



Université de Nouakchott Al Aasriya
Faculté des Sciences et Techniques
Département Mathématiques et Informatiques



Mémoire de Projet Big Data



Scraping et analyse quantitative du discours politique sur le site ami.mr (2007–2022)

Réalisé par :

Mohamed Lemine Abdallahi Tah

C12896

Encadré par :

Dr Mohamed El Ghaly Beheitt

Année universitaire : 2025 – 2026

Projet réalisé dans le cadre du Master SSD – Statistiques et Sciences des Données

Table des matières

Introduction	2
1 Collecte des données par Web Scraping	3
1.1 Présentation de la source des données	3
1.2 Méthodologie de web scraping	3
1.3 Structure du jeu de données	3
1.4 Préparation pour le stockage NoSQL	4
2 Base de données MongoDB : Réplication et Sharding	5
2.1 Objectifs du stockage distribué	5
2.2 Environnement de travail	5
2.3 Architecture du cluster MongoDB	5
2.4 Serveurs de configuration (Config Servers)	6
2.5 Configuration des shards	6
2.6 Configuration du routeur Mongos	7
2.7 Configuration des shards	8
2.8 Activation du sharding sur la base de données	9
2.9 Stratégie de sharding et répartition des données	10
2.10 Importation et répartition des données	10
3 Visualisations des données	11
3.0.1 Top 30 des mots les plus fréquents	11
3.0.2 Analyse de sentiment moyen par année	11
3.0.3 Évolution du nombre absolu de mentions des thèmes politiques	12
3.0.4 Évolution du cadrage du mot « corruption »	13
3.0.5 Mentions de « démocratie » et « armée »	13
3.0.6 Langage de crise vs langage de stabilité	14
3.0.7 Discours politique vs Action politique	14
Conclusion	15

Introduction

L’essor du numérique et la multiplication des sources d’information en ligne ont conduit à une croissance rapide du volume de données disponibles, en particulier dans le domaine médiatique. L’exploitation efficace de ces données nécessite des techniques avancées de collecte automatisée, de stockage distribué et d’analyse statistique, au cœur des compétences en Statistiques et Sciences des Données.

Dans ce contexte, ce projet vise à mettre en œuvre une chaîne complète de traitement des données, allant du web scraping jusqu’à l’analyse statistique et la visualisation, en s’appuyant sur des données réelles issues du site officiel de l’Agence Mauritanienne d’Information (AMI).¹

Les données utilisées dans ce travail sont constituées de **9 042 articles politiques** collectés automatiquement à partir du site AMI.mr. Ces articles couvrent la période **2007–2022** et appartiennent exclusivement à la catégorie *Politique*. Les données ont été initialement extraites sous forme de fichiers **CSV**, puis transformées en format **JSON** afin de faciliter leur stockage et leur manipulation dans une base de données NoSQL.

Afin de répondre aux problématiques de volume, de performance et de disponibilité, une base de données **MongoDB** distribuée a été mise en place en combinant les mécanismes de *sharding* et de *réplication*. Cette architecture permet d’assurer une répartition équilibrée des données, une tolérance aux pannes et une scalabilité horizontale adaptée aux grands volumes d’informations.

Enfin, le projet s’achève par une phase d’analyse statistique et de visualisation des données, visant à mettre en évidence des tendances temporelles, des répartitions thématiques et des caractéristiques textuelles des articles politiques mauritaniens. Les résultats obtenus illustrent l’intérêt des bases de données distribuées et des outils d’analyse dans le cadre de l’exploitation de données massives issues du web.

1. <https://ami.mr>

Chapitre 1

Collecte des données par Web Scraping

1.1 Présentation de la source des données

Les données utilisées dans ce projet proviennent du site officiel de l'Agence Mauritanienne d'Information (AMI), accessible à l'adresse : <https://ami.mr>. Ce site constitue la principale source d'information institutionnelle en Mauritanie et publie quotidiennement des articles couvrant plusieurs thématiques telles que la politique, l'économie, la société et la culture.

Dans le cadre de ce travail, la collecte s'est concentrée exclusivement sur les articles appartenant à la catégorie *Politique*, publiés entre les années 2007 et 2022.

1.2 Méthodologie de web scraping

La collecte automatisée des données a été réalisée à l'aide du langage **Python**, en utilisant principalement les bibliothèques suivantes :

- **requests** pour l'envoi des requêtes HTTP et **BeautifulSoup** pour l'analyse et l'extraction du contenu HTML
- **pandas** pour la structuration et le stockage des données et **concurrent.futures** pour le traitement parallèle

Pour chaque page, les informations principales de chaque article sont extraites, puis le contenu détaillé de l'article est récupéré séparément afin d'optimiser les performances.

Afin de limiter la charge sur le serveur cible et d'éviter tout blocage, des délais aléatoires ont été introduits entre les requêtes, et un mécanisme de reprise automatique (*retry*) a été mis en place en cas d'échec temporaire des requêtes.

1.3 Structure du jeu de données

Les données collectées ont été initialement stockées dans un fichier au format **CSV**, chaque ligne représentant un article. Les champs suivants ont été extraits et conservés :

- **tag** : catégorie de l'article (Politique)
- **title** : titre de l'article
- **url** : lien URL vers l'article original
- **datetime** : date et heure de publication

- **summary** : résumé de l'article
- **article_body** : contenu textuel principal de l'article

Il est important de souligner que, bien que le champ **article_body** contienne le texte intégral de l'article, celui-ci n'est pas reproduit dans le présent rapport afin d'éviter toute surcharge textuelle.

Aperçu du jeu de données

La Figure 1.1 présente un extrait du fichier CSV obtenu après la phase de web scraping. Elle illustre la structure tabulaire du jeu de données ainsi que les différents champs collectés.

	A	B	C	D	E
	title	url	datetime	summary	article_body
1	title	url	datetime	summary	article_body
2	Les partis de la majorité commémor	https://ami.mr/fr/archives/2022-08-06	2022-08-06	Les partis de la majorité ont commémoré, vendredi soir, le	Les partis de la majorité ont commémoré, vendredi soir, le troisièm
3	La présidente du Conseil régional de	https://ami.mr/fr/archives/2022-07-30	2022-07-30	La présidente du Conseil régional de Nouakchott, Mme Vatimetou	La présidente du Conseil régional de Nouakchott, Mme Vatimetou
4	Célébration de la journée mondiale d	https://ami.mr/fr/archives/2022-07-30	2022-07-30	Le Commissaire adjoint aux droits de l'homme, à l'Action humanita	Le Commissaire adjoint aux droits de l'homme, à l'Action humanita
5	Le ministre de l'Équipement et des T	https://ami.mr/fr/archives/2022-07-24	2022-07-24	Le ministre de l'Équipement et des Transports, M. Mactar Ould Ah	Le ministre de l'Équipement et des Transports, M. Mactar Ould Ah
6	Sélibabi : Atelier pour la coordinati	https://ami.mr/fr/archives/2022-07-21	2022-07-21	Un atelier associant autorités locales, comités villageois et clubs	Un atelier associant autorités locales, comités villageois et clubs "C
7	Décoration de l'ancien ambassadeur	https://ami.mr/fr/archives/2022-07-21	2022-07-21	Au nom de Sa Majesté l'Empereur du Japon, l'ambassadeur du Ja	Au nom de Sa Majesté l'Empereur du Japon, l'ambassadeur du Ja
8	Le ministre espagnol des Affaires ét	https://ami.mr/fr/archives/2022-07-20	2022-07-20	Le ministre des Affaires étrangères, de l'Union européenne et de l	Le ministre des Affaires étrangères, de l'Union européenne et de l
9	Fin de la visite du ministre espagnol	https://ami.mr/fr/archives/2022-07-20	2022-07-20	SEM. José Manuel Alparis, ministre des Affaires étrangères, de l'U	SEM. José Manuel Alparis, ministre des Affaires étrangères, de l'U
10	Le président du parti Insaf : La strat	https://ami.mr/fr/archives/2022-07-18	2022-07-18	Le président du parti Insaf, M. Mohamed Melaynine Ould Eyih, a	Le président du parti Insaf, M. Mohamed Melaynine Ould Eyih, a ir
11	La Commission de l'Orientation Islan	https://ami.mr/fr/archives/2022-07-17	2022-07-17	La Commission de l'Orientation Islamique, des Ressources Humain	La Commission de l'Orientation Islamique, des Ressources Humain
12	L'ambassadeur de France: la France	https://ami.mr/fr/archives/2022-07-15	2022-07-15	L'ambassadeur de France en Mauritanie, SEM Robert Moulié a org	L'ambassadeur de France en Mauritanie, SEM Robert Moulié a org
13	L'ambassade algérienne organise un	https://ami.mr/fr/archives/2022-07-05	2022-07-05	L'ambassade de la République démocratique et populaire algérien	L'ambassade de la République démocratique et populaire algérien
14	Le député Biram Dah Abédi appelle	https://ami.mr/fr/archives/2022-07-04	2022-07-04	Le député et président de l'Initiative pour la Résurrection du Mou	Le député et président de l'Initiative pour la Résurrection du Mou
15	Le parti "Insaf" entame un nouveau	https://ami.mr/fr/archives/2022-07-04	2022-07-04	Le Parti "Insaf" (anciennement Union pour la République, UPR) a	Le Parti "Insaf" (anciennement Union pour la République, UPR) a e
16	Le parti Union des Forces pour le Pr	https://ami.mr/fr/archives/2022-07-04	2022-07-04	Le Parti de l'Union des Forces pour le Progrès (UFP) a organisé u	Le Parti de l'Union des Forces pour le Progrès (UFP) a organisé u

FIGURE 1.1 – Aperçu du fichier CSV contenant les articles politiques issus de AMI.mr

1.4 Préparation pour le stockage NoSQL

Après la phase de collecte, le fichier CSV a été converti au format **JSON**, plus adapté au stockage dans une base de données NoSQL orientée documents telle que MongoDB.

Chaque article est ainsi représenté sous forme d'un document JSON indépendant, facilitant son indexation, sa distribution et son analyse dans un environnement distribué.

Chapitre 2

Base de données MongoDB : Réplication et Sharding

2.1 Objectifs du stockage distribué

Le volume important des données collectées à partir du site de l'Agence Mauritanienne d'Information (AMI), composé de plus de 9 000 articles politiques, nécessite une solution de stockage capable d'assurer à la fois la scalabilité, la disponibilité et la tolérance aux pannes. MongoDB, en tant que base de données NoSQL orientée documents, offre des mécanismes avancés de réplication et de partitionnement horizontal (*sharding*) répondant à ces exigences.

2.2 Environnement de travail

La mise en place du cluster MongoDB a été réalisée dans l'environnement suivant :

- **Système d'exploitation** : Windows 11
- **Version MongoDB** : 8.2.2
- **Outils utilisés** :
 - MongoDB Server (`mongod`)
 - MongoDB Shell (`mongosh`)
 - MongoDB Compass (interface graphique)
 - MongoDB Database Tools (`mongoimport`)

2.3 Architecture du cluster MongoDB

L'architecture déployée repose sur un cluster MongoDB sharded combinant réplication et partitionnement horizontal. Le point d'entrée unique du cluster est assuré par un routeur `mongos`, chargé de distribuer les requêtes vers les shards appropriés.

Vue d'ensemble de l'architecture

Chaque shard est configuré comme un *Replica Set* composé d'un nœud primaire et de deux nœuds secondaires, assurant ainsi la haute disponibilité des données.

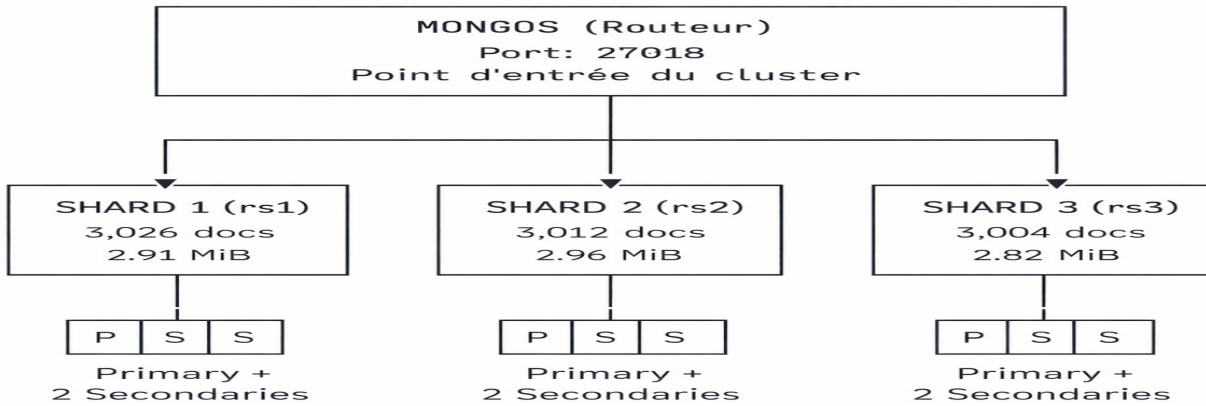


FIGURE 2.1 – Le cluster MongoDB : architecture distribuée avec les composants

2.4 Serveurs de configuration (Config Servers)

Les serveurs de configuration stockent les métadonnées du cluster et la configuration du sharding. Un replica set dédié, nommé `configReplSet`, a été mis en place avec 3 instances : Port 27019 , Port 27020 et Port 27021 ,

2.5 Configuration des shards

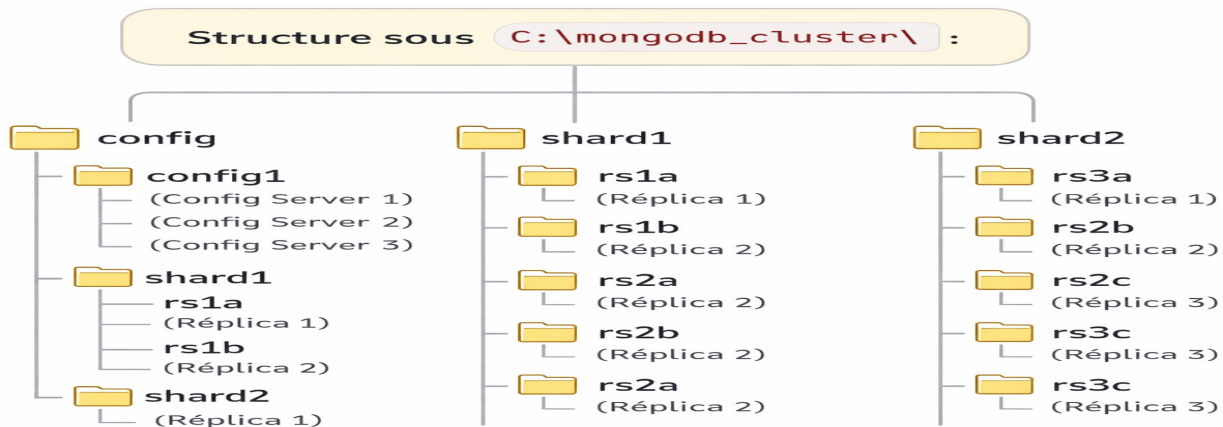


FIGURE 2.2 – Structure et Organisation des répertoires de données

Le cluster comprend trois shards, chacun configuré comme un replica set indépendant composé de trois instances :

Les serveurs de configuration stockent les métadonnées du cluster MongoDB. Trois instances ont été démarrées afin de garantir la haute disponibilité.

1. **Shard 1 (rs1)** : Ports utilisés : 27030, 27031, 27032
2. **Shard 2 (rs2)** : Ports utilisés : 27040, 27041, 27042
3. **Shard 3 (rs3)** : Ports utilisés : 27050, 27051, 27052

2.6 Configuration du routeur Mongos

Le routeur `mongos` constitue le point d'entrée du cluster. Il permet aux applications clientes d'interagir avec le cluster sans avoir à connaître la localisation physique des données. Le routeur est configuré pour se connecter au replica set des serveurs de configuration et écouter sur le port 27018.

```
mongos --configdb configReplSet/localhost:27019,localhost:27020,
      localhost:27021 --port 27018 --bind_ip localhost
```

Commandes de démarrage

```
mongod --configsvr --replSet configReplSet --port 27019 --dbpath "C:\
  mongodb_cluster\config\config1" --bind_ip localhost

mongod --configsvr --replSet configReplSet --port 27020 --dbpath "C:\
  mongodb_cluster\config\config2" --bind_ip localhost

mongod --configsvr --replSet configReplSet --port 27021 --dbpath "C:\
  mongodb_cluster\config\config3" --bind_ip localhost
```

NB : Une fois le routeur démarré, les trois shards ont été ajoutés au cluster.

Initialisation du Replica Set

```
mongosh --port 27019

rs.initiate({
  _id: "configReplSet",
  configsvr: true,
  members: [
    { _id: 0, host: "localhost:27019" },
    { _id: 1, host: "localhost:27020" },
    { _id: 2, host: "localhost:27021" }
  ]
})

rs.status()
```

La commande `rs.status()` permet de vérifier que chaque shard est correctement enregistré et opérationnel.

2.7 Configuration des shards

Shard 1 (rs1)

```
mongod --shardsvr --replSet rs1 --port 27030 --dbpath "C:\
  mongodb_cluster\shard1\rs1a" --bind_ip localhost
mongod --shardsvr --replSet rs1 --port 27031 --dbpath "C:\
  mongodb_cluster\shard1\rs1b" --bind_ip localhost
mongod --shardsvr --replSet rs1 --port 27032 --dbpath "C:\
  mongodb_cluster\shard1\rs1c" --bind_ip localhost
```

Listing 2.1 – Démarrage des instances du Shard 1

```
mongosh --port 27030

rs.initiate({
  _id: "rs1",
  members: [
    { _id: 0, host: "localhost:27030" },
    { _id: 1, host: "localhost:27031" },
    { _id: 2, host: "localhost:27032" }
  ]
})
```

Listing 2.2 – Initialisation du Replica Set rs1

Shard 2 (rs2)

```
mongod --shardsvr --replSet rs2 --port 27040 --dbpath "C:\
  mongodb_cluster\shard2\rs2a" --bind_ip localhost
mongod --shardsvr --replSet rs2 --port 27041 --dbpath "C:\
  mongodb_cluster\shard2\rs2b" --bind_ip localhost
mongod --shardsvr --replSet rs2 --port 27042 --dbpath "C:\
  mongodb_cluster\shard2\rs2c" --bind_ip localhost
```

Listing 2.3 – Démarrage des instances du Shard 2

```
mongosh --port 27040

rs.initiate({
  _id: "rs2",
  members: [
    { _id: 0, host: "localhost:27040" },
    { _id: 1, host: "localhost:27041" },
    { _id: 2, host: "localhost:27042" }
  ]
})
```

Listing 2.4 – Initialisation du Replica Set rs2

Shard 3 (rs3)

```
mongod --shardsvr --replSet rs3 --port 27050 --dbpath "C:\
  mongodb_cluster\shard3\rs3a" --bind_ip localhost
mongod --shardsvr --replSet rs3 --port 27051 --dbpath "C:\
  mongodb_cluster\shard3\rs3b" --bind_ip localhost
mongod --shardsvr --replSet rs3 --port 27052 --dbpath "C:\
  mongodb_cluster\shard3\rs3c" --bind_ip localhost
```

Listing 2.5 – Démarrage des instances du Shard 3

```
mongosh --port 27050

rs.initiate({
  _id: "rs3",
  members: [
    { _id: 0, host: "localhost:27050" },
    { _id: 1, host: "localhost:27051" },
    { _id: 2, host: "localhost:27052" }
  ]
})
```

Listing 2.6 – Initialisation du Replica Set rs3

Ajout des shards au cluster

```
mongosh --port 27018

sh.addShard("rs1/localhost:27030,localhost:27031,localhost:27032")
sh.addShard("rs2/localhost:27040,localhost:27041,localhost:27042")
sh.addShard("rs3/localhost:27050,localhost:27051,localhost:27052")

sh.status()
```

2.8 Activation du sharding sur la base de données

Le sharding a été activé sur la base de données `ami_politique`. Un index de type *hashed* a été créé sur le champ `_id` afin de garantir une distribution uniforme des documents entre les shards. La collection `articles` a ensuite été shardée sur cet index.

```
sh.enableSharding("ami_politique")
use ami_politique
db.articles.createIndex({ "_id": "hashed" })
sh.shardCollection("ami_politique.articles", { "_id": "hashed" })
```

2.9 Stratégie de sharding et répartition des données

Le sharding mis en œuvre dans ce projet repose sur une stratégie de type *Hashed Sharding*, utilisant le champ `_id` comme clé de partitionnement. Après l'importation des données, la répartition observée est équilibrée, avec environ **3 000 articles par shard** et une taille comprise entre **2.8 et 3.0 MB**. Cette homogénéité confirme l'efficacité de la stratégie de sharding adoptée.

Trois chunks ont été créés, MongoDB a automatiquement découpé les données en ces trois *chunks*, chacun associé à un shard du cluster. Chaque chunk couvre un intervalle distinct de valeurs hachées du champ `_id`, garantissant une répartition uniforme des documents.

```
chunks: [
  { min: { _id: MinKey() }, max: { _id: Long("-3074457345618258602") }, "on shard": "rs3" },
  { min: { _id: Long("-3074457345618258602") }, max: { _id: Long("3074457345618258602") }, "on shard": "rs1" },
  { min: { _id: Long("3074457345618258602") }, max: { _id: MaxKey() }, "on shard": "rs2" }
]
```

Listing 2.7 – Répartition des chunks entre les shards

Cette approche permet une distribution uniforme des données, évite les *hotspots* et s'adapte efficacement aux insertions massives de documents, même avec des identifiants séquentiels.

2.10 Importation et répartition des données

Les données ont été importées dans le cluster MongoDB à l'aide de l'outil `mongoimport`, en se connectant au routeur `mongos` afin de permettre une distribution automatique des documents entre les shards.

```
mongoimport --port 27018 \
--db ami_politique \
--collection articles \
--file "C:\Users\Mohame\Lemine\Desktop\ssd_projects\
  articles_mauritanie.json" \
--jsonArray
```

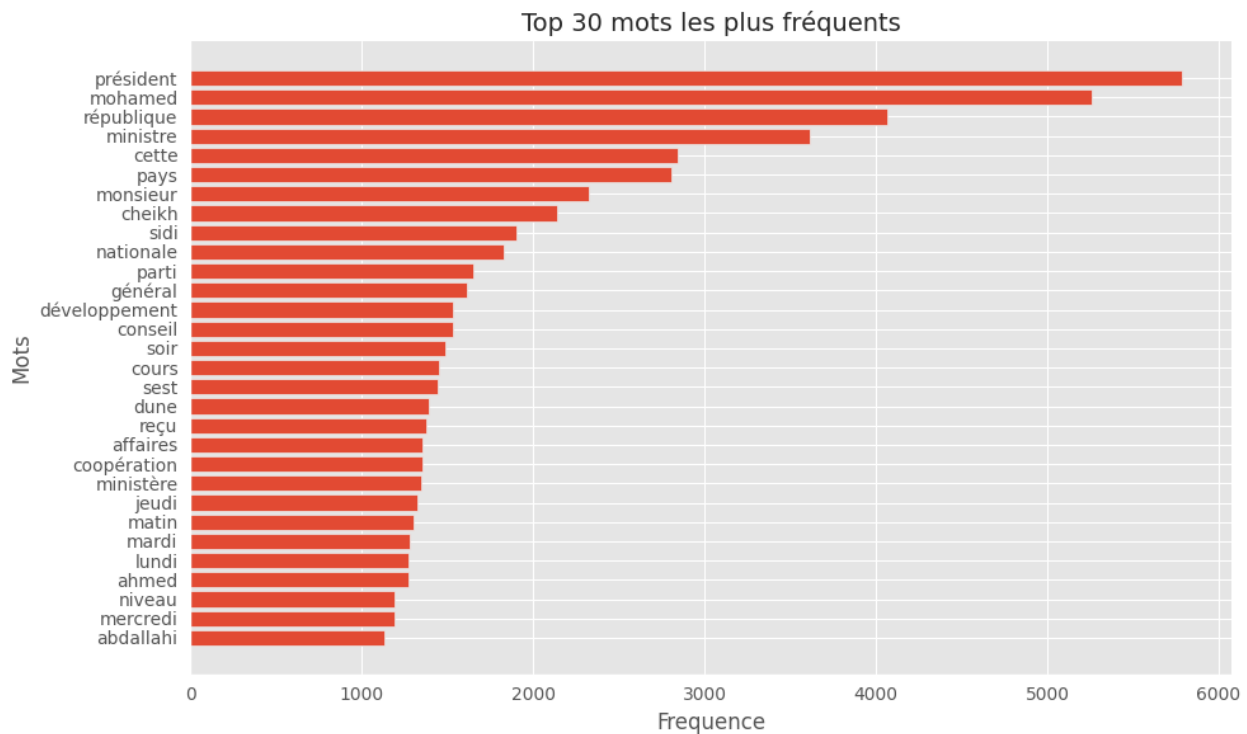
Après l'importation, la commande `db.articles.getShardDistribution()` a été utilisée pour analyser la répartition des données. Les résultats montrent une distribution très équilibrée : chaque shard contient environ un tiers des documents, avec un écart inférieur à 2%. Cette homogénéité confirme l'efficacité du sharding basé sur un hachage du champ `_id`.

Chapitre 3

Visualisations des données

3.0.1 Top 30 des mots les plus fréquents

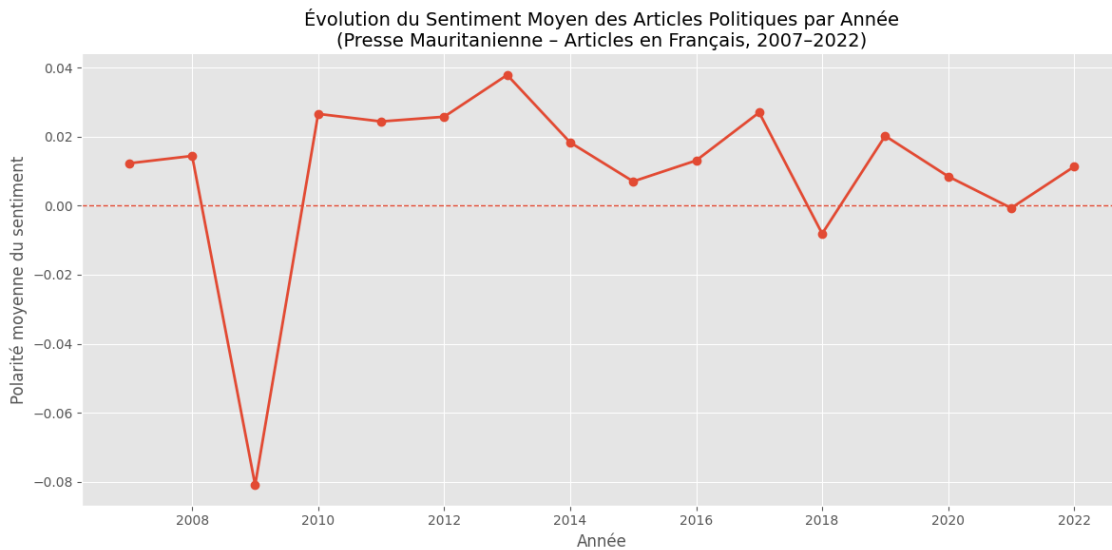
Cette visualisation montre les 30 mots les plus utilisés dans le corpus après nettoyage et suppression des stopwords. Elle permet d'identifier les thèmes dominants du discours politique.



3.0.2 Analyse de sentiment moyen par année

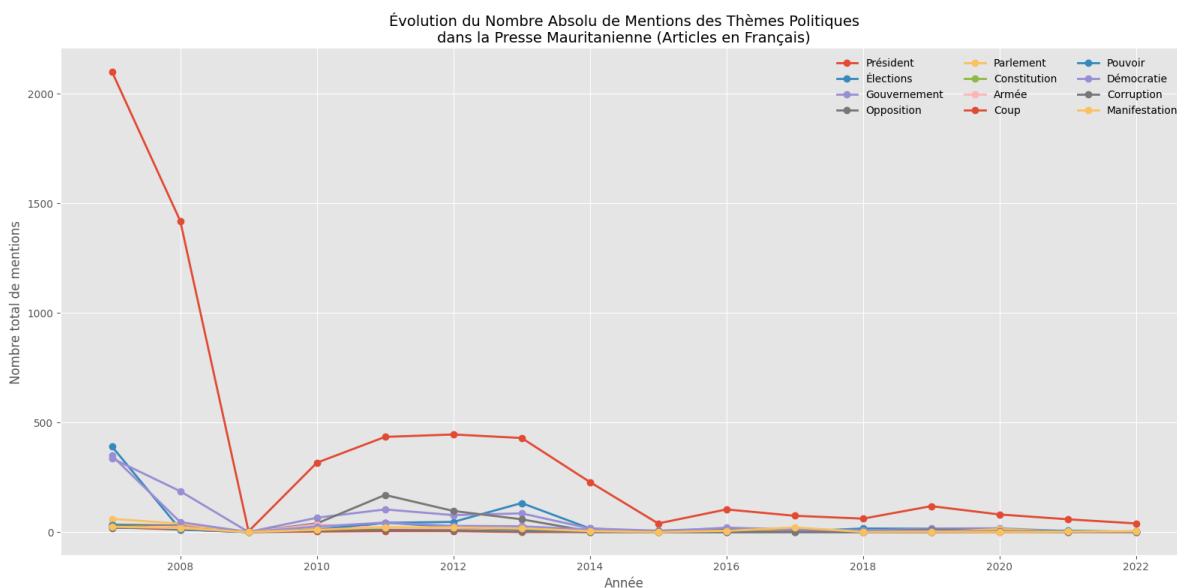
La figure montre que le sentiment moyen est globalement légèrement positif sur la période étudiée. Une chute marquée en 2009 révèle un discours fortement négatif, probablement liée à un contexte politique sensible. Entre 2010 et 2017, le ton devient majoritairement positif

avec des fluctuations modérées. Les années récentes indiquent une stabilisation autour de la neutralité, suggérant un discours plus équilibré.

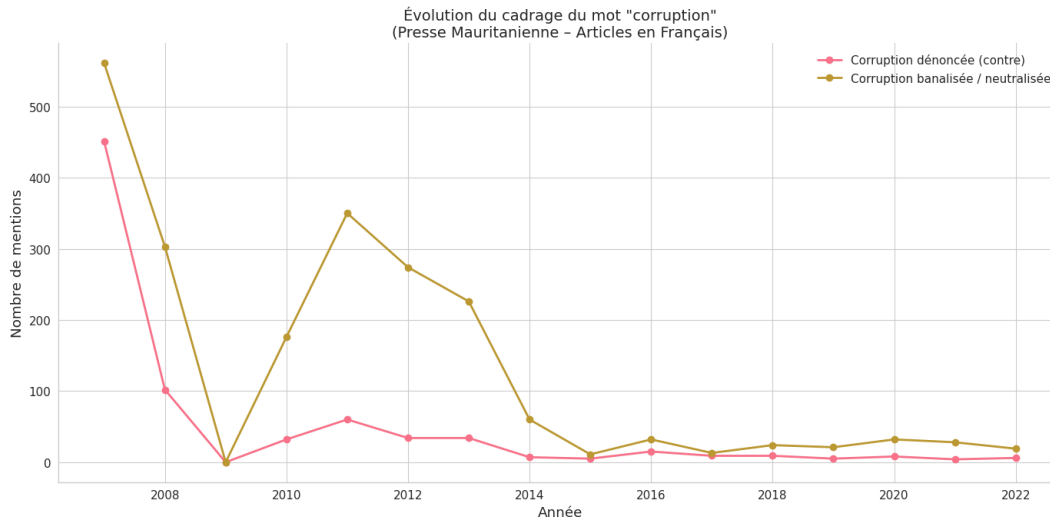


3.0.3 Évolution du nombre absolu de mentions des thèmes politiques

La figure montre une forte domination du thème *président*, particulièrement marquée en 2007 et 2008. Les thèmes liés aux *élections*, au *gouvernement* et à l'*opposition* connaissent des pics ponctuels, reflétant des périodes de forte activité politique. À partir de 2015, le volume global des mentions diminue, suggérant une baisse de l'intensité du débat politique dans la presse.



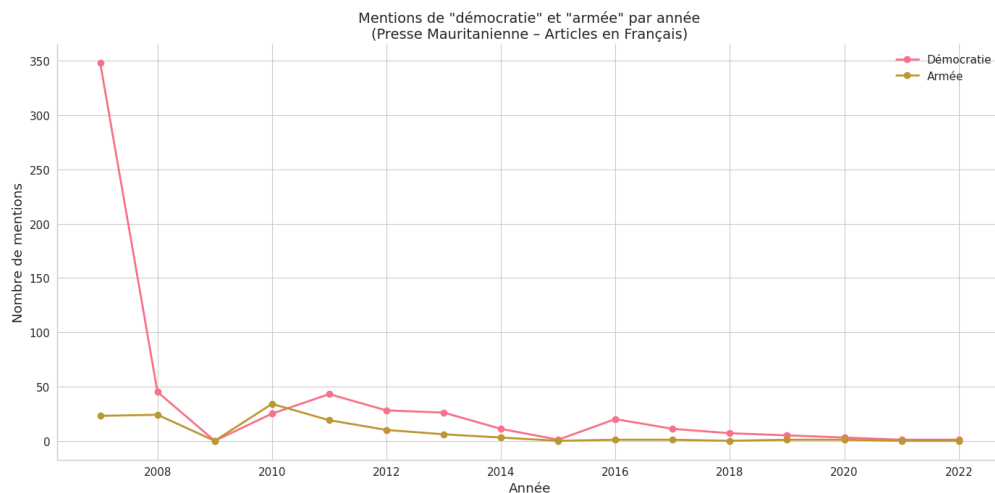
3.0.4 Évolution du cadrage du mot « corruption »



la corruption est majoritairement traitée sous un angle banalisé, notamment au début de la période étudiée. Les mentions dénonçant explicitement la corruption restent présentes, mais à un niveau nettement inférieur. Après 2014, Cette évolution suggère une diminution de la centralité du thème de la corruption dans l'agenda médiatique.

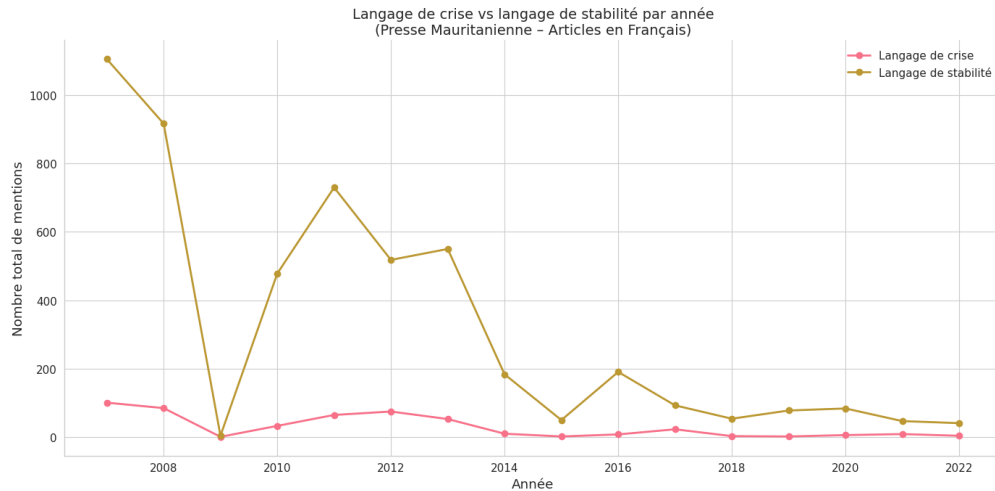
3.0.5 Mentions de « démocratie » et « armée »

Cette figure illustre l'évolution des mentions des termes « démocratie » et « armée » dans les articles entre 2008 et 2022. On observe un pic très marqué en 2008 pour « démocratie » (période post-coup d'État), suivi d'une chute rapide, tandis que les mentions de « armée » restent faibles et stables.



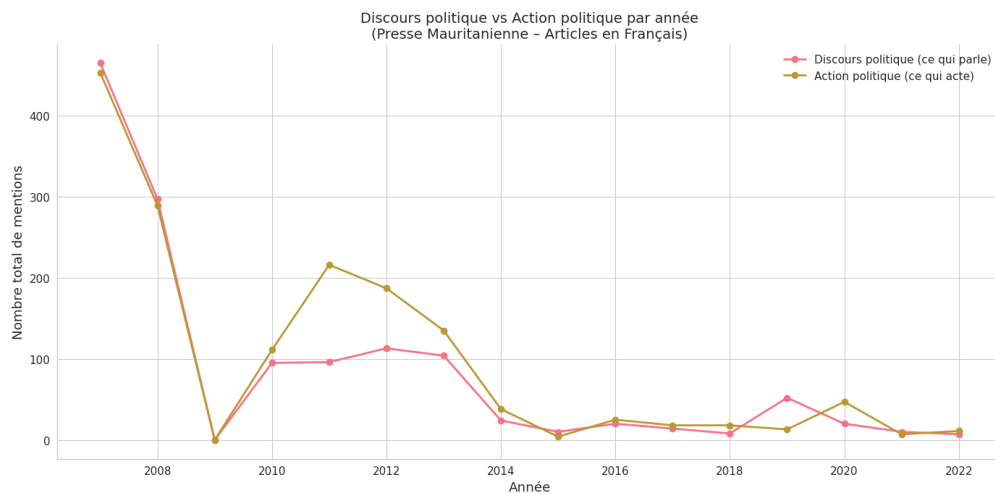
3.0.6 Langage de crise vs langage de stabilité

Cette visualisation agrège les mots liés à la crise (ex. *coup*, *violence*, *chaos*, *instabilité*) et à la stabilité (ex. *paix*, *ordre*, *réforme*, *développement*, *sécurité*) pour chaque année entre 2008 et 2022. Elle révèle une prédominance très forte du langage de stabilité en 2008–2012 (après le coup d’État puis pendant la transition), suivie d’une nette diminution progressive, tandis que le langage de crise reste marginal et relativement stable sur toute la période.



3.0.7 Discours politique vs Action politique

Cette figure distingue les mentions de mots liés au discours (ex. *promesse*, *déclaration*, *discours*) et à l’action (ex. *décision*, *mise en œuvre*, *réforme appliquée*) par année entre 2008 et 2022. Elle révèle un pic initial élevé pour les deux en 2008 (période de transition post-coup d’État), suivi d’un déclin marqué du discours politique, tandis que l’action reste faible et fluctuante, soulignant un écart persistant entre rhétorique et concrétisation effective du pouvoir.



Conclusion

Ce projet a permis de mettre en pratique les compétences acquises en web scraping, en gestion de bases de données distribuées et en analyse statistique appliquée aux données textuelles. L'exploitation d'articles politiques réels issus du site *ami.mr* a offert une meilleure compréhension du discours politique dans la presse mauritanienne. Les résultats obtenus mettent en évidence l'évolution des thématiques abordées, du sentiment global et du cadrage médiatique au fil du temps. Ce travail constitue ainsi une base solide pour des analyses futures plus approfondies, notamment à l'aide de méthodes avancées de traitement automatique du langage naturel.

Remerciements. Nous adressons nos sincères remerciements à notre enseignant pour son encadrement, ses conseils précieux et son accompagnement tout au long de la réalisation de ce projet.