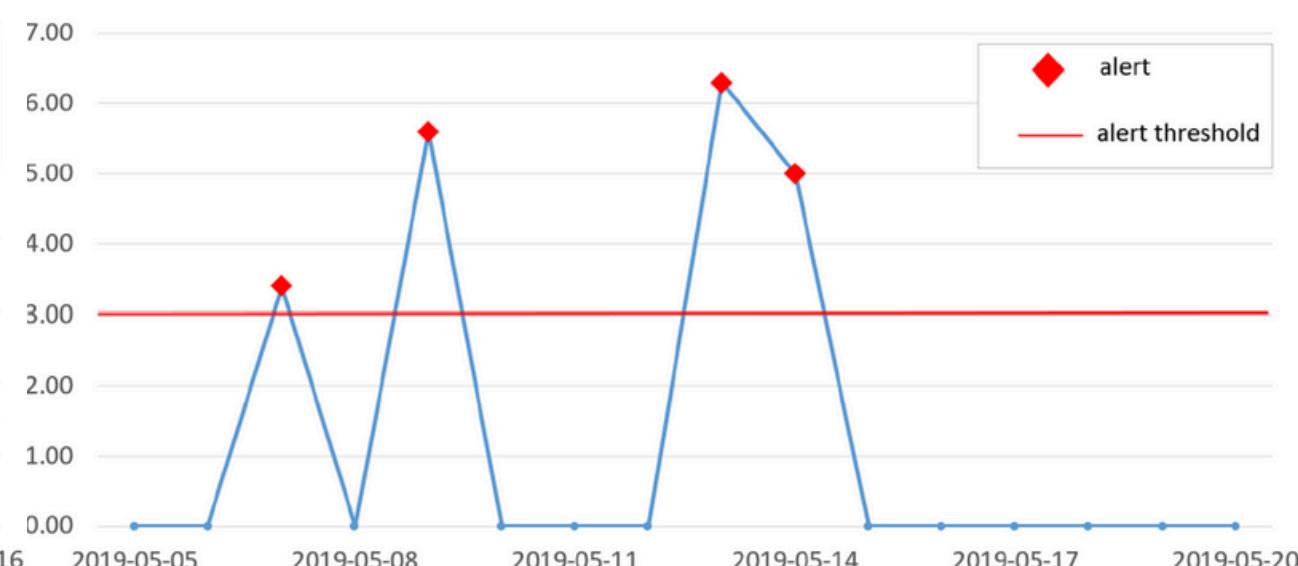


Résultats - Système d'alertes

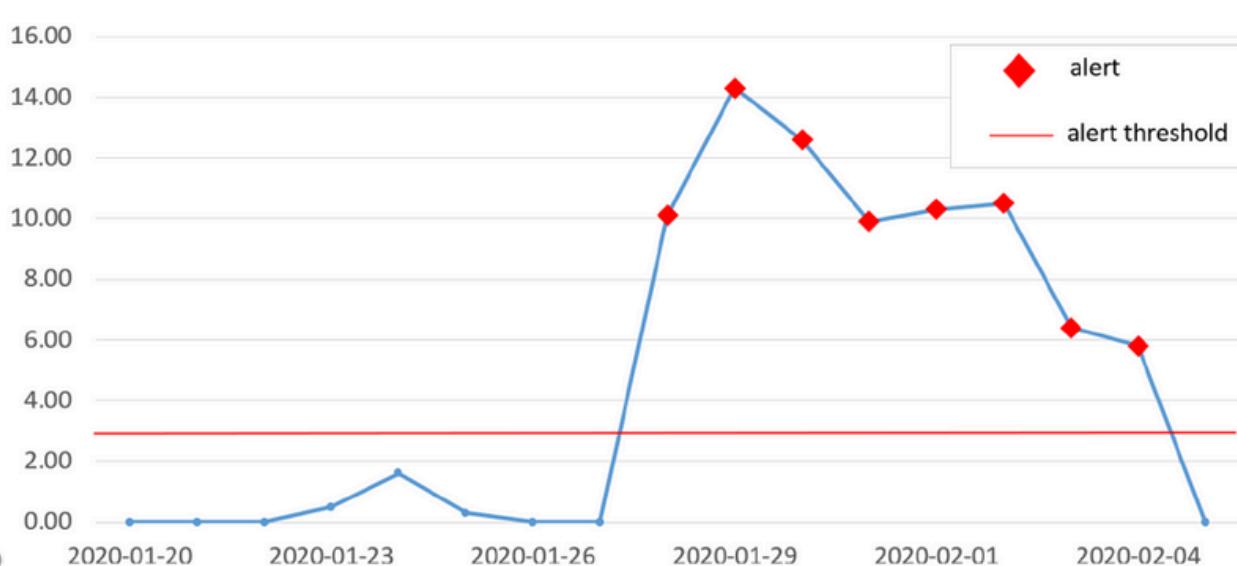
2016 U.S. Presidential Elections



U.S.-China trade war



Covid-19 outbreak



Pourcentage de tweets assignés et alertes associées

- Les marqueurs rouges indiquent les alertes générées, tandis que la ligne rouge horizontale représente le seuil d'alerte.
- La courbe atteint un maximum exactement à la date de l'événement réel.

Seuil de détection d'évenements

$$\Delta_d = \frac{|close_{(d+7)} - close_d|}{close_d} > 2\%$$

Conclusion

