Projet de certificat de spécialisation de données massives

Conservatoire National des Arts et Métiers (CNAM)

2021/2022 - second semestre

Micout, Frédéric

frederic.micout.auditeur@lecnam.net

12 septembre 2022

Table des matières

Table des figures

Abréviations

CNAM Conservatoire National des Arts et Métiers

RSS Really Simple Syndication

SGBDR Système de Gestion de Base de Donnée Relationnelle

1 Présentation

1.1 Contexte

Toutes les rédactions ont depuis longtemps pris le virage du numérique et publient régulièrement de nouveaux articles en lien avec l'actualité. La publication passe par différents canaux, outre la version papier lorsqu'elle existe. Le plus connu et le plus adapté à la diffusion directe dans un navigateur web est la page HTML. Le flux RSS (Really Simple Syndication) permet aussi de diffuser des contenus et offre des possibilités en termes de collecte automatisée.

1.2 Problématiques à résoudre

L'actualité est un objet en constant renouvellement, les nouvelles actualités chassant les plus anciennes. Ce renouvellement permanent se traduit, entre autre, par une certaine forme d'oubli de l'actualité passée. Ainsi, et même en cherchant bien, il est parfois impossible de se souvenir de l'actualité dominante sur une période passée de quelques mois. La première problématique à résoudre consiste à faire remonter l'ensemble des actualités dominantes présentes dans les titres de flux RSS.

L'actualité est un objet qui peut être traité de différentes manières. Chaque rédaction fait des choix éditoriaux (met en avant tel ou tel type d'actualité, traite l'information selon un certain angle, ...). La seconde problématique à résoudre consiste, sinon à cartographier précisément les orientations de chaque flux RSS, au moins à proposer une approche permettant cette analyse.

1.3 Description de la source de données

Les données utilisées dans le cadre de ce projet sont issues de la collecte personnelle de flux RSS de plusieurs sites d'information des diverses tendances d'opinion :

- L'Humanité
- Médiapart
- Le Figaro
- Ouest France (flux des actualités nationales)
- Télégramme de Brest (flux des actualités nationales)

Les données sont stockées dans une base PostgreSQL. Pour chaque article, les données suivantes sont conservées dans une table :

- Le titre de l'article.
- L'identifiant du flux RSS d'origine.
- La date de collecte au format YYYY-MM-DD

Les différents enregistrements sont indépendants les uns des autres. L'identifiant du flux d'origine est une clé étrangère faisant référence aux enregistrements dans une seconde table contenant les URL des flux RSS et une brève description.

```
fluxRSS=# \d+ feed
                                    Table "public.feed"
 Colonne |          Type          |                     Modificateurs                      | Stockage
---------+------------------------+--------------------------------------------------------+----------
 link    | character varying(300) |                                                        | extended
```

```
 id      | integer                | non NULL Par defaut, nextval('feed_id_seq'::regclass) | plain
 themes  | character varying(300) |                                                       | extended


fluxRSS=# d+ sauvegardefluxbrut
                                    Table "public.sauvegardefluxbrut"
 Colonne |  Type   |                          Modificateurs                          | Stockage
---------+---------+-----------------------------------------------------------------+----------
 id      | integer | non NULL Par defaut, nextval('sauvegardefluxbrut_id_seq'::regclass) | plain
 id_feed | integer |                                                                 | plain
 title   | text    |                                                                 | extended
 pubdate | date    |                                                                 | plain
Index :
    "sauvegardefluxbrut_pkey" PRIMARY KEY, btree (id)
```

Note : le contenu des articles a été écarté de ce projet de collecte. En effet, l'accès complet à cette partie du contenu est variable d'une source à l'autre, l'accès complet étant généralement conditionné par la souscription d'un abonnement. Cela pose aussi la question de la légalité des opérations de collecte, de stockage et de réutilisation de cette donnée. Ces questions se posent moins si l'on ne considère que le titre des articles, cette donnée étant par nature destinée à être largement diffusée sans restrictions.

Les données issues de ces flux sont collectées depuis le 1er mai 2017 (le début de collecte d'une partie des flux est antérieure à cette date mais ne sera pas analysée ici afin de permettre une analyse comparative). Cela représente près de 400 000 enregistrements stockés dans une base PostgreSQL. Les données collectées sont essentiellement composées en langue Française.

```
fluxRSS=# SELECT * FROM sauvegardefluxbrut WHERE id_feed IN (24,25,30,34,50) ORDER BY pubdate DESC LIMIT 10;
   id   | id_feed |                                                                                                  title                                                                                                  |  pubdate
--------+---------+---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+------------
 930188 |      25 |  Vols de chiens : les élevages sont aussi concernés                                                                                                                                                     | 2022-06-12
 930192 |      50 |  L'épave du galion San José dévoile un peu plus ses innombrables trésors                                                                                                                                | 2022-06-12
 930186 |      25 |  Le vrai-faux des législatives : les 15 choses à savoir avant d'aller voter                                                                                                                             | 2022-06-12
 930187 |      25 |  Cinq innovations qui vont façonner l'hôtellerie du futur                                                                                                                                               | 2022-06-12
 930189 |      25 |  Les vols de chats et chiens, troisième trafic après les stups et les armes                                                                                                                             | 2022-06-12
 930190 |      25 |  Un convoyeur de fonds jugé pour le vol de trois millions d'euros d'un fourgon blindé                                                                                                                    | 2022-06-12
 930179 |      50 |  24 Heures du Mans : l'accident de l'acteur Michael Fassbender en vidéo                                                                                                                                 | 2022-06-12
 930184 |      50 |  Mali : au moins 5 douaniers et civils tués dans une «attaque terroriste»                                                                                                                               | 2022-06-12
 930194 |      24 |  Législatives. « Jamais les députés n'ont été autant bousculés », selon cette journaliste de LCP                                                                                                        | 2022-06-12
 930193 |      24 |  DIRECT. Législatives 2022 : les bureaux de vote sont ouverts dans toute la France métropolitaine                                                                                                       | 2022-06-12
(10 rows)
```

1.4 Démarche envisagée

Afin de répondre à la problématique, l'analyse proposée initialement repose sur les étapes suivantes :

- Export des données depuis la base PostgreSQL vers un cluster PostgreSQL.
- Pré-traitement des données (mise en forme, gestion codage des caractères, traduction langage SMS, ...)
- Extraction d'entités primaires / Étiquetage grammatical / Extraction d'entités nommées
- Analyse syntaxique
- Lemmatisation et racinisation
- Représentation vectorielle de texte
- Identification des thématiques abordées, liées à des événements d'actualité en particulier ou à des sujets sociétaux par exemple
- Analyse de sentiments sur des thématiques extraites de celles trouvées précédemment et comparatif entre les flux RSS traités.

Le processus de traitement est présenté de manière générale en figure 1.1. Le processus hors ligne rassemble tous les pré-traitements permettant une présentation des données plus optimale. Cette étape regroupe les opérations consommatrices de ressources (temps, puissance de calcul) et dont les résultats n'ont plus à être recalculés ensuite. Le processus en ligne rassemble lui toutes les opérations d'analyse

moins coûteuses unitairement mais à exécuter autant de fois que nécessaire (en fonction de l'objectif de l'analyse en cours). Ce découpage permet donc d'optimiser le temps de traitement tout en conservant une certaine flexibilité pour l'analyse.

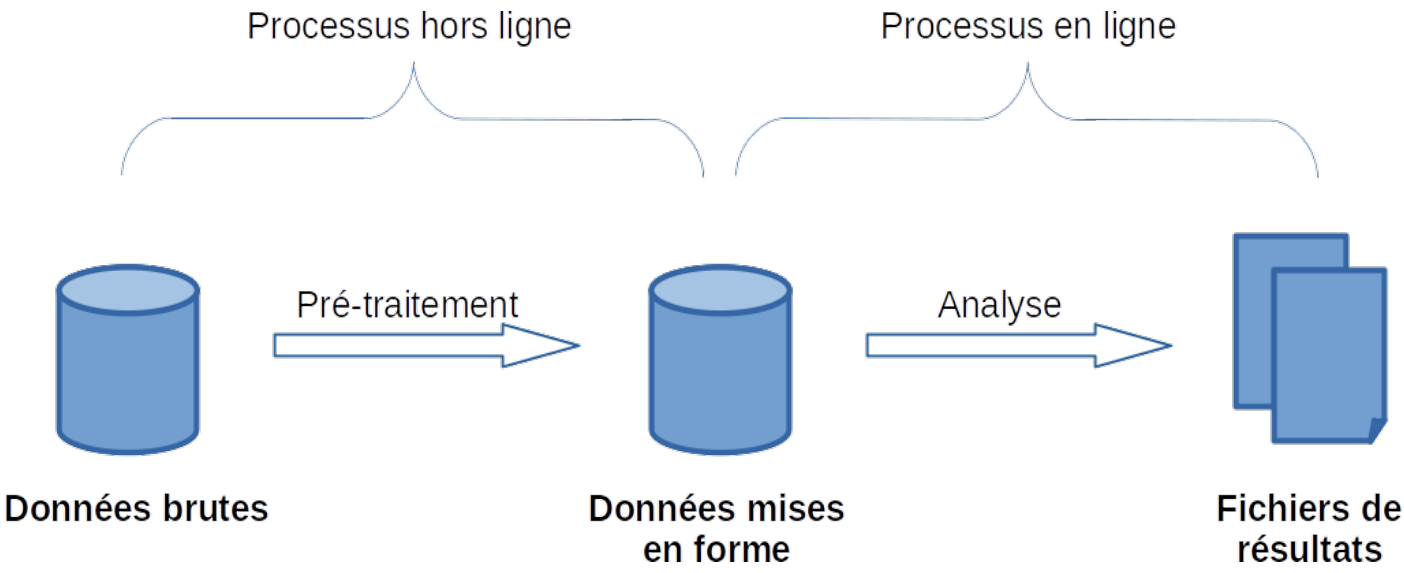

FIGURE 1.1 – Processus général envisagé

1.5 Plateforme matérielle mise en œuvre

Les différents nœuds déployés dans le cadre de ce projet sont exécutés sur des conteneurs LXC dans un hyperviseur de type 1 (Proxmox VE [14] - basé sur Debian 11) hébergé localement et installé sur un ordinateur portable bureautique reconditionné (processeur core i5 3360 ; disque SSD de 240 Go ; 8 Go de mémoire RAM). Par défaut, tous les nœuds s'appuient sur un système Debian 11 [4] (comme le système hôte). Ces machines disposent chacunes de ressources identiques (1 vCPU ; 8 Go d'espace disque ; 512 Mo de RAM) et communiquent entre elles par le biais d'un réseau privé de classe C (192.168.0.0/24). Une vue générale de l'interface de Proxmox est présentée en figure 1.2.

Les traitements réalisés notament dans Spark sont opérés sur un second ordinateur portable bureautique (Xubuntu LTS 22.04 ; core i5 6300U ; disque SSD de 120 Go ; 8 Go de RAM ; carte graphique intégrée Intel HD Graphics 520). La stack logicielle complète installée sur ces équipements et utilisée dans ce projet est présentée en annexe A.

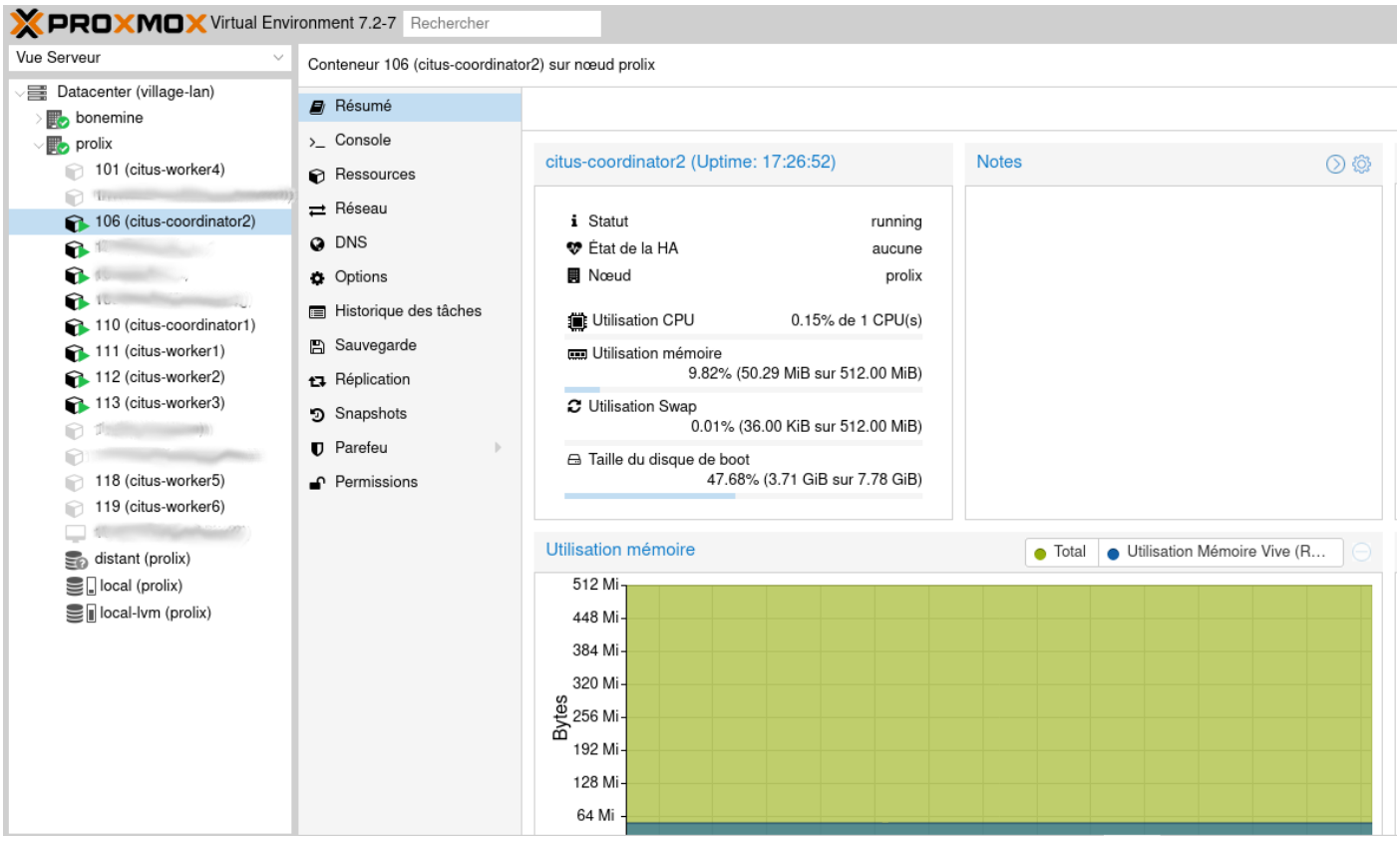

FIGURE 1.2 – Interface de l'hyperviseur Proxmox VE

2 Stockage des données

2.1 Description des données et du besoin

Les données brutes sont organisées dans deux tables initialement stockées sur un seul nœud. Le système cible doit être en mesure de traiter des données qui ne pourraient plus être convenablement stockées et/ou traitées sur un seul nœud. Plusieurs nœuds doivent donc être utilisables. Le nouveau système doit permettre l'ajout ou la suppression de nœuds à tout moment. Il doit enfin être tolérant à la panne et accepter la perte d'un ou plusieurs nœuds sans que cela ne bloque l'accès aux données restantes.

Il est à noter que les données à stocker de manière distribué peuvent théoriquement être les données brutes et/ou les données issues de la phase de pré-traitement. La distribution des données brutes présente relativement peu d'intérêt car elles sont traitées une seule fois et hors ligne. La distribution des données mises en forme est bien plus intéressante, et cela pour trois raisons :

1. La volumétrie sur disque des données une fois mises en forme est bien plus importante que celle des données brutes (avec en particulier l'ajout d'un vecteur de 200 dimensions à chaque enregistrement, chaque dimension étant codée par un nombre flottant occupant 4 octets). La distribution des données permet donc de dépasser les limites d'une seule machine.
2. Il faut envisager qu'un service potentiellement populaire puisse être basé sur l'exploitation de ces données. Le service en question doit pouvoir gérer un nombre important de requêtes en simultané.
3. Toujours dans le cas d'un service basé sur l'exploitation de ces données, la distribution des données permet d'envisager un mode dégradé en cas de perte d'un nœud.

2.2 Choix de la solution de stockage

PostgreSQL [7] est un Système de Gestion de Base de Donnée Relationnelle (SGBDR) qui n'est par nature pas immédiatement adapté pour fonctionner en mode distribué. Malgré cela, cette solution est fortement utilisée y compris dans des contextes où la volumétrie des données est importante. PostgreSQL est un outil sous licence libre et un certain nombre d'éditeurs proposent à leur tour des solutions basées dessus ou étendant ses possibilités. C'est le cas de Citus.

Citus [3] est une solution venant étendre les possibilités de PostgreSQL en termes de passage à l'échelle et cela, sans casser l'organisation relationnelle des données. Cette dernière caractéristique est objectivement peu décisive dans notre cas d'usage mais mérite d'être notée. Nous rechercherons ici à obtenir de la scalabilité horizontale (Sharding). La répartition des données entre les différents nœuds du cluster peut être effectuée en se basant explicitement sur une clé définie dans chaque table. Le point d'entrée est la machine qui coordonne les requêtes. L'accès au service est dépendant du bon fonctionnement de ce nœud coordinateur (c'est un point unique de défaillance ; SPOF (Single Point of Failure)). Du point de vu des outils externes, la solution se présente comme une base de donnée PostgreSQL classique.

2.3 Stratégie de distribution des données

La stratégie de distribution des données est un point important et est fonction des objectifs fixés :

— Afin de servir un maximum de requêtes en parallèle, il est possible de **regrouper les données** qui sont généralement accédées en même temps. Cela permet potentiellement d'éviter l'interrogation de l'ensemble des nœuds du cluster à chaque requête et ainsi d'utiliser les ressources disponibles plus efficacement.
— Afin de permettre un mode dégradé en cas de panne (dans le cas où les données ne seraient pas redondées et où l'on souhaiterait être en mesure de fournir au moins une réponse un peu moins bonne en l'absence d'une partie des nœuds), il a autant que possible tout intérêt à **ventiler les données** entre les différents nœuds du cluster et cela de manière aléatoire.

Ces objectifs ne semblent pas conciliables à première vue. Il est cependant possible de remarquer que procéder à un regroupement optimal des données est une opération délicate qui suppose de connaître à l'avance la nature des requêtes formulées lors de l'analyse. Cette information n'est pas connue à priori. Un bon regroupement pour un utilisateur, par exemple basé sur le mois de publication des titres, ne l'est pas pour un autre utilisateur qui sera plutôt amené à rechercher des informations uniquement sur un flux dans toute la période couverte par le jeu de données. Le regroupement des actualités par date de publication semble être une option qui couvre des besoins variés. Dans la mesure où l'id des titres est attribué progressivement, il est donc aussi simple de distribuer les données selon cette colonne.

La **répliquation des données** entre les nœuds pourrait être envisagée. Par ce moyen, il serait à la fois possible de répartir la charge entre différents nœuds contenant les mêmes données et de palier avec plus d'efficacité à la perte accidentelle de l'un des nœuds. Si l'on s'arrête à la documentation de la version de Citus employée (v11), aucun mécanisme permettant la réplication des données entres les membres du cluster n'est proposée nativement. La documentation renvoit aux mécanismes de réplication natifs de PostgreSQL. En particulier, la solution proposée consiste à utiliser la réplication continue ("Streaming Replication") vers un nœuds worker en attente. Le WAL du nœud actif (journal contenant toutes les transactions avant écriture sur disque) est constament envoyé vers le nœud inactif associé. Dans l'absolu, c'est cette solution officielle et pérenne qui devrait être implémentée en production.

Cela étant dit, la documentation de la version 10 de Citus fait référence à un mécanisme de réplication intégré [5]. Les motivations liées au retrait de cette fonctionnalité de la documentation ne sont pas abordées. Du moins, il n'a pas été possible initialement de les retrouver, alors même que cet aspect touche le cœur de Citus en tant que solution permettant la distribution de données Postgres (des limitations seront finalement vues à l'usage). Ainsi, la version 11 de Citus accepte le paramètre de réplication des shards sur les nœuds workers du cluster tel que décrit dans la version 10 de la documentation. Sur le nœud coordinateur, indiquer le paramètre suivant (puis redémarrer postgresql) avant de procéder à la distribution des tables sur les nœuds workers :

```
SET citus.shard_replication_factor = 2;
```

À noter que ce paramètre peut aussi être fixé dans le fichier de configuration de Postgresql directement (nécessite un redémarrage de postgresql).

```
/etc/postgresql/14/main/postgresql.conf
```

La table contenant les informations générales des flux est conservée sur le nœud coordinateur. Elle aurait éventuellement pu être copiée sur chaque worker (Small Cross Tenant Table) mais il n'y a pas de réels gains à procéder ainsi. Cette table n'a pour ainsi dire pas à être utilisée dans des opérations de jointures avec la table des titres lors de la phase d'analyse. Si cela avait été le cas, dupliquer cette table sur les workers aurait permi de réaliser les opérations de jointure directement sur les workers et ainsi d'éviter de nombreux échanges de données entre tous les workers et le nœud coordinateur.

2.4 Implémentation de solution de stockage retenue

2.4.1 Préparation des nœuds

PostgreSQL 14 et Citus 11 sont installés sur le nœud coordinateur, suivant les instructions du site officiel. Un certain nombre de nœuds workers sont préalablement provisionnées (ici 3 nœuds, respectivement nommés *citus-worker1*, *citus-worker2* et *citus-worker3*, en écoute sur le port tcp 5432). La préparation de ces nœuds est aussi réalisée en suivant les instructions officielles et avec les mêmes versions de PostgreSQL et de Citus. Les opérations décrites par la suite sont réalisées dans la base de donnée par défaut "postgres". Si une autre base de donnée est employée, elle doit être créée sur l'ensemble des nœuds.

2.4.2 Configuration du coordinateur

Le coordinateur traite les demandes en externes et gère l'état du cluster (cluster au sens Physique et non au sens Postgres). Il assure la répartition des données sur les différents nœuds (ainsi que le sharding). Comme évoqué précédemment, ce nœud est un SPOF. S'il est acceptable de perdre accidentellement un worker (dans la mesure où un mécanisme de réplication des données est mis en œuvre), cela n'est pas le cas pour le coordinateur. Un mécanisme de "Streaming replication" est donc mis en place et permet de synchroniser un second serveur coordinateur avec le premier. D'une part, cela permet de conserver un accès aux données si l'un des serveurs n'est plus accessible et d'autre part, cela permet d'envisager de la haute disponibilité.

À noter que seul le premier nœud coordinateur (*citus-coordinator1*) dispose d'un accès en lecture/écriture sur la base. Le second nœud coordinateur (*citus-coordinator2*) ne dispose lui que d'un accès en lecture. Ce choix est réalisé par simplicité de mise en œuvre et parce que le cas d'usage ne nécessite pas de conserver impérativement un accès en écriture à chaque instant sur la base. Si le serveur autorisé en écriture vient à ne plus être accessible, le service est donc en mode dégradé mais toujours actif. Si l'état de l'accès en écriture avait été plus critique, il aurait été nécessaire de mettre en place un mécanisme de failover permettant de promouvoir un nœud standby en master. La configuration à appliquer est la suivante :

Sur le nœud master, dans le fichier */etc/postgresql/14/main/postgresql.conf*, ajouter :

```
wal_level = replica
wal_log_hints = on
max_wal_senders = 8
max_wal_size = 100MB
hot_standby = on
```

Sur le nœud master, créer le rôle postgres utilisé pour la réplication :

```
postgres=# CREATE USER replication REPLICATION LOGIN CONNECTION LIMIT 1 ENCRYPTED PASSWORD 'replication';
```

Sur le nœud master, dans le fichier */etc/postgresql/14/main/pg_hba.conf*, ajouter ce qui suit afin de permettre les opérations de réplication pour l'utilisateur et depuis une machine présente sur le réseau local :

```
host    replication     replication      192.168.0.1/24 md5
```

Le service PostgreSQL peut ensuite être démarré sur le nœud master. Sur chaque nœud de réplication (ici uniquement un nœud : *citus-coordinator2*), le fichier */etc/postgresql/14/main/postgresql.conf* doit lui aussi être modifié (ajout des instructions suivantes permettant notamment des spécifier les paramètres de connexion TCP avec le nœud master) :

```
listen_addresses = '*'
wal_level = replica
max_wal_senders = 10
hot_standby = on
primary_conninfo = 'host=192.168.0.61 port=5432 user=replication password=replication'
```

2.4.3 Export et import des données brutes

L'export des données de la base d'origne PostgreSQL se fait en deux temps. Tout d'abord, le schéma des données (en plain text) puis les données elles-mêmes (format custom (binaire)). Ici, la base est exportée dans son intégralité.

```
$ pg_dump -d fluxRSS --format=plain --no-owner --schema-only --file=dump_fluxrss_schema.sql
$ pg_dump -d fluxRSS --format=custom --no-owner --data-only --file=dump_fluxrss_data.dump
```

De la même manière, lors de la phase d'import (opération réalisée au niveau du nœud coordinateur1 disposant d'un accès en écriture), le shéma des données est importé avant les données.

```
postgres=# \i /home/fred/dump_fluxrss_schema.sql
postgres=# \i /home/fred/dump_fluxrss_data.sql
```

2.4.4 Application de la stratégie de distribution des données

Les données brutes ne sont pas distribuées et sont uniquement présentes sur le nœud coordinateur. Les scripts développés lors de la phase de pré-traitement interrogent le nœud coordinateur de Citus en l'état. Comme indiqué par la suite, le nombre de requêtes lors de cette phase est très modéré (environ 5 par seconde en moyenne en lecture et autant en écriture pour sauvegarder les résultats dans la table *donnees_pour_analyse*. Les données issues de la phase de pré-traitement sont employées lors de la phase d'analyse et c'est cette table qui fait l'objet d'une distribution sur les différents workers citus.

Dans l'hypothèse où l'on souhaite distribuer les données sur 3 workers et les découper en 32 shards (valeur par défaut), les instructions suivantes sont à donner :

```
postgres=# SELECT * FROM citus_add_node('citus-worker1', 5432);
postgres=# SELECT * FROM citus_add_node('citus-worker2', 5432);
postgres=# SELECT * FROM citus_add_node('citus-worker3', 5432);
postgres=# SELECT create_distributed_table('donnees_pour_analyse','id',shard_count => 32);
```

Comme il est possible de le voir sur la trace suivante, les shards sont effectivement présents sur 2 nœuds workers :

```
postgres=# SELECT * FROM citus_shards LIMIT 10;
      table_name      | shardid |         shard_name          | citus_table_type | colocation_id |   nodename    | nodeport | shard_size
----------------------+---------+-----------------------------+------------------+---------------+---------------+----------+------------
 donnees_pour_analyse |  102547 | donnees_pour_analyse_102547 | distributed      |            13 | citus-worker2 |     5432 |   31924224
 donnees_pour_analyse |  102547 | donnees_pour_analyse_102547 | distributed      |            13 | citus-worker1 |     5432 |   31924224
 donnees_pour_analyse |  102548 | donnees_pour_analyse_102548 | distributed      |            13 | citus-worker2 |     5432 |   31735808
 donnees_pour_analyse |  102548 | donnees_pour_analyse_102548 | distributed      |            13 | citus-worker3 |     5432 |   31735808
 donnees_pour_analyse |  102549 | donnees_pour_analyse_102549 | distributed      |            13 | citus-worker1 |     5432 |   31637504
 donnees_pour_analyse |  102549 | donnees_pour_analyse_102549 | distributed      |            13 | citus-worker3 |     5432 |   31637504
 donnees_pour_analyse |  102550 | donnees_pour_analyse_102550 | distributed      |            13 | citus-worker2 |     5432 |   31694848
 donnees_pour_analyse |  102550 | donnees_pour_analyse_102550 | distributed      |            13 | citus-worker1 |     5432 |   31694848
 donnees_pour_analyse |  102551 | donnees_pour_analyse_102551 | distributed      |            13 | citus-worker3 |     5432 |   31580160
 donnees_pour_analyse |  102551 | donnees_pour_analyse_102551 | distributed      |            13 | citus-worker2 |     5432 |   31580160
(10 rows)
```

À ce stade, la configuration du nœud standby est à réaliser. Dans la mesure où ce nœud ne doit contenir que les données du master, le dossier contenant d'éventuelles données est vidé préventivement et une copie des données est réalisée immédiatement afin d'initialiser le nœud. Afin que le nœud soit identifié comme un standby, un fichier *standby.signal* doit être présent à la racine du répertoire où se trouvent les données. Une fois ces opérations réalisées, le service Postgres peut être démarré sur le nœud de standby. L'architecture complète mise en œuvre dans le cadre de ce projet est présentée en figure 2.1.

```
$ rm -rf /var/lib/postgresql/14/main/*
$ pg_basebackup -h 192.168.0.61 -D /var/lib/postgresql/14/main/ -X stream -c fast -U replication -W
$ touch /var/lib/postgresql/14/main/standby.signal
```

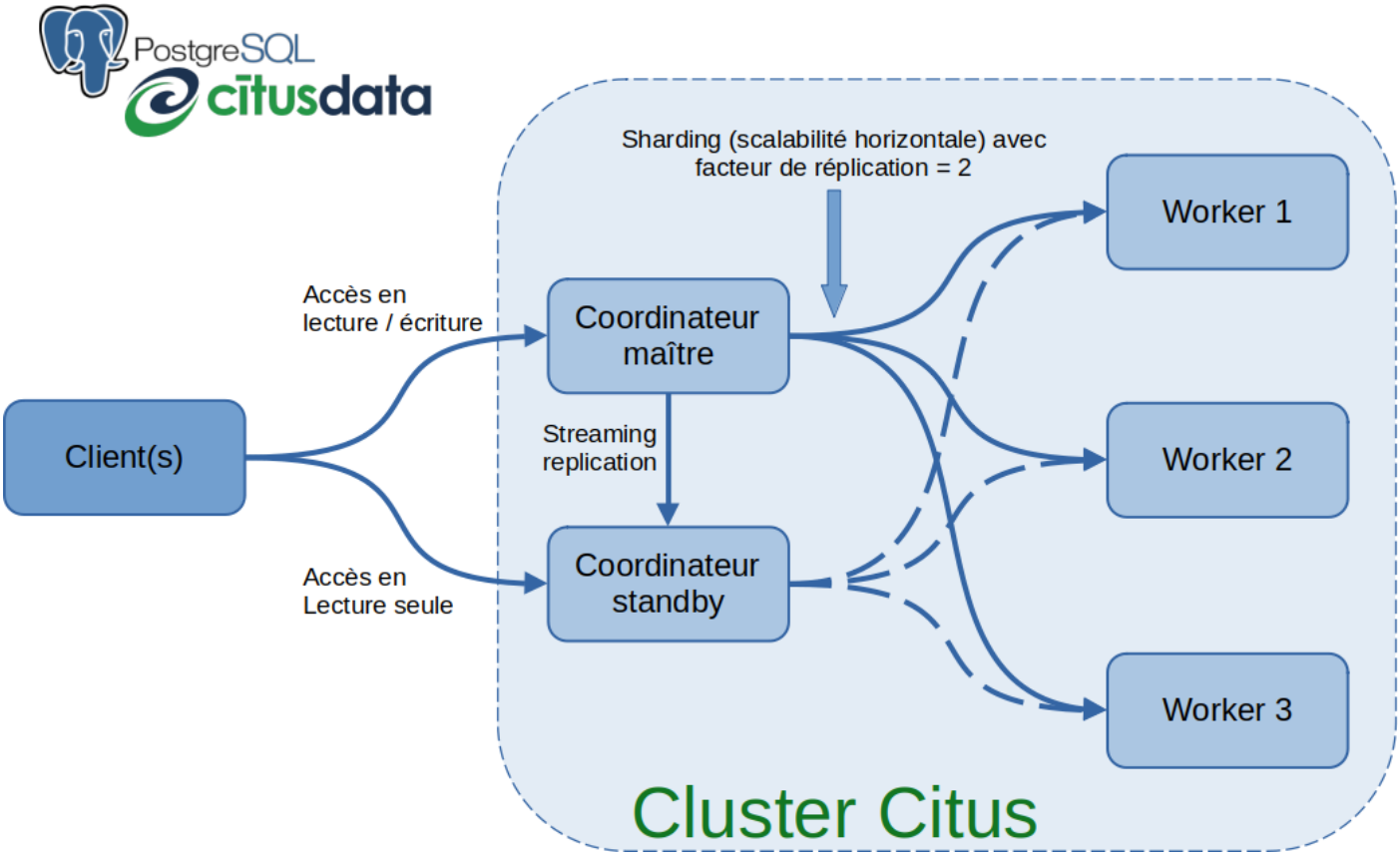

FIGURE 2.1 – Architecture du cluster Citus

3 Algorithmique distribuée

Spark [1] est la solution retenue pour le calcul distribué. Afin de simplifer la réalisation, le framework est exécuté localement via pyspark sur un PC portable fonctionnant sur Xubuntu LTS 22.04. Les divers scripts mis en œuvres sont rédigés au travers d'une instance Jupyter-lab installée localement. Cela permet de travailler directement via des notebooks Jupyter [13] facilement exportables.

La connexion à la base de données PostgreSQL distante peut être réalisée directement en python, de manière classique. Le nœud coordinateur est alors interrogé et masque la distribution des données. Toutefois de cette manière, Spark ne gère pas complètement les requêtes effectuées sur la base de donnée. Toutes les requêtes doivent nécessairement être traitées par le nœud coordinateur (master ou standby), y compris celles nécessitant un accès en écriture (la création de **vues** est notamment concernée, voir ci-dessous). De plus, Spark n'est alors pas en mesure d'optimiser les diverses requêtes afin d'utiliser au mieux les ressources physiques dont il dispose. Il est donc judicieux de profiter du connecteur Postgresql de Spark à cet effet. En travaillant directement sur des Dataframes dans Spark, il est ainsi possible de travailler avec une base de donnée en lecture seule et de profiter pleinement des performances de Spark.

```
postgres=# CREATE VIEW toto AS (SELECT COUNT(*) FROM donnees_pour_analyse);
ERROR:  cannot execute CREATE VIEW in a read-only transaction
```

Spark-nlp [10] est l'outil mis en œuvre dans le cadre de ce projet afin de procéder aux traitement en langage naturel. Cette bibliothèque permet plusieurs choses dont l'extractions de mots clés ou bien encore l'analyse de sentiments. Il est à noter que Spark-nlp suppose de charger de modèles pré-entraînés pouvant dépasser le Gigaoctet. Pour traiter de tels fichiers, outre l'espace de stockage sur disque, Spark doit disposer d'une quantité de mémoire vive conséquente (au moins 16 Go dans le cas présent). La machine physique employée ici ne disposant que de 8 Go de RAM, il a donc été nécessaire d'allouer une partition de swap de 12 Go pour que les traitements puissent aboutir sur de tels fichiers. Dans la mesure où cette partition d'échange est prise sur un disque SSD, les performances restent relativement acceptables.

4 Pré-traitement des données

Notebook Jupyter correspondant à ce chapitre : Pre-traitementDesDonnees.ipynb . Il contient l'ensemble des scripts de pré-traitement des données brutes.

4.1 Objectif

L'objectif de cette phase est de produire, par le biais de traitements hors ligne, les données qui seront proposées en entrée de la phase d'analyse. En effet, les données brutes ne permettent pas d'accéder directement à diverses informations nécessaires à l'analyse (mots clés, sentiment) et ne permettent pas certains traitements (comparaison objective des titres, classification). Aussi, l'étape de pré-traitement doit offrir une vue plus facilement manipulable des données et doit permettre de pré-calculer un certain nombre d'informations, ceci afin de gagner du temps ensuite.

4.2 Processus mis en place lors de la phase de pré-traitement des données

L'ensemble des opérations réalisées et leur articulation est présenté dans la figure 4.1.

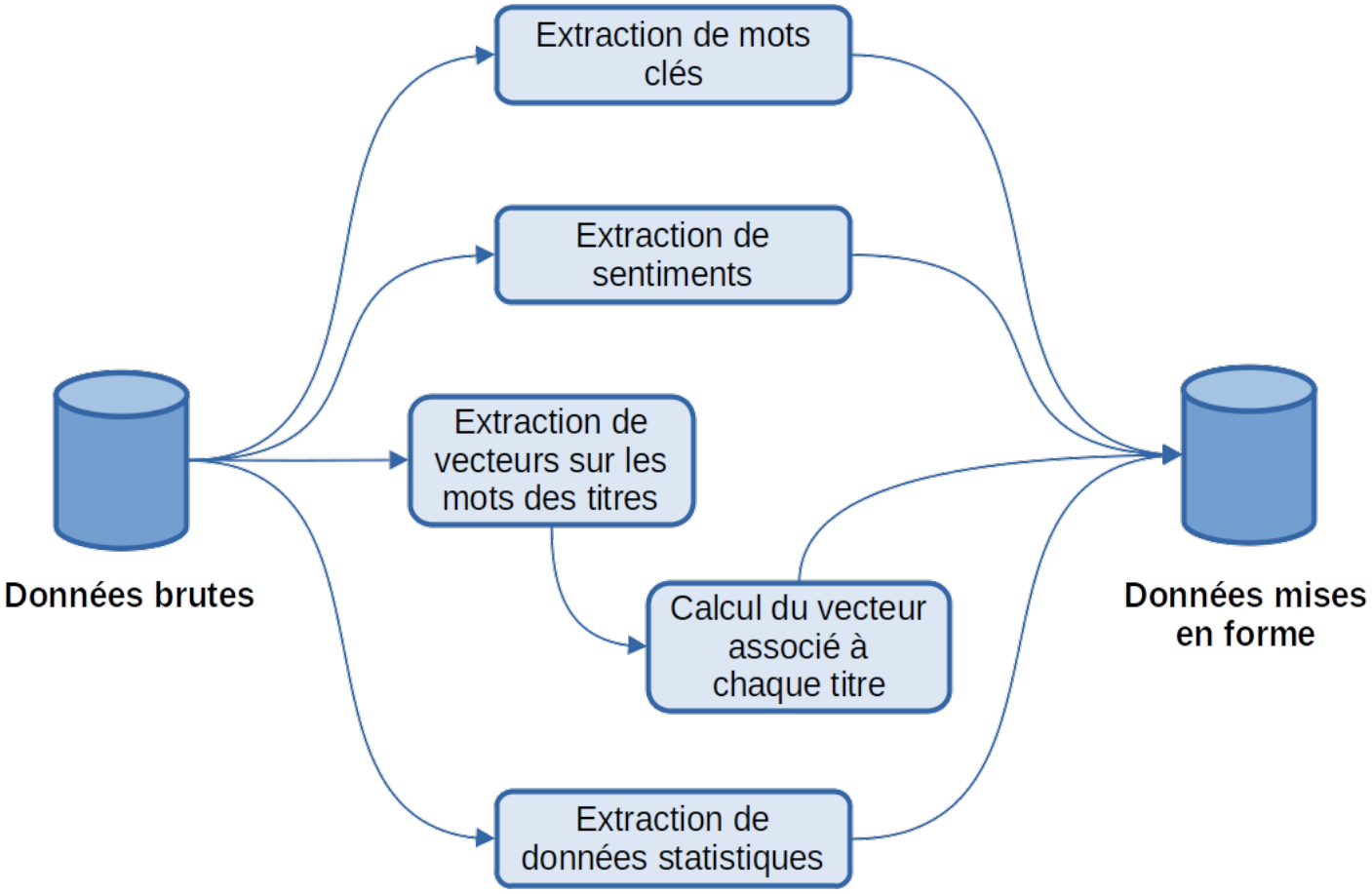

FIGURE 4.1 – Processus de pré-traitement

4.3 Adaptation de la structure de la base de donnée

L'ajout de nouvelles tables semble s'imposer afin de conserver une trace des traitements intermédiaires. Les traitements en langage naturel sont relativement chronophages. Dans les conditions matérielles de ce travail, l'extraction des mots clés l'analyse de sentiments dure environ 250 ms par titre, soit environ 4 jours pour l'ensemble du corpus considéré. La charge en termes de nombre de requêtes SQL d'insertion est constante et suit le même rythme (4 requêtes par seconde, ce qui n'est pas critique ici). Le principe mis en œuvre ici est de rapprocher le calcul des données à traiter. Les données brutes n'étant pas amenées à être modifiées, il semble donc pertinent d'effectuer ces traitements une seule fois et de stocker les résultats en base de données, l'interrogation d'une table étant bien plus rapide.

Le choix est fait ici de créer diverses tables dans la base de donnée. L'ensemble des données stockées dans ces diverses tables seront ensuite rassemblées dans une nouvelle table amenée à être consultée lors de la phase d'analyse. Cette table sera alors distribuée sur le cluster Citus afin de profiter des avantages de la scalabilité horizontale.

4.4 Recherche de mots clés

La gestion du codage ainsi que la traduction du langage SMS par exemple sont des opérations à gérer en premier lieu. Dans le cas présent, les articles sont propres et ne nécessitent pas de traitements bas niveau particuliers.

Spark-nlp procède à des traitements intermédiaires. L'outil permet l'extraction d'entités (primaires ou nominales), l'étiquetage grammatical, l'analyse syntaxique ainsi que la lemmatisation et la racinisation des mots. La collecte et la sauvegarde de ces divers éléments apporte peu à l'analyse et seuls des mots clés sont conservés.

L'approche employée ici consiste à s'appuyer sur des modèles pré-entraînés proposés au travers de Spark-nlp. Ici, comme l'ensemble des données sont en langue Française, un modèle en Francais est donc retenu. Sur l'ensemble du corpus traité, l'outil retourne en moyenne entre un et deux mots clés. A noter que les mots clés dont il est question ici peuvent être des expressions composées (ce point peut avoir une certaine importance pour la suite).

Dans les conditions de réalisation de ce projet, le traitement de l'ensemble des données sur cette partie aura duré plus de quatre jours. Les 4 cœurs de la machine sont mobilisé à 100% durant ce traitement, ce qui tend à démontrer que Spark utilise pleinement les ressources dont il dispose.

Le contrôle de la qualité de l'extraction des mots clé est réalisée par prélèvement. Il en ressort que les mots présentés sont généralement pertinents vis à vis de l'actualité dont sont issus ces mots. Un processus d'échantillonnage plus poussé ne semble pas nécessaire ici. Toutefois, plusieurs problèmes sont à prendre en compte. Pour certains titres, aucune expression clé n'est remonté. Selon les titres analysés, des expressions à priori identiques peuvent différer d'un caractère (par exemple, un guillemet accolé au mot). De plus, une recherche de correspondance exacte entre des mots clés de titres différents ne permet donc pas nécessairement de faire le lien entre mots manifestement identiques. Enfin, certains mots à priori banals sont écartés alors qu'ils ont un sens et une importance particulière lorsqu'il se trouvent en présence d'autres mots. L'expression "Gilet jaune" par exemple ne ressort jamais directement car le mot jaune est écarté systématiquement. Cela peut très certainement s'expliquer par le fait que le mouvement des gilets jaunes en France s'est déroulé après que le modèle sur lequel repose l'analyse ait été construit. Ainsi, une analyse reposant uniquement sur l'extraction de mots clés unitaires risque de produire des résultats incomplets. Il convient donc de coupler cette approche avec d'autres.

4.5 Analyse de sentiments

L'analyse de sentiments mise en œuvre dans ce cadre repose également sur un modèle pré-entraîné. Les titres sont présentés à ce modèle et en retour, Spark-nlp retourne une valeur binaire (sentiment positif ou négatif). Cette analyse donne des résultats globalement attendus. Toutefois, certains titres peuvent être affublés d'une mauvaise étiquette. Cela peut s'expliquer de diverses manières. Pour certains titres, le sentiment devrait plutôt être neutre (par exemple, des bulletins météo annonçant des prévisions conformes aux normales de saison). Pour d'autres titres, le sentiment est aussi fonction de l'observateur. La classification proposée n'est donc ni bonne ni mauvaise. Certains mots peuvent faire basculer des titres majoritairement vers un sentiment ou vers un autre (par exemple, racisme (sentiment majoritairement négatif) ou à l'inverse joie (sentiment majoritairement positif)). Enfin, l'outil peut se tromper dans un sens ou dans l'autre sans raisons apparentes.

Dans la mesure où le sentiment est une information prenant simplement deux modalités et dans la mesure où la phase de comparaison des fluxRSS dépend partiellement de ce paramètre, il est important de savoir jusqu'à quel point il est possible de lui faire confiance. Il n'est pas envisageable de contrôler l'ensemble des labels de sentiment sur tous les titres, en particulier parce que cette approche ne passe pas à l'échelle. Il est donc préférable de procéder à un échantillonnage.

Le processus d'échantillonnage nécessite de préciser différents paramètres. Par simplicité, l'échantillon sera pris aléatoirement. Le nombre d'éléments de l'échantillon permet d'obtenir une certaine représentativité de l'ensemble du jeu de données. Comme le nombre de titres est important (> 400000), la taille de l'échantillon est indépendante de ce paramètre. Si l'on souhaite une marge d'erreur à 5% et un niveau de confiance des résultats à 95% (selon la loi normale centrée réduite), le nombre d'élements à échantillonner est obtenu de la manière suivante :

$$n = \frac{niveau\_confiance^2 * p * (1-p)}{marge\_erreur^2}$$

L'analyse de sentiment réalisée par Spark-NLP identifie environ 60% de titre comme relevant d'un sentiment négatif. la variable p vaut donc 0,6. Cela donne donc :

$$n = \frac{1,96^2 * 0,6 * (1-0,6)}{0,05^2} = 368,79 \approx 370$$

Les données issues de cet échantillonnage aléatoire ont été étiquetées manuellement afin de poser un sentiment "positif" ou "négatif" sur chaque titre. La pose d'un sentiment est, comme indiqué précédemment, un processus relativement subjectif dans certains cas. Ce processus manuel reste réalisable relativement rapidement (1h de temps pour labelliser 370 titres). Il en résulte que 78% des titres ont été adossés au même sentiment entre la phase automatique par Spark-NLP et celle réalisée manuellement. Il y a donc une probabilité de 0,95 que le taux de bonnes prédictions par Spark-NLP soit dans notre cas compris entre 73% et 83%. Ces chiffres sont meilleurs qu'en appliquant une labellisation aléatoire du sentiment sur les titres (où l'on serait à 50% de taux de bonnes prédictions pour p = 0,5). Malheureusement, l'erreur reste importante, ce qui limite les possibilitées liées à l'usage de ce paramètre par la suite.

	Négatif (spark-nlp)	Positif (spark-nlp)
Négatif (manuellement)	VN = 180 / 49%	FP = 58 / 16%
Positif (manuellement)	FN = 25 / 7%	VP = 107 / 29%

À noter que dans l'échantillon traité, 58 titres ont été classifié comme négatif manuellement et positifs par Spark-NLP (ce sont les faux positifs). Ils sont 25 à avoir été jugés positifs manuellement et négatifs par Spark-NLP (ce sont les faux négatifs). Dans ce cas d'usage, Spark-NLP semble donc avoir tendance à se tromper plutôt dans un sens que dans l'autre.

4.6 Représentation vectorielle de texte

À ce stade, chaque titre d'article dispose d'un ensemble de mots clés. Afin de pouvoir exploiter ces données, et en particulier de les comparer, il est nécessaire de les représenter sous forme de vecteurs. Cela permet ensuite d'introduire une notion de distance essentielle ici. La distance cosinus est utilisée afin de quantifier cette distance.

La représentation des titres sous forme de vecteurs peut s'opérer de diverses manières. La manière naïve de procéder est de construire un vecteur où chaque mot est représenté par l'une des composantes du vecteur. Cette approche est totalement inadaptée dans le cas présent dans la mesure où elle amène à produire des vecteurs de grande dimension (de la taille du nombre de mots usuels de la langue, soit plusieurs milliers de mots dans le cas du Français) et que la matrices très creuses (les titres sont composés au plus de quelques dizaines de mots). Outre le volume de données à manipuler qui peut vite devenir important, le risque ici est de tomber dans la malédiction de la dimension. La distance entre les vecteurs varie alors assez peu autour d'une valeur moyenne, ce qui a pour effet de rendre cette information totalement inutile.

L'approche retenue dans le cadre de ce projet vise à contourner ces problèmes afin que la notion de distance entre vecteurs représentatifs soit utilisable dans les phases d'analyse suivantes. Via Spark-NLP et des modèles pré-entraînés, il est possible de construire un vecteur de dimension fixe représentatif d'un mot ou d'un groupe de mots. Ces modèles sont entraînés sur des corpus de textes importants et sont capables de produire des vecteurs tenant compte du contexte dans lequel les mots sont habituellement utilisés. Ainsi, deux mots utilisés habituellement dans un même contexte seront associés à des vecteurs proches. À l'inverse, deux mots n'ayant aucun contexte en commun seront associés à des vecteurs dont la distance est importante.

À l'usage, le modèle pré-entraîné mis en œuvre garde quelques défauts. Ainsi, les mots non présents dans le corpus d'entraînement du modèle ne peuvent être traités. Le problème est général à cette approche et concerne des noms propres ou bien des mots n'existant pas dans la langue au moment de l'entraînement. Le vecteur résultant est donc un vecteur nul. Les mots usuels comportant des accents sont aussi ignorés. Ce point n'ayant pas été vu immédiatement, un certain nombre de mots n'ont pas été mis en forme avant d'être soumis au modèle. Une nouvelle itération de la phase d'extraction des vecteurs représentatifs tenant compte de cela permettrait donc de produire des résultats certainement plus fidèle à la réalité (un modèle n'étant de toute manière qu'une représentation imparfaite de la réalité). Il est à noter que si le modèle est effectivement entraîné sur des textes en langue française, des mots en langue anglaise peuvent aussi être représentés par des vecteurs non nuls. Malgré ces quelques quelques défauts, l'approche par vecteur reste pertinente. On veillera toutefois à ne pas faire reposer l'analyse uniquement sur cette approche.

```
print(mod.predict("Trump")["word_embedding_word2vec_wac_200"].iloc[0])
[ 7.51417756e-01 -1.17967200e+00  4.13671613e-01  6.43185496e-01
 -1.41080379e+00 -1.42210746e+00 -1.54268968e+00 -6.52850866e-01
...
 -4.65887524e-02  1.23917198e+00 -5.18829487e-02  9.84116644e-02
  6.48881048e-02  6.88070953e-01  2.91203737e-01 -1.10472870e+00]

print(mod.predict("Micout")["word_embedding_word2vec_wac_200"].iloc[0])
[0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
...
```

```
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0.]

print(mod.predict("Coronavirus")["word_embedding_word2vec_wac_200"].iloc[0])
[-1.53459275e+00 -1.41011357e+00  9.32128727e-01 -2.06217751e-01
  1.02102622e-01 -7.41026044e-01 -1.21519208e+00  4.69679952e-01
...
 -1.23606336e+00 -2.56932855e+00 -1.48761183e-01 -2.45899141e-01
 -1.63329601e+00 -2.42774558e+00 -7.76029229e-02 -3.47993612e-01]

print(mod.predict("Covid")["word_embedding_word2vec_wac_200"].iloc[0])
[0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
...
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0.]

print(mod.predict("element")["word_embedding_word2vec_wac_200"].iloc[0])
[ 8.41904804e-02 -4.19549376e-01  9.18343067e-01  1.74338222e-01
 -2.34180287e-01  2.19448757e+00 -1.11910582e+00 -3.56103927e-01
...
  1.38163090e-01 -3.61478239e-01 -9.82962489e-01  6.11730099e-01
  1.33659565e+00 -8.76732111e-01 -5.06609082e-01  6.28046691e-01]

print(mod.predict("élément")["word_embedding_word2vec_wac_200"].iloc[0])
[0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
...
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0.]

print(mod.predict("Hello")["word_embedding_word2vec_wac_200"].iloc[0])
[ 4.12235022e-01 -7.01087594e-01  5.76080501e-01 -4.26186204e-01
  6.25010490e-01 -9.85967755e-01  5.51805675e-01 -1.72511256e+00
...
 -8.48927200e-01 -2.64785826e-01 -9.26363096e-03  1.42057776e-01
 -1.28089607e+00  4.10665959e-01  2.43178904e-01 -1.86839819e-01]
```

Dans le cadre de ce projet, un titre sera représenté par la somme des vecteurs des mots qui le compose. En fonction du nombre de mots du titre, la norme des vecteurs résultants peut être plus ou moins grande. Si le titre est composé de mots habituellement associés dans des contextes similaires, le lien du titre avec le contexte en question sera renforcé. Il est à noter que le contexte en question n'est pas connu en tant que tel. Cette notion de contexte fait plus référence à une représentation interne du modèle qui ne nous est pas accessible. L'un des enjeux ensuite sera donc de réussir à identifier et à étiqueter les contextes qui semblent intéressants.

La recherche des vecteurs associés à chaque titre est une opération coûteuse en temps de calcul. Ici, et dans les mêmes conditions que lors de la phase d'extraction des mots clés, un peu plus de 3 jours auront été nécessaires pour obtenir ces vecteurs. Cette information est donc stockée en base de donnée afin de pouvoir être réutilisée dans la phase d'analyse.

4.7 Extraction d'informations quantitatives des titres bruts

Longueur des titres :

```
UPDATE donnees_pour_analyse
SET nb_char = length(titre)-1;
```

Nombre de caractères en majuscule :

```
CREATE VIEW mots_maj AS (SELECT id, COUNT(*) AS nb_maj
FROM (
    SELECT id, regexp_matches(titre, '[A-Z]', 'g')
    FROM donnees_pour_analyse) AS blop
    GROUP BY id
);

UPDATE donnees_pour_analyse
SET nb_char_maj = mots_maj.nb_maj
FROM mots_maj
WHERE donnees_pour_analyse.id = mots_maj.id;
```

Nombre de mots dans le titre :

```
CREATE VIEW mots AS (SELECT id, COUNT(*) AS nb_mots
FROM (
    SELECT id, regexp_matches(titre, ' ', 'g')
    FROM donnees_pour_analyse) AS blop
    GROUP BY id
);

UPDATE donnees_pour_analyse
SET nb_mots = mots.nb_mots
FROM mots
WHERE donnees_pour_analyse.id = mots.id;
```

Nombre de chiffres :

```
CREATE VIEW chiffres AS (SELECT id, COUNT(*) AS nb_chiffres
FROM (
    SELECT id, regexp_matches(titre, '[0-9]', 'g')
    FROM donnees_pour_analyse) AS blop
    GROUP BY id
);

UPDATE donnees_pour_analyse
SET nb_chiffres = chiffres.nb_chiffres
FROM chiffres
WHERE donnees_pour_analyse.id = chiffres.id;
```

Ces données extraites directement des titres bruts peuvent éventuellement permettre de compléter l'analyse des actualités. L'usage permettra de définir si ces informations ont ou non une importance significative dans ce cadre.

5 Recherche des actualités dominantes

Notebooks Jupyter correspondant à ce chapitre : RechercheActualitesDominantes-partie1.ipynb (identification de communautés de titres) et RechercheActualitesDominantes-partie2.ipynb (mise en évidence d'actualités à partir des communautés identifiées). La table contenant les données issues de la phase de pré-traitement est disponible en téléchargement [9].

Le processus complet d'analyse de la phase de recherche d'actualité et de la phase de recherche de différences de traitement de l'information entre les flux RSS est présenté en figure 5.1.

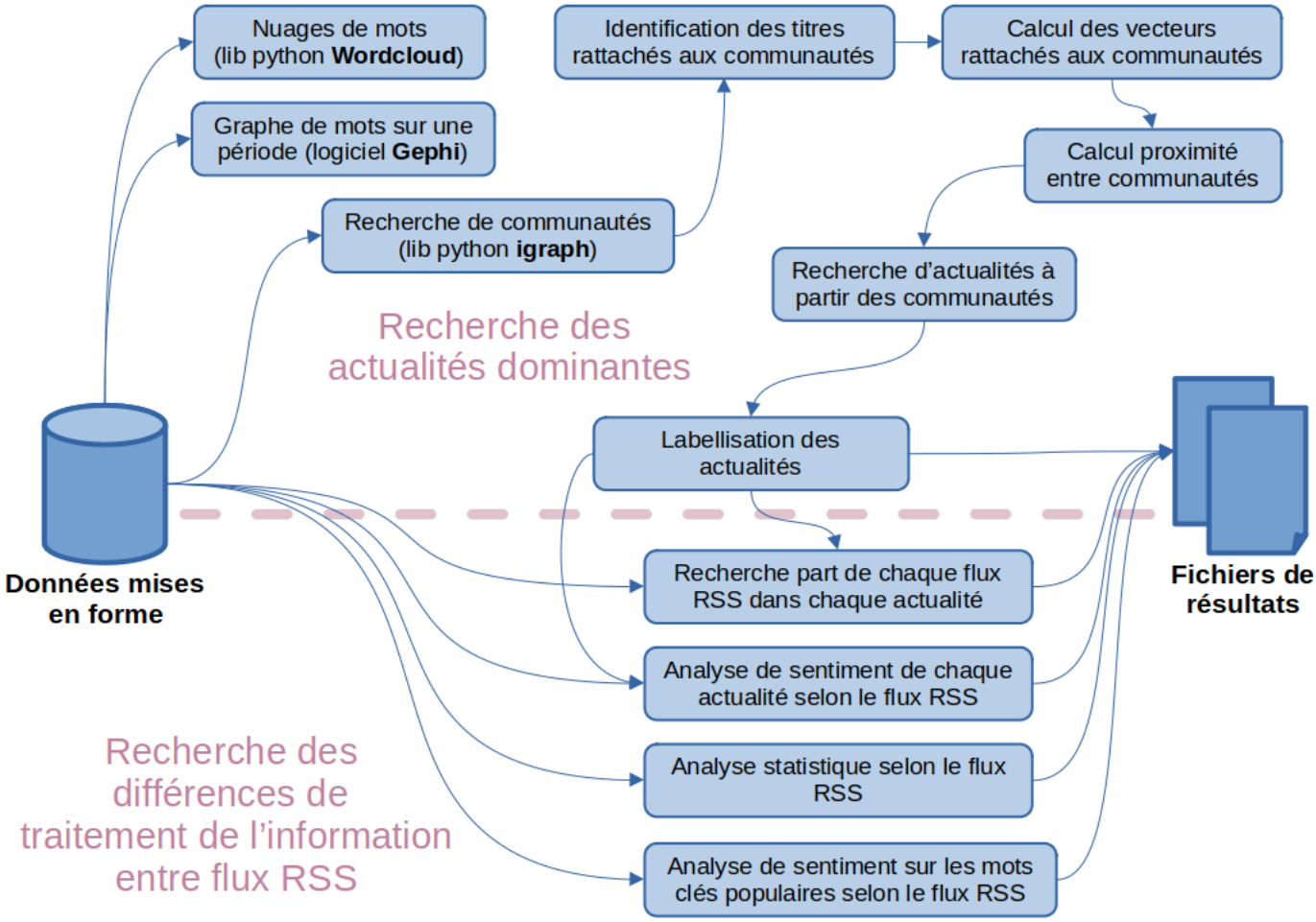

FIGURE 5.1 – Processus d'analyse complet

5.1 Objectif

Cette phase d'extraction d'informations doit permettre de répondre à la première problématique posée. La source de donnée est la table produite à l'issue de la phase de pré-traitement, stockée en base de donnée Postgres distribuée par Citus. Ces données de base ne sont jamais modifiées lors de cette étape et ne sont donc accédées qu'en lecture. Spark manipule des Dataframes qui ont comme propriété de ne pas être modifiables une fois créés (comme peuvent l'être des listes ou des tableaux). Cette

contrainte forte, en apparence, permet à Spark d'optimiser les accès sur ces éléments et de paralléliser les traitements. Un autre avantage des Dataframes dans Spark est qu'il est possible de créer des vues manipulables ensuite en SQL.

5.2 Définitions et hypothèses de travail

Ce projet suppose de définir précisément ce qu'est une actualité dominante, notamment dans notre contexte. Une actualité est comprise ici comme un "Evénement actuel intéressant un domaine d'activité" [11]. Le domaine d'activité n'est pas précisé implicitement mais dans la mesure où les flux RSS traitent d'informations générales en France, le domaine d'activité est la France en général. De la même manière, l'adjectif dominant est compris ici comme quelque chose "Qui a la prépondérance par l'influence, le nombre, l'étendue" [12]. Cette notion est très subjective et suppose de définir un seuil de décision (basé sur une métrique à définir) lors du process. La contrainte peut théoriquement être définie soit sur des paramètres d'entrée (nombre d'occurences d'une expression, nombre de liens entre mots clés, ...), soit sur des paramètres de sortie (nombre d'actualités à retenir, ...).

Dans l'absolu, l'usage fait de ce jeu de données mis en forme n'est pas connu par avance (ces données pouvant être utilisées dans d'autres buts que ceux fixés dans ce projet). Le stockage éventuel des données produites (tableaux ou visualisations) est donc totalement libre et dépend des besoins des utilisateurs finaux. Ici, le choix est fait de travailler si nécessaire avec des exports en CSV pour les tableaux de données et des exports en PNG pour les diverses visualisations.

Dans l'ensemble du jeu de données, le nombre de mots clés est important. De même, la période couverte par cet ensemble est de 5 ans environ. Vu les objectifs du projet, un découpage des enregistrements sur des tranches temporelles semble donc pertinent pour une partie des traitements au moins. Dans le travail réalisé, la tranche temporelle retenue est celle du mois, ce qui revient à découper le jeu de données en 60 fragments environ.

Les différents exemples pris dans ce chapitre traitent la même période temporelle (à savoir février 2022). Ce choix est arbitraire. Il permet de présenter et de comparer les résultats obtenus par différentes approches.

5.3 Nuages de mots

La première approche de représentation consiste à produire des nuages des mots à partir des mots clés. Cela permet de représenter très rapidement à l'utilisateur final une partie de l'information contenue dans les données. Un nuage de mots sur toute la période considérée peut donner une tendance générale (voir figure 5.2).

L'outil employé ici est WordCloud. Plusieurs paramètres peuvent être ajustés. Il est notamment possible de choisir le nombre de mots à représenter, la taille maximale de la police, le masque (donnant une forme particulière au nuage)... Dans le contexte de ce projet, il faut tenir compte du fait que certains mots clés sont en fait plusieurs mots séparés par un caractère d'espace. Wordcloud propose notamment deux approches concernant les expressions composées :

- Traiter chaque mot de manière indépendante. Cela qui conduit à mettre en avant des mots de liaison hors de leur contexte. Par exemple, l'expression *tour de France* fait ressortir les mots *tour*, *de* et *France*. Le mot *de* n'a pas d'intérêt particulier hors de son contexte ici.
- Traiter les mots seuls et conserver les paires de mots. Cela est utile pour ne pas séparer des paires qui ont du sens conjointement mais peut devenir contre-productif dès lors que l'expression

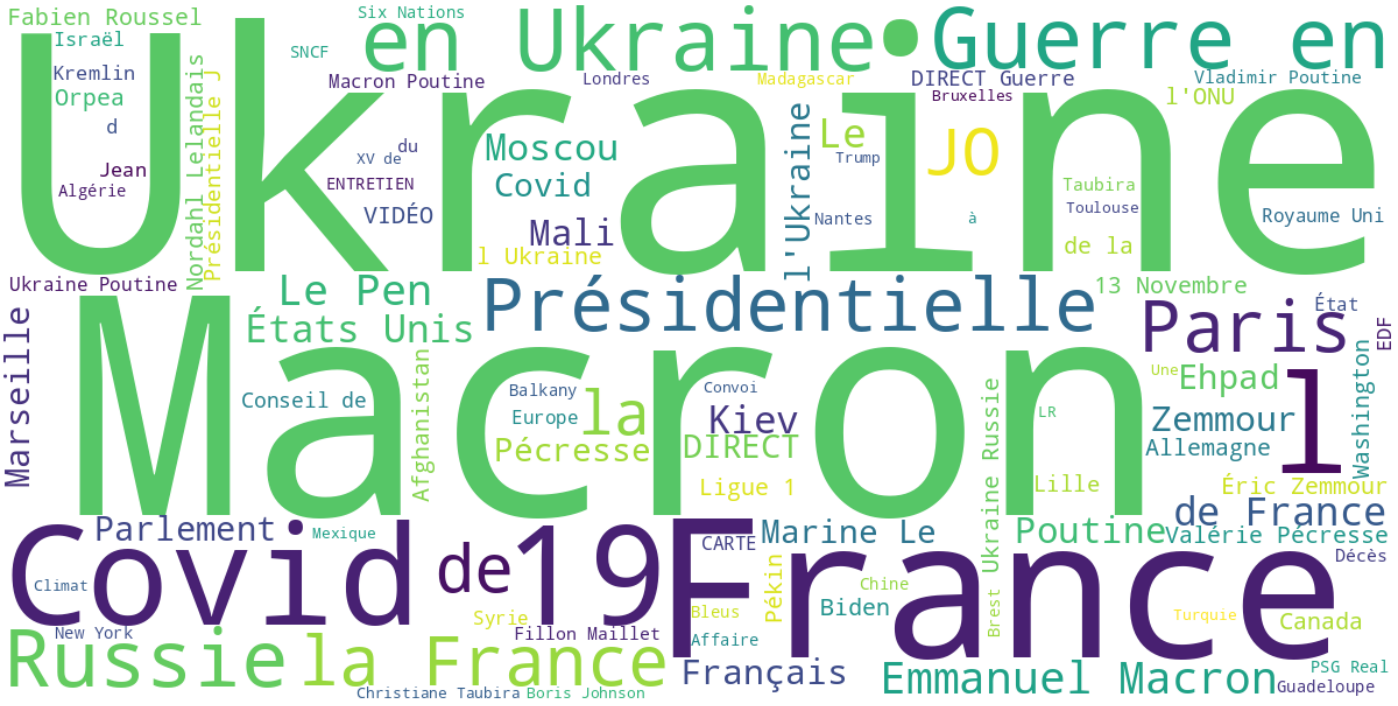

FIGURE 5.2 – Nuage de 100 mots - Février 2022

contient 3 mots ou plus. Dans l'exemple *tour de France*, les paires *tour de* et *de France* sont remontées. D'une part, le mot de liaison *de* ne peut être supprimé ici et d'autre part, cette approche conduit à sur-représenter les mots intermédiaires dans les expressions de plus de deux mots. Ce défaut reste relativement peu pénalisant à l'usage dans la mesure où les expressions de trois mots ou plus restent minoritaires. Cette seconde approche semble donc préférable.

L'utilisation du nuage de mot est intéressante mais s'accompagne de quelques limitations :

- L'évolution temporelle d'un ensemble de mots clés n'est pas aisé car la position de mêmes mots entre deux nuages successifs n'est pas commune. Un exemple de périodes successives représentées sous forme de nuages de mots est présenté en annexe B.
- La moindre modification du jeu de donnée en entrée peut modifier complètement la répartition des mots.
- Il faut produire autant de nuages de mots que d'ensembles à traiter, ce qui peut vite rendre les résultat indigestes.
- Certains titres ne contiennent aucun mot clé et ne sont donc pas représentés.
- Certains titres contiennent plus d'un mot clé et sont donc sur-représentés.

5.4 Graphes des mots clés associés sur une période donnée

Une expression peut faire référence à une actualité précise. Pour autant, cela n'est pas systématique. Trouver des liens forts entre des expressions permet de remonter à un niveau sémantique supérieur. Dit autrement, la combinaison de plusieurs mots clés (ou expressions) peut permettre de remonter à une actualité (ou du moins, à gagner en précision quand à la description de cette actualité). Il est envisageable de ne retenir qu'un nombre limité de couples de mots clés (expression) par période en ne sélectionnant par exemple que ceux ayant 3 liens ou plus (voir figure 5.3). A partir de ces sous-ensembles les plus employés remontés par périodes, il est possible de construire des graphes qui permettent d'identifier assez facilement, de manière visuelle, des liens forts entre des expressions sur une période donnée. L'avantage de la limitation aux expressions les plus présentes est que le graphe gagne en lisibilité.

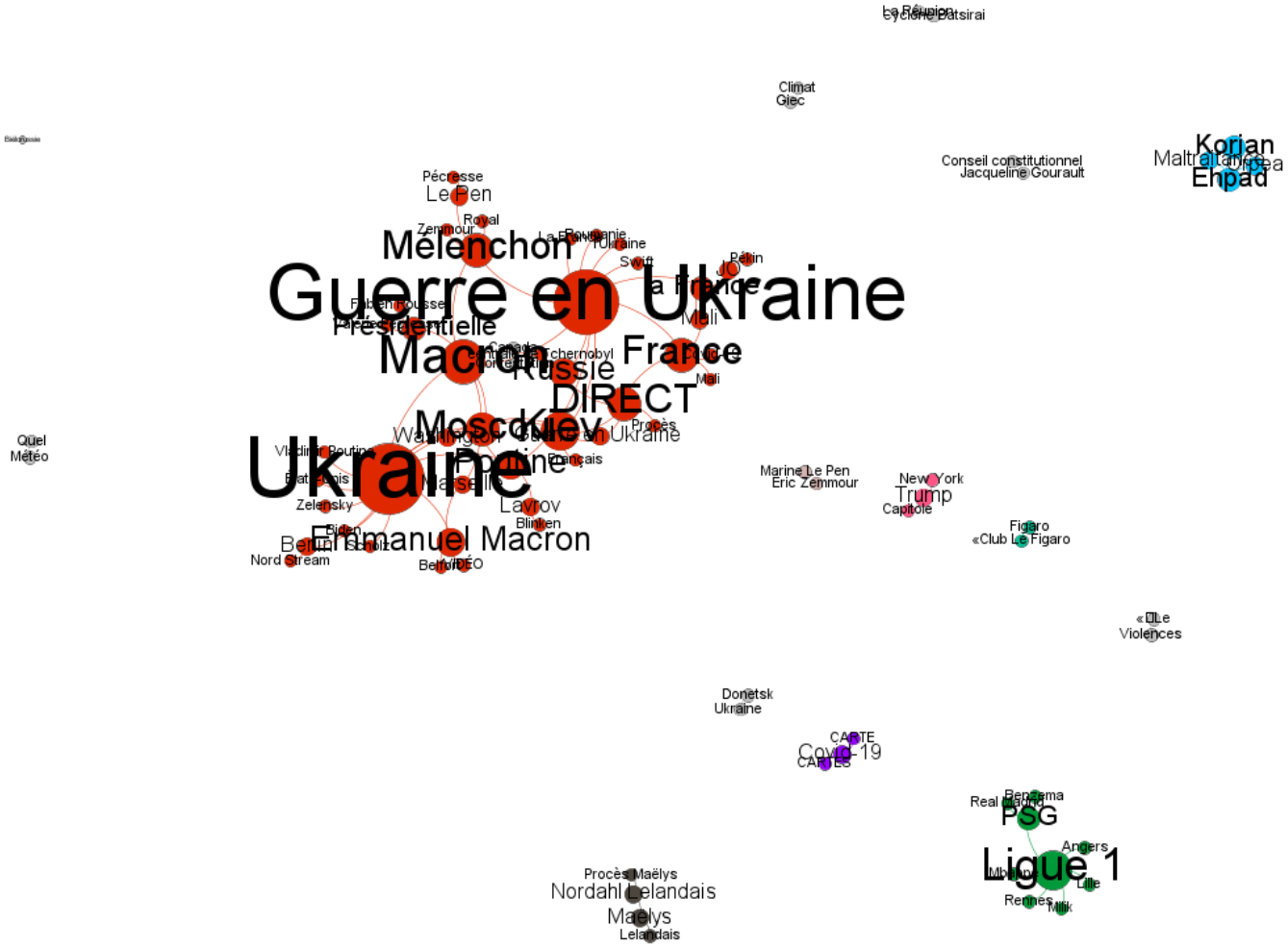

FIGURE 5.3 – Graphe des liens entre les expressions les plus présentes en février 2022

Le graphe en figure 5.3, mis en forme via le logiciel Gephi [6] semble plus pertinent que le nuage de mots présenté en figure 5.2 couvrant la même période. L'outil permet de mettre en évidence les expressions clés les plus employées mais permet aussi d'identifier des groupes partiellement, voir totalement disjoints. Ces sous-ensembles sont clairement liés à des actualités dominantes sur la période considérée et ont donc un intérêt ici.

5.5 Détection automatisée de communautés

Deux approches sont envisageables pour constituer des groupes de mots/expressions clés à partir d'un graphe où ne sont conservés que les nœuds reliés par un nombre minimum de liens. La première approche consiste simplement à regrouper tous les nœuds connectés dans le graphe. Comme il est possible de le voir sur la figure 5.3, cela conduit à rassembler des expressions qui n'ont finalement que peu de choses en commun (JO de Pékin, présidentielles 2022 et guerre en Ukraine). La seconde approche consiste à exécuter des algorithmes de détection de communauté. Globalement, ces différents algorithmes permettent de rassembler les nœuds fortement connectés et amènent à découper des graphes connectés en communautés.

La présentation sous forme de graphe via le logiciel Gephi conserve certains problèmes déjà identifiés lors de la présentation en nuages de mots et en introduit de nouveaux :

— La mise en forme des données via Gephi est un process partiellement manuel.

— Le rendu visuel est fonction d'un critère d'ordre esthétique, à la fois subjectif et difficilement automatisable.
— L'évolution temporelle d'un ensemble de mots clés n'est pas aisé car la position de mêmes mots entre deux graphes successifs n'est pas commune et peut très rapidement être modifié en fonction de l'algorithme de spacialisation employé.
— Il faut produire autant de graphes que d'ensembles à traiter, ce qui peut vite rendre les résultat indigestes.
— Certains titres ne contiennent aucun mot clé et ne sont donc pas représentés.
— Certains titres contiennent plus d'un mot clé et sont donc sur-représentés.
— La détection de communautés reste possible visuellement mais reste empirique et très subjective. Il est toutefois possible de d'exécuter un algorithme de détection de communautés sur un graphe dans Gephi mais cela reste une procédure à réaliser manuellement.

Cela dit, la mise en forme en graphe reste très intéressante dans le cadre de la mise en œuvre d'algorithmes de détection de communautés. La visualisation en tant que telle des graphes passe alors au second plan. Ici, la librairie igraph [8] est utilisée. Cet outil permet de produire et de manipuler directement des graphes à partir de données présentées sous forme de dataframes. Différents algorithmes de détection de communautés sont directement intégrés à cette bibliothèque. L'approche employée sur cette partie peut se représenter sous la forme du process suivant (à réaliser sur chaque période considérée) :

1. Extraction des expressions et du nombre d'occurences associées.
2. Extraction des paires d'expressions et du nombre de liens associés.
3. Recherche de communautés en spécifiant un nombre de liens minimum (ici, on ne conserve que les couples ayant plus de 7 liens).
4. Pour chaque communauté identifiée, associer les id des titres rattachés et leur nombre, ainsi que les expressions associées dans le groupe. A noter qu'un même id peut être associé à deux communautés différentes.

Comme on peut le voir, cette approche implique de spécifier un seuil sur le nombre de liens. Plus ce seuil diminue, plus les nœuds du graphe sont liés entre eux et plus ce seuil augmente, moins le graphe comporte de nœuds et de laisons entre les nœuds restants. Ce choix a été réalisé de manière empirique et permet ici de conserver généralement entre 4 et 10 communautés par période (un seuil important peut par ailleurs réduire le temps de calcul nécessaire à la phase de détection de communautés). Les expressions clés qui représentent les groupes sont celles qui doivent permettre ensuite de définir le label associé à l'actualité.

Sur le graphe (figure 5.4 : 18 expressions et 17 liens, regroupements en 7 communautés), le regroupement des mots en communautés peut conduire à différents cas de figure. Certaines communautés sont naturellement bien séparées des autres puisqu'elle ne contiennent pas un nombre de liens suffisants avec les expressions des autres communautés. Les groupes d'expressions reliées par des liens peuvent, s'il sont trop importants, être découpés en différentes communautés. Ce découpage est fonction de l'algorithme de détection de communauté employé.

La fonction de détection de communautés utilisée ici est une implémentation de la méthode de Louvain [2] dans igraph. La méthode de Louvain permet de regrouper les nœuds d'un graphe en communautés. Cet algorithme cherche à maximiser la modularité totale du graphe. Cela revient à maximiser le rapport entre la somme des arêtes des nœuds dans une communauté d'une part et la somme des arêtes entre les nœuds de communautés différentes d'autre part. L'algorithme est itératif et alterne deux phases (optimisation de la modularité et aggrégation de communautés). L'algorithme se termine lorsque la

modularité atteint un maximum local (cela signifie donc que la solution proposée n'est pas nécessairement la solution optimale). Une portion des résultats (couvrant le premier semestre 2022) est présenté en annexe C. Les résultats couvrant le mois de février 2022 sont présentés à suivre :

annee	mois	member	mots_cles
2022	2	0	[Poutine, Ukraine, Macron, Moscou, Kiev]
2022	2	1	[Russie, Guerre en Ukraine]
2022	2	2	[JO, Pékin]
2022	2	3	[Zemmour, Le Pen, Pécresse]
2022	2	4	[Maëlys, Nordahl Lelandais]
2022	2	5	[DIRECT, Guerre en Ukraine]
2022	2	6	[Ehpad, Orpea]

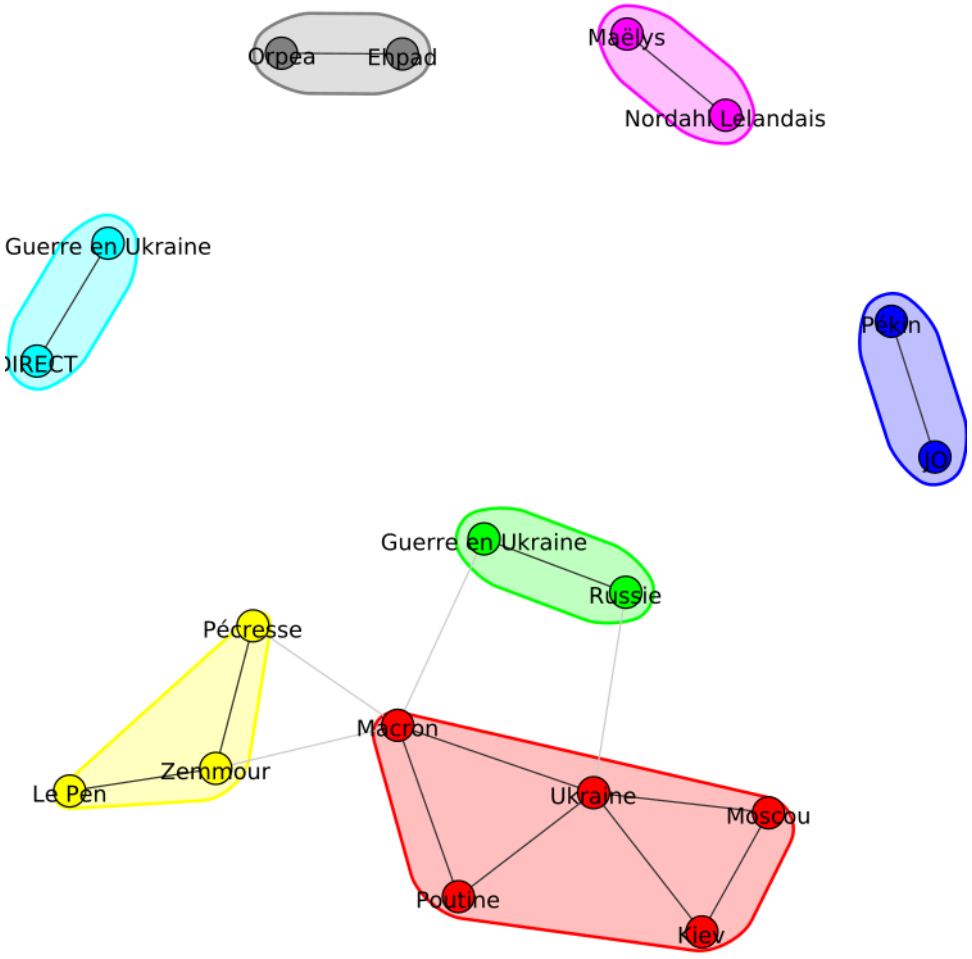

FIGURE 5.4 – Communautés identifiées en février 2022

Comme il est possible de la voir sur la période prise en exemple, les communautés 0, 1 et 5 traitent de la même thématique (la guerre en Ukraine) et pourraient tout à fait être rattachées. Toutefois, cela n'est pas automatique. Ici par exemple la communauté 5 n'a pas de liens avec les communautés 0 et 1 sur le graphe. Cela est lié à la présence d'un caractère non imprimable en fin de chaîne sur "Guerre en Ukraine " qui fait que cette expression est dissociée de celle sans le caractère non imprimable. A priori, ce caractère non imprimable est introduit lors de la phase de détection des mots clés. Comme

on peut le voir ensuite, ce type de problème n'est pas réellement pénalisant. Autre point notable dans cet exemple, la communauté 3 traite d'une actualité différente de celle traitée par les communautés 0 et 1 (les élections présidentielles de 2022 en France d'une part et la guerre en Ukraine d'autre part). Le découpage par l'algorithme de détection de communauté a été en mesure d'identifier ce point malgré la présence de liens notables sur le graphe.

5.6 Associer des communautés proches

Les sous-ensembles d'expressions clés associées à chaque groupe sont limités. Mécaniquement, tous les titres de la période analysée ne sont pas représentés dans les différentes communautés, soit parce qu'ils ne contiennent pas de mots clés fortement connectés à d'autres, soit parce qu'ils ne contiennent pas du tout de mots clés remontés lors de l'analyse par Spark-NLP. Cela n'est toutefois pas problématique, au contraire. Les communautés proposées sont réellement représentatives de liens forts et particuliers entre certaines expressions.

On peut noter par ailleurs qu'il peut exister des passerelles plus ou moins importantes entre les communautés remontées. En effet, certains titres peuvent tout à fait être présents dans plus d'une communautés (en fonction de la répartition de leurs expressions clés dans les diverses communautés). Il semble donc pertinent de rechercher, par une approche complémentaire à celle de l'analyse des graphes, quelle distance peut exister entre les communautés. Ce commentaire est tout à fait valable pour identifier la proximité entre des communautés présentes sur des tranches temporelles adjacentes.

Pour répondre à ces besoins d'analyse de proximité, les vecteurs obtenus lors de la phase de prétraitement semblent tout à fait indiqués. Ici, chaque communautés (approximativement 500 dans les conditions de l'analyse) est donc associée à un vecteur. Chaque vecteur est construit à partir de la somme terme à terme des vecteurs de tous les titres de la communauté qu'il représente. Le calcul de distance peut ensuite être opéré uniquement entre les communautés les plus proches temporellement (ce qui a par ailleurs pour effet de réduire le nombre de calculs à réaliser). Le critère de distance retenu est la distance cosinus (plus précisément, un critère de similarité = 1 - la distance cosinus) car il ne tient pas compte de la longueur des vecteurs mais uniquement de leur orientation. Le point d'origine de l'ensemble des vecteurs est unique. Il est ainsi possible de comparer des communautés comportant des nombres très différents de membres.

Pour revenir sur la problématique liée à la présence de caractères non imprimables dans les expressions clés, cela n'a pas d'incidence sur la génération des vecteurs sur les mots. En effet, le process de recherche de vecteur se base sur les titres bruts. Les mots des titres sont correctement découpés et pris en compte sans qu'intervienne un caractère non imprimable. Ainsi, comme il est possible de la voir de la tableau ci-dessous, la communauté 2022_2_5 (correspondant à la communauté 5 précédente) est très proche des communautés 2022_2_0 et 2022_2_1 (correspondant respectivement aux communautés 0 et 1 précédentes ; similarité égale à 0.76 avec l'une et 0.73 avec l'autre). A l'inverse, la communauté 2022_2_3 (correspondant à la communauté 3 précédente) est peu similaire aux autres communautés identifiés. Ce critère de similarité permet donc de rapprocher les communautés proches comportant des mots employés dans des contextes similaires et d'éloigner les communautés dont les mots sont employés dans des contextes différents.

communaute_A	nb_titres_A	mots_cles_A	communaute_B	nb_titres_B	mots_cles_B	similarite_cosinus
2022_2_0	423	Poutine;Ukraine;M...	2022_2_1	210	Russie;Guerre en ...	0.92521954
2022_2_0	423	Poutine;Ukraine;M...	2022_2_5	67	DIRECT;Guerre en ...	0.762179
2022_2_1	210	Russie;Guerre en ...	2022_2_5	67	DIRECT;Guerre en ...	0.73869276
2022_2_0	423	Poutine;Ukraine;M...	2022_2_2	86	JO;Pékin	0.3681345
2022_2_1	210	Russie;Guerre en ...	2022_2_2	86	JO;Pékin	0.36713156

2022_2_2	86	JO;Pékin	2022_2_5	67	DIRECT;Guerre en ...	0.31236634
2022_2_4	33	Maëlys;Nordahl Le...	2022_2_5	67	DIRECT;Guerre en ...	0.30849737
2022_2_5	67	DIRECT;Guerre en ...	2022_2_6	49	Ehpad;Orpea	0.28645578
2022_2_3	87	Zemmour;Le Pen;Pé...	2022_2_4	33	Maëlys;Nordahl Le...	0.2813962
2022_2_0	423	Poutine;Ukraine;M...	2022_2_4	33	Maëlys;Nordahl Le...	0.2730139
2022_2_4	33	Maëlys;Nordahl Le...	2022_2_6	49	Ehpad;Orpea	0.25320098
2022_2_3	87	Zemmour;Le Pen;Pé...	2022_2_5	67	DIRECT;Guerre en ...	0.23998488
2022_2_0	423	Poutine;Ukraine;M...	2022_2_6	49	Ehpad;Orpea	0.23418868
2022_2_2	86	JO;Pékin	2022_2_4	33	Maëlys;Nordahl Le...	0.22020483
2022_2_1	210	Russie;Guerre en ...	2022_2_6	49	Ehpad;Orpea	0.21545415
2022_2_1	210	Russie;Guerre en ...	2022_2_4	33	Maëlys;Nordahl Le...	0.21413825
2022_2_0	423	Poutine;Ukraine;M...	2022_2_3	87	Zemmour;Le Pen;Pé...	0.21310654
2022_2_3	87	Zemmour;Le Pen;Pé...	2022_2_6	49	Ehpad;Orpea	0.18257065
2022_2_2	86	JO;Pékin	2022_2_6	49	Ehpad;Orpea	0.15418316
2022_2_1	210	Russie;Guerre en ...	2022_2_3	87	Zemmour;Le Pen;Pé...	0.100984745
2022_2_2	86	JO;Pékin	2022_2_3	87	Zemmour;Le Pen;Pé...	0.04482174

Le commentaire réalisé entre les communautés identifiées sur la même période est aussi valable entre communautés adjacentes. Le tableau ci-dessous présentant la similarité cosinus entre les communautés de février 2022 et celles de mars 2022 donne en effet des résultats similaires (voir annexe D). La similarité cosinus est proche de 1 pour les communautés très proches et inférieure à 0,5 pour celles éloignés (0 étant la valeur minimale de la similarité). Cela confirme qu'il est intéressant de suivre ce paramètre afin de regrouper des communautés proches temporellement (soit sur une même tranche temporelle, soit sur une tranche adjacente).

Là aussi, une représentation sous forme de graphe présente un intérêt (communaute_A, communaute_B, similarite_cosinus). En fixant un seuil en dessous duquel les liens ne sont pas conservés, il est possible de consituer des ensembles de communautés fortement liées entre elles et de fait, de remonter à des actualités dominantes. Le seuil retenu est 0,7 car comme il est possible de la voir sur la figure 5.5, assez peu de distances calculées sont présentes autour de cette valeur. De plus, le nombre de liens dans chaque tranche de 1% commence à s'accroître significativement dès lors que l'on continue de descendre en dessous de ce seuil.

La représentation du graphe constitué à partir des communautés similaires (seuil = 0,7) est présenté en figure 5.6. Ce graphe contient beaucoup d'informations mais est difficilement exploitable. Toutefois, il permet de visualiser rapidement les différentes topologies d'actualités dominantes que l'on peut retrouver dans le jeu de données.

5.7 Caractériser les actualités dominantes obtenues

Le cas d'actualités dominantes représentées par un seul mot clé n'est pas considéré ici. Ce choix est justifié par le fait qu'une expression seule ne peut pas réellement représenter un fait d'actualité surtout si ce fait d'actualité est dominant (c'est à dire développé de manière approfondie et de manière régulière sur une période donnée). Ce sont les liens d'une expression avec d'autres qui permettent de traiter une actualité selon différents angles de vue. Les différentes étapes d'analyse montrent que l'on peut trouver divers types d'actualités que l'on peut qualifier de dominantes.

5.7.1 Actualités couvrant des périodes temporelles larges

Un certain nombre d'actualités couvrent une étendue temporelle assez large. Dans ce cas, plusieurs communautés identifiées sur des tranches temporelles successives sont identifiées comme proches. Chacune de ces communautés est caractérisée par des liens forts entre plusieurs expressions clés.

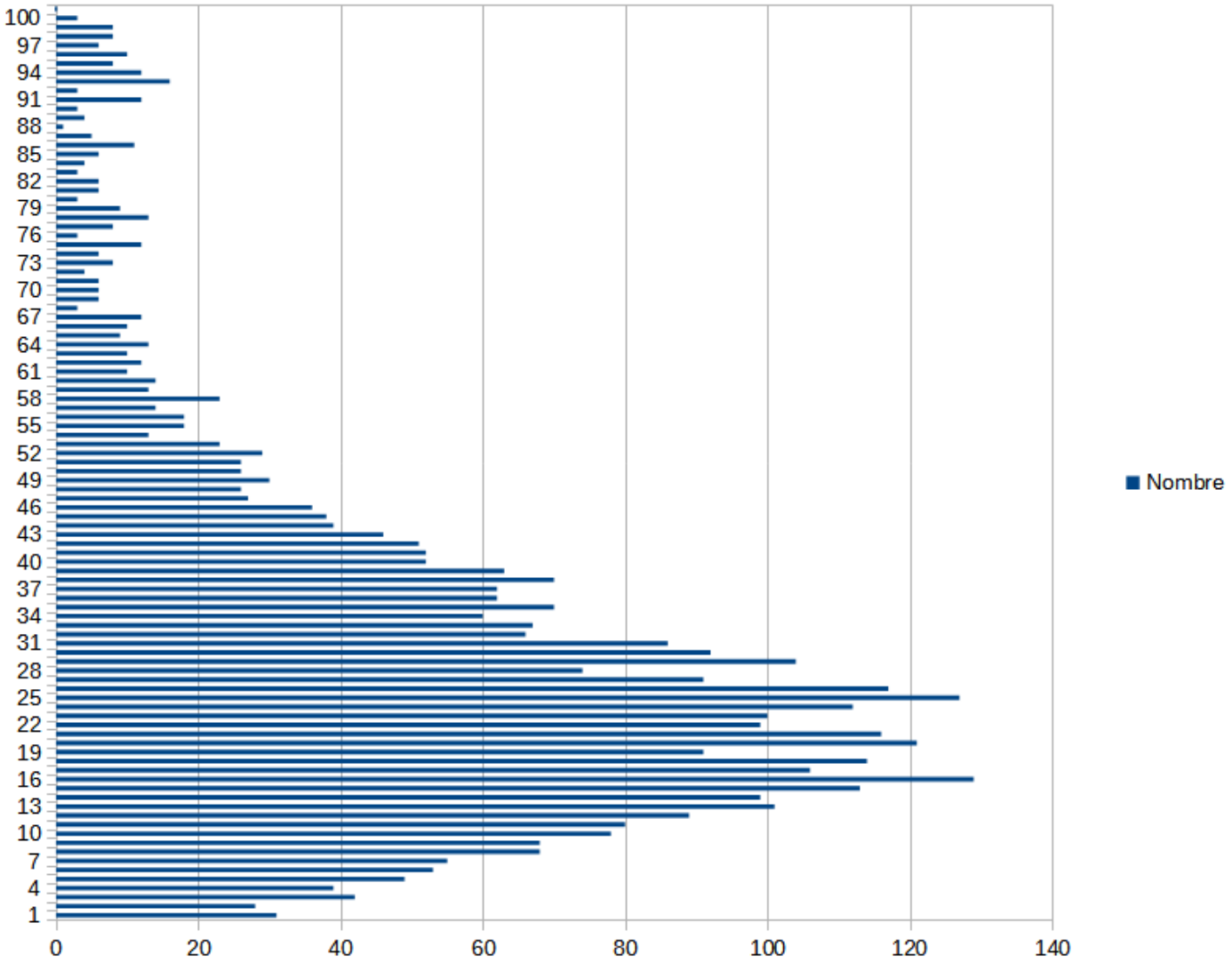

FIGURE 5.5 – Répartition des distances calculées entre communautés (avec regroupement des valeurs de similarité dans des intervalles de 1%)

Afin de caractériser ces différentes actualités, il est nécessaire de préciser la période temporelle couverte et l'ensemble des identifiants des titres rattachés à cette actualité. Ce dernier élément permet d'extraire les mots clés les plus employés afin de générer une étiquette à rattacher à l'actualité traitée. De manière empirique, les 7 mots/expressions clés les plus présents sont conservés pour générer cette étiquette. Ici, chaque sous-graphe connecté constitue une actualité (voir en annexe E). Les paramètres d'entrée permettent de trouver 60 actualités distinctes sur la période étudiée. Les résultats du script d'extraction des labels d'actualités sont présentés en annexe F.

5.7.2 Actualités couvrant des périodes temporelles ciblées

Un certain nombre d'autres actualités sont plus isolées. Elles sont uniquement représentées par une seule communauté caractérisée par quelques expressions. Ces actualités couvrent donc des périodes très ciblées.

Le processus de caractérisation de ces actualités est tout à fait comparable à celui mis en place dans le cadre des actualités couvrant des périodes temporelles étendus, à ceci près que la période est directement donnée par la communauté traitée. Les paramètres d'entrée permettent de trouver 255 actualités distinctes sur la période étudiée. Les résultats de la phase d'extraction des labels d'actualités

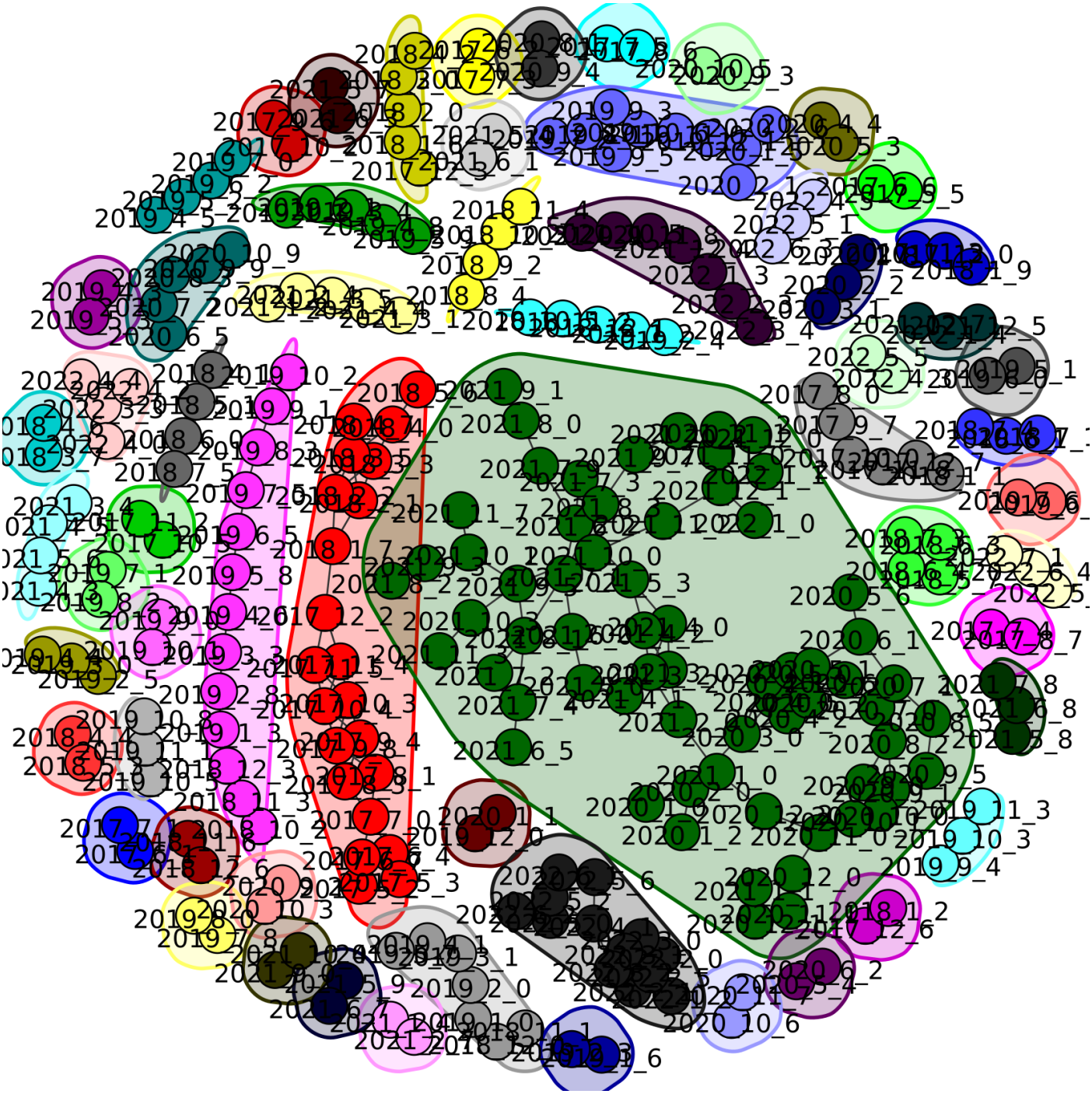

FIGURE 5.6 – Groupes de communautés identifiées sur l'ensemble des périodes

sont présentés en annexe G.

Dans la mesure où le nombre de titres rattaché à chaque actualité est inférieur au cas précédent, le nombre de mots clés conservés pour constituer le label de l'actualité est fixé à 5. Les mots/expressions clés sont présentés de manière décroissante par rapport au nombre de fois où ils sont présents dans les titres traités. Très rapidement dans certains cas, les mots remontés n'ont finalement rien de spécifiques vis à vis de l'actualité à représenter.

6 Analyse des différences de traitement de l'information

Notebook Jupyter correspondant à ce chapitre : RechercheDifferencesTraitementActualites.ipynb . Le processus correspondant à cette phase est présenté en figure 5.1 (présentation commune avec le processus de la phase de recherche d'actualités dominantes).

6.1 Objectif

L'objectif de cette partie est de déterminer dans quelle mesure des différences dans le traitement de l'information entre les différents flux RSS suivis peuvent être identifiées. Ce travail repose sur les données produites lors de la phase de recherche des actualités dominantes et sur plusieurs informations non exploitées jusqu'à présent, à savoir le flux RSS d'origine, le sentiment associé à chaque titre et diverses caractéristiques collectées sur les titres bruts (nombre de caractères, nombre de majuscules, nombre de mots, nombre de chiffres).

6.2 Analyse générale sur la période

6.2.1 Statistiques des données brutes

Le jeu de données analysé comporte 413130 titres issus de 5 flux RSS. La répartition de ces données n'est pas homogène entre les flux RSS. Il est donc déjà possible d'imaginer que la phase de découverte des actualités dominantes réalisée lors de la phase précédente soit plus influencée par certains flux RSS (Le Figaro et Ouest-France) que par d'autres (Mediapart en particulier). Cela ne constitue pas forcement un biais dans le traitement de l'information par un média ou un autre mais peu influer sur la préception que l'on peut en avoir.

link	id_feed	nombre_titres
http://www.ouest-france.fr/rss.xml	24	140589
http://www.letelegramme.fr/france/rss.xml	25	42626
http://www.mediapart.fr/articles/feed	30	18504
http://www.humanite.fr/rss/actu.rss	34	42426
http://www.lefigaro.fr/rss/figaro_flash-actu.xml	50	168985

summary	sentiment	nb_mots	nb_char	nb_char_maj	nb_chiffres
count	413130	413130	413130	413130	413130
mean	null	11.05735482777818	73.31731900370343	3.3168276329484665	0.7080192675429041
stddev	null	3.116258718528414	19.595846878132065	2.3961381679988984	1.557048167020605
min	Negatif	1	2	0	0
max	Positif	45	250	129	19

6.2.2 Métriques reccueillies

Les données produites lors de la phase d'extraction des métriques (reccueillies à l'issue de la phase de pré-traitement) sont présentées en annexe H.

La figure 6.1, qui présente la proportion de sentiments positifs pour chaque flux RSS, met en évidence une sur-représentation de textes identifiés comme positifs sur le site de l'Humanité (proche d'un titre sur deux (48% des titres)). À l'inverse, le flux RSS du Télégramme présente une sous-représentation de la proportion de titres identifiés avec un sentiment positif (seulement 34%). Les valeurs retrouvées pour les trois autres flux semblent plus homogènes (entre 40% et 42%). Ces informations peuvent servir de référence pour l'analyse d'actualités précises.

La figure 6.2, qui présente l'influence de la valeur moyenne des différentes métriques observées sur le sentiment identifié, ne montre pas d'écart significatifs entre les différents flux RSS. On notera que les titres identifiés comme ayant un sentiment positif comportent en moyenne légèrement plus de chiffres que ceux dont le sentiment est identifié comme négatif. Toutefois, ces écarts étant peu significatifs, on considèrera qu'il n'est pas nécessaire d'en tenir compte pour la suite.

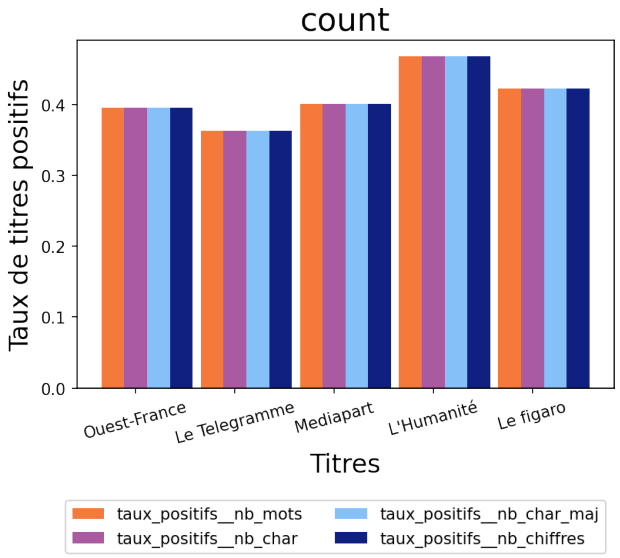

FIGURE 6.1 – Proportion de titres avec sentiment positif par flux RSS

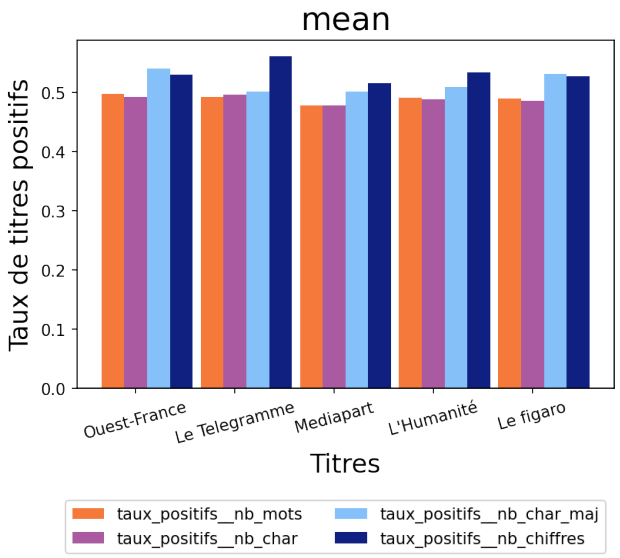

FIGURE 6.2 – Influence de la valeur moyenne des métriques observées sur le sentiment identifié

6.2.3 Analyse de sentiments sur les mots clés les plus populaires

Chaque mot ou expression clé est associé à un ensemble de titres et chaque titre est associé à un sentiment remonté de manière automatique par Spark-NLP. Un mot clé ou une expression clé est plus ou moins utilisé dans des contextes positifs. Il apparaît comme intéressant de voir pour quels mots ou expressions le taux de présence dans un contexte positif varie le plus d'un flux RSS à un autre. L'objectif ici est d'identifer les mots potentiellement les plus clivants entre les flux RSS.

Afin de réduire la quantité de mots à traiter, seules les expressions présentes au moins 30 fois sur un flux RSS (quel que soit le sentiment identifié) sont conservés. Les données ainsi extraites sont conservées uniquement pour les mots présents dans au moins 4 flux RSS, ceci afin de pouvoir établir des comparaisons. Les résultats présentés en figure 6.3 mettent en évidence les mots clés pour lesquels l'écart de taux de titres positifs est maximal entre deux des flux. Le choix de ces seuils est arbitraire et est fonction d'un compromis entre la précision désirée et la complexité d'analyse.

Les écarts maximum constatés entre les taux de sentiments positifs ne peuvent pas toujours s'expliquer de manière simple. Ainsi, le mot clé "Cannes", à priori assez neutre, semble relativement clivant entre le flux de l'Humanité (près de 2/3 de titres positifs) et celui de Ouest-France (moins d'un tiers de titres

mot_cle	Ouest-France	Le Télégramme	Mediapart	L'Humanité	Le Figaro	Écart maxi entre flux RSS
l'Assemblée	82,23	50	46,97	36,51		45,72
Cannes	30,86	39,68		64,71	47,3	33,85
États-Unis	34,88	66,04		68,48	43,89	33,6
Mali	31,65	41,07	7,89	12,24	39,25	33,18
Enquête	43,1	10		22,73	21,89	33,1
Parlement	37,85	56,36	33,33	26,67	43,06	29,69
Marseille	31,99	28,31	34,92	56,25	36,7	27,94
Seine-Saint-Denis	10,48	6,67		31,82	17,03	25,15
Sarkozy	34,69	14,13	22,73		38,24	24,11
Législatives	39,71	39,47		29,36	51,67	22,31
Toulouse	26,29	16,24	24,24		38,22	21,98
Lyon	41,82	26,53		36,54	48,17	21,64
Emmanuel Macron	61,75	71,43	65,98	50,3	70,13	21,13
Israël	40,7		38,71	30,3	51,41	21,11
Strasbourg	45,92	36,45		57,14	46,1	20,69
la France	54,33	55,47	39,63	53,37	60,18	20,55
PS	27,39	23,27	30,51	42,86	28,93	19,59
Paris	36,72	31,9	35,88	51,12	44,61	19,22
Algérie	47,41		66,2	54,31	64,6	18,79
Allemagne	31,1		45,45	27,18	36,29	18,27
Poutine	30,32		29,41	21,43	39,44	18,01
Mélenchon	26,71	30,41	44,62		36,24	17,91
CGT	17,09	15,49		33,33	17,37	17,84
Noël	61,22	68		59,38	51,02	16,98
Europe	42,93	33,33		45,58	50,15	16,82
Ukraine	32,75		48,75	40,17	39,81	16
Sénat	25,72	19,69		15,94	31,76	15,82
Macron	48,12	46,03	42,7	39,16	54,92	15,76
Syrie	24,27	26,67	34,29	31,48	39,98	15,71
La France	49,91	48,02	53,95	53,06	62,96	14,94
LR	34,38	27,52	38,16	34,62	41,64	14,12
SNCF	14,81	16,29	18,18	28,79	18,12	13,98
Tunisie	30		32,31	24,53	37,78	13,25
Blanquer	14,52	24,56	19,61	11,54	13,79	13,02
France	53,81	51,81	41,77	54,12	54,54	12,77
Justice	22,62	12,22		13,12	24,71	12,49
FN	25,26	19,35	25		31,58	12,23
Covid-19	27,96	30,95	22,33	21,61	33,67	12,06
LREM	26,13	21,13	32,63		33,16	12,03
Présidentielle	63,2	62,5		68,53	74,42	11,92
Turquie	24,1		29,55	20,54	32,3	11,76
Jean Castex	36,34	35,29		31,37	42,98	11,61
RN	30,53	39,51	28,09	30	37,36	11,42
Violences	9,15	8,29	2,84	6,88	14,23	11,39
Royaume-Uni	22,92	33,93		26,67	30,58	11,01
Climat	50,85	58,46		50,49	61,21	10,72
Procès	8,94	11,54		17,57	18,49	9,55
Italie	29,72		37,93	28,57	31,56	9,36
Trump	31,13	37,93	40,15	32,82	35,47	9,02
Santé	54,33	58,44		49,71	52	8,73
Les	32,47	36,95	33,62	38,78	40,83	8,36
Ehpad	28,74	27,12		35,19	29,92	8,07
l'ONU	54,37				62,44	8,07
Marine Le Pen	33,46	29,49	30,88	33,87	37,46	7,97
Brésil	29,42		35,44	28,75	36,69	7,94
Brexit	17,43	12,5	16,67	13,33	20,35	7,85
Retraites	26,34	20,31		27,27	27,54	7,23
Covid	32,24		30,43	25,54	32,08	6,7
Français	53,05	46,91		52,63	53,43	6,52
Russie	28		33,85	32,65	34,4	6,4
Gilets	21,76	16,32		22,45	21,05	6,13
la Chine	89,78				91,22	1,44

FIGURE 6.3 – Taux de titres positifs par mot et par flux RSS

positifs). Pour d'autres mots clés tel que "Mali", les flux de Médiapart et de l'Humanité présentent chacun un taux titre positif très faible (respectivement 7,89% et 12,24%) tandis qu'il se situe entre 30% et un peu plus de 40% pour les autres flux RSS analysés. Cet écart peu potentiellement s'expliquer par la ligne éditoriale des diverses rédactions et leur positionnement vis à vis de l'opération militaire de la France au Mali. Ces deux exemples montrent que les mots clés seuls ne permettent pas de comprendre systématiquement toutes les situations. Il peuvent par contre donner des pistes intéressantes quand aux orientations éditoriales des rédactions à l'origine de ces flux d'informations.

6.3 Analyse générale des actualités

Deux aspects sont potentiellement intéressants à observer lorsqu'il s'agit du lien entre les actualités et les flux RSS. Le premier concerne l'importance donnée à telle ou telle actualité par chacun des flux RSS. Le second concerne le taux de sentiments positifs par actualité et par flux RSS. Les fichiers CSV générés par le script sur cette partie sont ensuite manipulés dans un tableur.

6.3.1 Importance des actualités selon le flux RSS

Pour réaliser une analyse de ces données, il est impératif de tenir compte du fait qu'il existe un déséquilibre en terme de nombre de titres entre les flux observés. Un seuil différent est donc appliqué pour chaque flux lors de l'opération de mise en forme conditionnelle. Ce seuil est déterminé en fonction de la part globale de chaque flux dans l'ensemble des actualités dominantes et permet d'une certaine manière de normaliser la visualisation des données. Les actualités sur-représentées par chaque flux RSS sont alors nettement distinguables.

Ouest-France	47,93%
Le Télégramme	11,75%
Médiapart	2,47%
L'Humanité	4,62%
Le Figaro	33,22%
TOTAL	100%

En spécifiant un mot dans le champ de label du fichier tableur mis en forme, il devient relativement facile de visualiser l'importance des différentes actualités contenant ce mot dans chaque flux RSS. Les figures 6.4 à 6.8 données en exemple mettent en évidence ce point. Il est ainsi possible de voir que parmis les actualités dominantes le sport est une thématique mise en avant par le Ouest-France et le Figaro (plutôt le football pour le premier et le tennis pour le second). Ces thèmes sont moins mis en avant dans les autres flux RSS. De la même manière, Médiapart en particulier met régulièrement en avant des actualités contenant le nom "Macron" depuis le début de la période étudiée. Le flux de l'Humanité semble faire de même mais cela est plus marqué depuis la fin du premier semestre 2021 (ce qui coïncide avec le début de la période pré-électorale en France). Comme Médiapart, Le Figaro semble s'intéresser de manière plus appuyée aux actualités contenant les noms "Trump" et "Poutine".

actualite	label	date_mini	date_maxi	nb	Ouest-France	Le Télégramme	Mediapart	L'Humanité	Le Figaro
0	Ligue,Football,PSG,DIRECT,Neymar,Mercato,Monaco	2017-05-01	2018-05-26	2880	75,14	1,67	0,21	0,69	22,29
17	DIRECT,PSG,Ligue,Ligue 1,Football Leaks,Lyon,Monaco	2018-08-04	2018-11-29	339	69,91	0,59	3,54	0,59	25,37
40	PSG,Ligue des champions,Football,Neymar,Stade Rennais,Coupe de France,Ligue	2020-06-01	2020-10-31	252	56,75	0,79		1,59	40,87
47	Football,PSG,Stade Rennais,Mbappé,Benzema,Bayern,Football - Ligue des champions	2021-03-01	2021-05-31	249	26,91	0,8		8,84	63,45
2021_3_8	Lyon,Violences,DIRECT,PSG,Football - Ligue 1,Suisse,Ligue,Beauvais,Football - PSG	2021-03-01	2021-03-31	106	31,13	5,66	2,83	15,09	45,28

FIGURE 6.4 – Couverture des actualités contenant le mot "Foot" dans les flux RSS

actualite	label	date_mini	date_maxi	nb	Ouest-France	Le Télégramme	Mediapart	L'Humanité	Le Figaro
3	Roland-Garros,Wimbledon,Tennis,Nadal,Mladenovic,Djokovic,Rafael Nadal	2017-06-01	2017-07-24	202	51,49				48,51
48	Roland-Garros,Tennis,Federer,Nadal,Djokovic,Français,Wimbledon	2021-05-09	2021-06-18	108	30,56			13,89	55,56
59	Roland-Garros,Nadal,Wimbledon,Djokovic,Alcaraz,Kyrgios,Tennis	2022-05-01	2022-07-10	133			0,75	6,77	92,48

FIGURE 6.5 – Couverture des actualités contenant le mot "Tennis" dans les flux RSS

L'analyse de la répartition des titres rattachés à chaque actualité entre les flux RSS permet manifestement de dégager des centres d'intérêts dans tel ou tel flux. Il permet aussi en creux d'identifier des actualités de moindre intérêt pour les diverses sources étudiées. Cette approche semble donc pertinente dans ce cas d'usage.

actualite	label	date_mini	date_maxi	nb	Ouest-France	Le Télégramme	Mediapart	L'Humanité	Le Figaro
1	Trump,FBI,Climat,États-Unis,Macron,Paris,Russie,Comey	2017-05-01	2017-06-30	219	35,16	1,83	4,11	2,28	56,62
2017_5_0	Macron,Présidentielle,Emmanuel Macron,DIRECT,Le Pen	2017-05-01	2017-05-31	654	41,28	21,1	7,8	6,12	23,7
5	Macron,Paris,Trump,Marseille,Français,G20,Philippe	2017-07-01	2017-08-31	677	33,53	16,69	4,73	5,91	39,14
4	Venezuela,Maduro,Constituante,Caracas,Colombie,Trump,Nicolas Maduro,Macron	2017-07-03	2017-08-31	154	27,92	0,65	5,84	12,34	53,25
7	Paris,JO,Bonbonnes,Macron,Jeux,Bourse,Sentinelle,Berlin,Loup Bureau	2017-09-01	2017-10-31	244	40,16	19,26	4,92	5,33	30,33
8	Macron,Violences,Philippe,Guyane,Juppé,Wauquiez,Français	2017-10-01	2017-11-30	364	24,18	27,2	4,4	6,59	37,64
9	Trump,Jérusalem,États-Unis,Poutine,Macron,Israël,Palestiniens	2017-11-01	2018-01-31	433	38,8	2,54	3,23	5,54	49,88
10	Syrie,Macron,Ghouta,Russie,Trump,Douma,Erdogan	2017-12-01	2018-04-30	824	31,92	5,58	5,58	1,33	55,58
11	France,la France,Météo,Gall,Neige,Norvège,Macron	2017-12-01	2018-01-31	148	48,65	23,65	2,7	1,35	23,65
2017_12_1	Johnny Hallyday,Saint-Barthélemy,Emmanuel Macron,Paris,Champs-Élysées,[Direct],IMAGES	2017-12-04	2017-12-30	135	49,63	29,63	0,74	1,48	18,52
2018_2_7	Macron,Corse,Syrie,Merkel,Poutine,Wauquiez	2018-02-01	2018-02-28	132	31,06	19,7	8,33	8,33	32,58
2018_3_11	Macron,Merkel,Affaire Skripal,Moscou,Philippe	2018-03-01	2018-03-29	103	20,39	16,5	11,65	3,88	47,57
2018_3_9	Trump,États-Unis,Stormy Daniels,Macron,Donald Trump,Affaire Skripal,Pékin,Pyongyang,Commerce,Miami	2018-03-01	2018-03-31	116	53,45	1,72	2,59		42,24
2018_6_5	Macron,Migrants,Lifeline,Malte,G7	2018-06-04	2018-06-30	219	40,18	17,35	6,39	4,11	31,96
2018_7_0	Affaire Benalla,Macron,l'Assemblée,Congrès,Collomb	2018-07-01	2018-07-31	274	37,96	29,93	6,93	2,55	22,63
19	Macron,Trump,Grand débat,Gilets,Français,États-Unis,Philippe	2018-10-01	2019-02-28	1073	29,26	22,93	5,03	4,38	38,4
2018_10_1	Trump,Khashoggi,Turquie,Macron,Saoudiens	2018-10-01	2018-10-31	146	23,29	2,74	2,05	2,74	69,18
20	Gilets,Paris,France,DIRECT,Notre-Dame,Macron,Ligue	2018-11-01	2019-05-31	3875	47,54	22,97	1,96	2,22	25,32
25	Macron,Européennes,Europ&eacute,,Français,RN,Grand débat,Corse	2019-04-01	2019-05-31	672	29,76	24,55	5,51	2,98	37,2
2019_6_10	Trump,Macron,États-Unis,G20,Poutine,Marseille,Kim,Elton John	2019-06-01	2019-06-30	172	16,28	10,47	5,23	2,33	65,7
31	Macron,G7,Amazonie,Emmanuel Macron,Brésil,Biarritz,Bolsonaro	2019-07-01	2019-08-31	394	36,29	16,24	3,3	1,78	42,39
33	Paris,Macron,Gilets,Municipales,Attaque,Extinction Rebellion,DIRECT,Tuerie &agrave,,Castaner	2019-09-01	2019-11-30	722	32,83	19,94	3,6	2,49	41,14
2020_6_4	Macron,Convention,Français,Poutine,la France,Emmanuel Macron,Philippe,Merkel	2020-06-01	2020-06-30	114	28,07	22,81	7,89	6,14	35,09
2020_8_0	Beyrouth,Liban,Explosions,Macron,Français	2020-08-04	2020-08-31	249	43,78	4,82	2,81	4,82	43,78
41	Biélorussie,Loukachenko,Minsk,Poutine,Russie,Emmanuel Macron,Moscou,Tikhanovskaïa,Alexandre Loukachenko	2020-08-07	2020-09-29	162	47,53		6,17	4,32	41,98
50	Emmanuel Macron,Macron,Polynésie,Rwanda,Covid-19,DIRECT,Drôme	2021-05-01	2021-07-30	217	35,94	11,98	3,69	12,9	35,48
52	Macron,Marseille,Emmanuel Macron,Bernard Tapie,Présidentielle,Biden,DIRECT	2021-09-01	2021-10-31	469	33,05	18,34	4,9	9,17	34,54
53	Zemmour,Macron,Pécresse,Le Pen,Présidentielle,Mélenchon,LR,DIRECT	2021-09-01	2022-03-31	566	10,07	17,14	10,07	12,01	50,71
2021_9_6	Présidentielle,Anne Hidalgo,Marine Le Pen,Éric Zemmour,LR,Valérie Pécresse,Emmanuel Macron,PS	2021-09-01	2021-09-30	127	52,76	6,3	4,72	25,2	11,02
55	Ukraine,Guerre en Ukraine,Russie,Macron,Kiev,France,Poutine	2022-01-01	2022-06-30	3496	15,5	5,15	5,81	9,35	64,19
56	Macron,Présidentielle,Emmanuel Macron,Marine Le Pen,Le Pen,DIRECT,Mélenchon	2022-03-01	2022-04-30	897	38,24	15,83	9,59	11,15	25,2
2022_6_0	Législatives,Nupes,RN,DIRECT,Macron	2022-06-01	2022-06-30	228	29,39	10,09	8,33	33,33	18,86

FIGURE 6.6 – Couverture des actualités contenant le nom “Macron” dans les flux RSS

actualite	label	date_mini	date_maxi	nb	Ouest-France	Le Télégramme	Mediapart	L'Humanité	Le Figaro
1	Trump,FBI,Climat,États-Unis,Macron,Paris,Russie,Comey	2017-05-01	2017-06-30	219	35,16	1,83	4,11	2,28	56,62
5	Macron,Paris,Trump,Marseille,Français,G20,Philippe	2017-07-01	2017-08-31	677	33,53	16,69	4,73	5,91	39,14
4	Venezuela,Maduro,Constituante,Caracas,Colombie,Trump,Nicolas Maduro,Macron	2017-07-03	2017-08-31	154	27,92	0,65	5,84	12,34	53,25
2017_8_4	Trump,Charlottesville,Corée du Nord,États-Unis,Tempête Harvey	2017-08-01	2017-08-31	111	33,33	0,9	4,5	4,5	56,76
2017_9_13	Trump,Corée du Nord,Pyongyang,Kim Jong-Un,Japon	2017-09-01	2017-09-30	136	42,65	0,74	1,47	4,41	50,74
9	Trump,Jérusalem,États-Unis,Poutine,Macron,Israël,Palestiniens	2017-11-01	2018-01-31	433	38,8	2,54	3,23	5,54	49,88
10	Syrie,Macron,Ghouta,Russie,Trump,Douma,Erdogan	2017-12-01	2018-04-30	824	31,92	5,58	5,58	1,33	55,58
2018_3_9	Trump,États-Unis,Stormy Daniels,Macron,Donald Trump,Affaire Skripal,Pékin,Pyongyang,Commerce,Miami	2018-03-01	2018-03-31	116	53,45	1,72	2,59		42,24
15	Trump,Donald Trump,Corée du Nord,États-Unis,Poutine,G7,Kim Jong-un	2018-05-01	2018-07-31	463	41,04	1,51	3,67	1,94	51,84
19	Macron,Trump,Grand débat,Gilets,Français,États-Unis,Philippe	2018-10-01	2019-02-28	1073	29,26	22,93	5,03	4,38	38,4
2018_10_1	Trump,Khashoggi,Turquie,Macron,Saoudiens	2018-10-01	2018-10-31	146	23,29	2,74	2,05	2,74	69,18
2019_6_10	Trump,Macron,États-Unis,G20,Poutine,Marseille,Kim,Elton John	2019-06-01	2019-06-30	172	16,28	10,47	5,23	2,33	65,7
2019_8_4	États-Unis,Trump,Donald Trump,G7,la Chine,Groenland,Paso,Etats-Unis	2019-08-01	2019-08-31	220	54,55	1,82	0,91	1,36	41,36
2019_10_0	Syrie,Turquie,Trump,France,Kurdes	2019-10-01	2019-10-31	484	28,31	3,51	3,72	1,86	62,6
2020_1_3	Iran,Boeing,Téhéran,Crash,Donald Trump	2020-01-01	2020-01-29	120	34,17	1,67			64,17
2020_1_7	Trump,États-Unis,Soleimani,Sénat,Iran	2020-01-02	2020-01-31	169	42,6	2,96	2,37	2,37	49,7
45	États-Unis,Joe Biden,Trump,Donald Trump,Capitole,Biden,Johnson	2021-01-01	2021-04-30	949	39,41	2,63	1,69	9,59	46,68

FIGURE 6.7 – Couverture des actualités contenant le nom “Trump” dans les flux RSS

actualite	label	date_mini	date_maxi	nb	Ouest-France	Le Télégramme	Mediapart	L'Humanité	Le Figaro
9	Trump,Jérusalem,États-Unis,Poutine,Macron,Israël,Palestiniens	2017-11-01	2018-01-31	433	38,8	2,54	3,23	5,54	49,88
2018_2_7	Macron,Corse,Syrie,Merkel,Poutine,Wauquiez	2018-02-01	2018-02-28	132	31,06	19,7	8,33	8,33	32,58
15	Trump,Donald Trump,Corée du Nord,États-Unis,Poutine,G7,Kim Jong-un	2018-05-01	2018-07-31	463	41,04	1,51	3,67	1,94	51,84
2019_6_10	Trump,Macron,États-Unis,G20,Poutine,Marseille,Kim,Elton John	2019-06-01	2019-06-30	172	16,28	10,47	5,23	2,33	65,7
2020_6_4	Macron,Convention,Français,Poutine,la France,Emmanuel Macron,Philippe,Merkel	2020-06-01	2020-06-30	114	28,07	22,81	7,89	6,14	35,09
41	Biélorussie,Loukachenko,Minsk,Poutine,Russie,Emmanuel Macron,Moscou,Tikhanovskaïa,Alexandre Loukachenko	2020-08-07	2020-09-29	162	47,53		6,17	4,32	41,98
55	Ukraine,Guerre en Ukraine,Russie,Macron,Kiev,France,Poutine	2022-01-01	2022-06-30	3496	15,5	5,15	5,81	9,35	64,19

FIGURE 6.8 – Couverture des actualités contenant le nom “Poutine” dans les flux RSS

6.3.2 Taux de sentiment positif des actualités selon le flux RSS

L'analyse de sentiments sur les mots clés n'est pas suffisante car il n'est pas possible de placer les mots ou les expressions dans leur contexte d'origine. Cette analyse est donc étendue aux actualités dominantes identifiées. Contrairement à l'analyse de la répartition des titres dans les actualités, ici l'échelle de couleur est commune aux cinq flux RSS.

actualite	label	date_mini	date_maxi	nb_titres_actu	Ouest-France	Le Télégramme	Mediapart	L'Humanité	Le Figaro
0	Ligue;Football;PSG;DIRECT;Neymar;Mercato;Monaco	2017-05-01	2018-05-26	2880	49,26	37,5	33,33	45	57,94
17	DIRECT;PSG;Ligue;Ligue 1;Football Leaks;Lyon;Monaco	2018-08-04	2018-11-29	339	53,59		8,33	50	48,84
40	PSG;Ligue des champions;Football;Neymar;Stade Rennais;Coupe de France;Ligue	2020-06-01	2020-10-31	252	53,15			50	51,46
47	Football;PSG;Stade Rennais;Mbappé;Benzema;Bayern;Football - Ligue des champions	2021-03-01	2021-05-31	249	46,27	50		40,91	55,06
2021_3_8	Lyon;Violences;DIRECT;PSG;Football - Ligue 1;Suisse;Ligue;Beauvais;Football - PSG	2021-03-01	2021-03-31	106	24,24	16,67			35,42

FIGURE 6.9 – Sentiment des actualités contenant le mot “Foot” dans les flux RSS

actualite	label	date_mini	date_maxi	nb_titres_actu	Ouest-France	Le Télégramme	Mediapart	L'Humanité	Le Figaro
3	Roland-Garros;Wimbledon;Tennis;Nadal;Mladenovic;Djokovic;Rafael Nadal	2017-06-01	2017-07-24	202	65,38				76,53
48	Roland-Garros;Tennis;Federer;Nadal;Djokovic;Français;Wimbledon	2021-05-09	2021-06-18	108	54,55			73,33	78,33
59	Roland-Garros;Nadal;Wimbledon;Djokovic;Alcaraz;Kyrgios;Tennis	2022-05-01	2022-07-10	133				66,67	68,29

FIGURE 6.10 – Sentiment des actualités contenant le mot “Tennis” dans les flux RSS

actualite	label	date_mini	date_maxi	nb_titres_actu	Ouest-France	Le Télégramme	Médiapart	L'Humanité	Le Figaro
1	Trump;FBI;Climat;États-Unis;Macron;Paris;Russie;Comey	2017-05-01	2017-06-30	219	32,47	25	22,22	60	45,97
2017_5_0	Macron;Présidentielle;Emmanuel Macron;DIRECT;Le Pen	2017-05-01	2017-05-31	654	59,63	60,87	60,78	50	65,81
5	Macron;Paris;Trump;Marseille;Français;G20;Philippe	2017-07-01	2017-08-31	677	51,1	38,05	43,75	37,5	57,36
4	Venezuela;Maduro;Constituante;Caracas;Colombie;Trump;Nicolas Maduro;Macron	2017-07-03	2017-08-31	154	32,56		44,44	31,58	51,22
7	Paris;JO;Bonbonnes;Macron;Jeux;Bourse;Sentinelle;Berlin;Loup Bureau	2017-09-01	2017-10-31	244	47,96	21,28	50	76,92	67,57
8	Macron;Violences;Philippe;Guyane;Juppé;Wauquiez;Français	2017-10-01	2017-11-30	364	48,86	36,36	31,25	20,83	57,66
9	Trump;Jérusalem;États-Unis;Poutine;Macron;Israël;Palestiniens	2017-11-01	2018-01-31	433	40,48	72,73	28,57	41,67	42,13
10	Syrie;Macron;Ghouta;Russie;Trump;Douma;Erdogan	2017-12-01	2018-04-30	824	29,66	36,96	30,43	36,36	41,48
11	France;la France;Météo;Gall;Neige;Norvège;Macron	2017-12-01	2018-01-31	148	51,39	48,57	50		60
2017_12_1	Johnny Hallyday;Saint-Barthélemy;Emmanuel Macron;Paris;Champs-Élysées;[Direct];IMAGES	2017-12-04	2017-12-30	135	80,6	97,5			88
2018_2_7	Macron;Corse;Syrie;Merkel;Poutine;Wauquiez	2018-02-01	2018-02-28	132	41,46	42,31	18,18	18,18	44,19
2018_3_11	Macron;Merkel;Affaire Skripal;Moscou;Philippe	2018-03-01	2018-03-29	103	47,62	41,18	50	50	40,82
2018_3_9	Trump;États-Unis;Stormy Daniels;Macron;Donald Trump;Affaire Skripal;Pékin;Pyongyang;Commerce;Miami	2018-03-01	2018-03-31	116	27,42	50	33,33		32,65
2018_6_5	Macron;Migrants;Lifeline;Malte;G7	2018-06-04	2018-06-30	219	36,36	21,05	28,57	22,22	57,14
2018_7_0	Affaire Benalla;Macron;l'Assemblée;Congrès;Collomb	2018-07-01	2018-07-31	274	17,31	32,93	26,32	42,86	43,55
19	Macron;Trump;Grand débat;Gilets;Français;États-Unis;Philippe	2018-10-01	2019-02-28	1073	47,13	36,99	33,33	31,91	48,3
2018_10_1	Trump;Khashoggi;Turquie;Macron;Saoudiens	2018-10-01	2018-10-31	146	23,53		66,67	25	37,62
20	Gilets;Paris;France;DIRECT;Notre-Dame;Macron;Ligue	2018-11-01	2019-05-31	3875	35,88	25,62	44,74	41,86	42,51
25	Macron;Européennes;Europé;Français;RN;Grand débat;Corse	2019-04-01	2019-05-31	672	44,5	46,67	37,84	35	48
2019_6_10	Trump;Macron;États-Unis;G20;Poutine;Marseille;Kim;Elton John	2019-06-01	2019-06-30	172	46,43	44,44	44,44	75	54,87
31	Macron;G7;Amazonie;Emmanuel Macron;Brésil;Biarritz;Bolsonaro	2019-07-01	2019-08-31	394	61,54	67,19	61,54	42,86	64,67
33	Paris;Macron;Gilets;Municipales;Attaque;Extinction Rebellion;DIRECT;Tuerie à;Castaner	2019-09-01	2019-11-30	722	34,18	25	38,46	50	43,77
2020_6_4	Macron;Convention;Français;Poutine;la France;Emmanuel Macron;Philippe;Merkel	2020-06-01	2020-06-30	114	62,5	69,23	55,56	28,57	62,5
2020_8_0	Beyrouth;Liban;Explosions;Macron;Français	2020-08-04	2020-08-31	249	42,2	41,67	57,14	33,33	55,96
41	Biélorussie;Loukachenko;Minsk;Poutine;Russie;Emmanuel Macron;Moscou;Tikhanovskaïa;Alexandre Loukachenko	2020-08-07	2020-09-29	162	46,75		50	14,29	45,59
50	Emmanuel Macron;Macron;Polynésie;Rwanda;Covid-19;DIRECT;Drôme	2021-05-01	2021-07-30	217	48,72	65,38	50	53,57	49,35
52	Macron;Marseille;Emmanuel Macron;Bernard Tapie;Présidentielle;Biden;DIRECT	2021-09-01	2021-10-31	469	49,68	59,3	60,87	44,19	53,7
53	Zemmour;Macron;Pécresse;Le Pen;Présidentielle;Mélenchon;LR;DIRECT	2021-09-01	2022-03-31	566	45,61	53,61	33,33	50	52,61
2021_9_6	Présidentielle;Anne Hidalgo;Marine Le Pen;Éric Zemmour;LR;Valérie Pécresse;Emmanuel Macron;PS	2021-09-01	2021-09-30	127	68,66	87,5	33,33	75	50
55	Ukraine;Guerre en Ukraine;Russie;Macron;Kiev;France;Poutine	2022-01-01	2022-06-30	3496	40,04	48,33	44,83	43,12	43,54
56	Macron;Présidentielle;Emmanuel Macron;Marine Le Pen;Le Pen;DIRECT;Mélenchon	2022-03-01	2022-04-30	897	64,43	64,79	52,33	52	65,49
2022_6_0	Législatives;Nupes;RN;DIRECT;Macron	2022-06-01	2022-06-30	228	44,78	43,48	47,37	34,21	44,19

FIGURE 6.11 – Sentiment des actualités contenant le nom “Macron” dans les flux RSS

actualite	label	date_mini	date_maxi	nb_titres_actu	Ouest-France	Le Télégramme	Médiapart	L'Humanité	Le Figaro
1	Trump;FBI;Climat;États-Unis;Macron;Paris;Russie;Comey	2017-05-01	2017-06-30	219	32,47	25	22,22	60	45,97
5	Macron;Paris;Trump;Marseille;Français;G20;Philippe	2017-07-01	2017-08-31	677	51,1	38,05	43,75	37,5	57,36
4	Venezuela;Maduro;Constituante;Caracas;Colombie;Trump;Nicolas Maduro;Macron	2017-07-03	2017-08-31	154	32,56		44,44	31,58	51,22
2017_8_4	Trump;Charlottesville;Corée du Nord;États-Unis;Tempête Harvey	2017-08-01	2017-08-31	111	18,92			20	34,92
2017_9_13	Trump;Corée du Nord;Pyongyang;Kim Jong-Un;Japon	2017-09-01	2017-09-30	136	39,66			16,67	42,03
9	Trump;Jérusalem;États-Unis;Poutine;Macron;Israël;Palestiniens	2017-11-01	2018-01-31	433	40,48	72,73	28,57	41,67	42,13
10	Syrie;Macron;Ghouta;Russie;Trump;Douma;Erdogan	2017-12-01	2018-04-30	824	29,66	36,96	30,43	36,36	41,48
2018_3_9	Trump;États-Unis;Stormy Daniels;Macron;Donald Trump;Affaire Skripal;Pékin;Pyongyang;Commerce;Miami	2018-03-01	2018-03-31	116	27,42	50	33,33		32,65
15	Trump;Donald Trump;Corée du Nord;États-Unis;Poutine;G7;Kim Jong-un	2018-05-01	2018-07-31	463	48,42	57,14	64,71	44,44	51,25
19	Macron;Trump;Grand débat;Gilets;Français;États-Unis;Philippe	2018-10-01	2019-02-28	1073	47,13	36,99	33,33	31,91	48,3
2018_10_1	Trump;Khashoggi;Turquie;Macron;Saoudiens	2018-10-01	2018-10-31	146	23,53		66,67	25	37,62
2019_6_10	Trump;Macron;États-Unis;G20;Poutine;Marseille;Kim;Elton John	2019-06-01	2019-06-30	172	46,43	44,44	44,44	75	54,87
2019_8_4	États-Unis;Trump;Donald Trump;G7;la Chine;Groenland;Paso;Etats-Unis	2019-08-01	2019-08-31	220	32,5	75		33,33	43,96
2019_10_0	Syrie;Turquie;Trump;France;Kurdes	2019-10-01	2019-10-31	484	30,66	52,94	44,44	44,44	29,04
2020_1_3	Iran;Boeing;Téhéran;Crash;Donald Trump	2020-01-01	2020-01-29	120	43,9	50			41,56
2020_1_7	Trump;États-Unis;Soleimani;Sénat;Iran	2020-01-02	2020-01-31	169	22,22	20	25		36,9
45	États-Unis;Joe Biden;Trump;Donald Trump;Capitole;Biden;Johnson	2021-01-01	2021-04-30	949	39,57	48	68,75	43,96	49,66

FIGURE 6.12 – Sentiment des actualités contenant le nom “Trump” dans les flux RSS

actualite	label	date_mini	date_maxi	nb_titres_actu	Ouest-France	Le Télégramme	Médiapart	L'Humanité	Le Figaro
9	Trump;Jérusalem;États-Unis;Poutine;Macron;Israël;Palestiniens	2017-11-01	2018-01-31	433	40,48	72,73	28,57	41,67	42,13
2018_2_7	Macron;Corse;Syrie;Merkel;Poutine;Wauquiez	2018-02-01	2018-02-28	132	41,46	42,31	18,18	18,18	44,19
15	Trump;Donald Trump;Corée du Nord;États-Unis;Poutine;G7;Kim Jong-un	2018-05-01	2018-07-31	463	48,42	57,14	64,71	44,44	51,25
2019_6_10	Trump;Macron;États-Unis;G20;Poutine;Marseille;Kim;Elton John	2019-06-01	2019-06-30	172	46,43	44,44	44,44	75	54,87
2020_6_4	Macron;Convention;Français;Poutine;la France;Emmanuel Macron;Philippe;Merkel	2020-06-01	2020-06-30	114	62,5	69,23	55,56	28,57	62,5
41	Biélorussie;Loukachenko;Minsk;Poutine;Russie;Emmanuel Macron;Moscou;Tikhanovskaïa;Alexandre Loukachenko	2020-08-07	2020-09-29	162	46,75		50	14,29	45,59
55	Ukraine;Guerre en Ukraine;Russie;Macron;Kiev;France;Poutine	2022-01-01	2022-06-30	3496	40,04	48,33	44,83	43,12	43,54

FIGURE 6.13 – Sentiment des actualités contenant le nom “Poutine” dans les flux RSS

L’interprétation de ces éléments est complexe à réaliser. En effet, le sentiment dont il est question n’est pas lié directement à l’opinion exprimée dans l’ensemble des titres d’actualité. Ce sentiment est lié aux mots qui composent chaque titre et au fait que ces mots soient utilisés ou non dans des contextes positifs. De plus, l’analyse par échantillonnage a montré que le label de sentiment était potentiellement mal posé dans un peu plus de 20% des cas (avec une marge d’erreur à plus ou mois 5%). Ces chiffres sont meilleurs qu’une affectation aléatoire des labels de sentiment et il est envisageable de prendre en compte des écarts importants de taux de sentiments positifs entre des flux RSS sur une actualité (quelques dizaines de pourcents). Par contre, des écarts plus minimes, de l’ordre de quelques pourcents ne peuvent donner lieu à des interprétations fiables. À noter que les actualités sont ordonnées de manière chronologique (tri sur la date minimum des titres sur la période). Dans une certaine mesure, il est donc possible de voir l’évolution du sentiment des titres dans le temps.

S’agissant des exemples présentés dans les figures 6.9 et 6.10, il est possible de constater que les actualités dont le label contient le mot “Foot” donnent lieu à des titres moins positifs que ceux dont le label contient le mot “Tennis”. La différence est ici assez nette et peut potentiellement s’expliquer par le fait que les actualités autour du football traitent autant de sport que d’affaires en lien. Les actualités autour du tennis se concentrent plus sur le sport en lui même. Ces deux premiers exemples ne pointent donc pas des spécificités propres au traitement de l’information par des flux RSS.

S'agissant des exemples présentés dans les figures 6.11, 6.12 et 6.13, des différences notables du taux de sentiment positif sont visibles entres les flux RSS pour certaines actualités. Le Figaro semble plus enclin que les autres flux à employer le nom "Macron" dans des contextes positifs sur l'ensemble de la période considérées. Cette analyse pourrait potentiellement être affinée en tenant compte de l'évolution dans le temps des taux de titres positifs. D'éventuelles différences de traitement sont moins nettes lorsque l'on examine les résultats obtenus avec le nom "Trump" et il convient probablement de ne pas tirer de conclusions des résultats affichés. Lorsque l'on observe les actualités dont la label contient le nom "Poutine", Médiapart et L'Humanité sont les seuls pour lesquels existent des actualités ayant un faible taux de sentiment positif. Cela n'est pas systématique mais permet d'envisager qu'il existe une spécificité pour eux dans le traitement de l'information concernant Vladimir Poutine.

7 Test des propriété de l'architecture

7.1 Scalabilité

On s'intéresse à la capacité du système à suivre les variations de charge dans le cadre d'un fonctionnement normal. La montée en charge dont il est question ici peut concerner un accroissement du volume de données à gérer ou bien encore un accroissement du nombre de requêtes par les clients. Pour rappel, le process de pré-traitement n'est pas considéré ici car ce traitement n'est à réaliser qu'une seule fois par le gestionnaire du service et que la ressource la plus critique est alors la puissance CPU disponible. On cherche ici à assurer la scalabilité lors de la phase d'analyse. Des questions identiques se posent en cas de réduction du nombre de requêtes. On considère par contre que le volume de données à gérer ne peut pas décroître lors de la phase d'analyse.

L'ajout de workers permert d'accroître la capacité de stockage disponible et de répondre potentiellement à plus de requêtes (ce paramètre étant aussi dépendant de la distribution des données sur les nœuds dans le cluster). Comme indiqué lors de la phase de préparation, l'ajout de worker est une opération triviale avant la distribution des données. Dans le cas où les données sont déjà distribuées sur un certain nombre de nœuds workers, la marche à suivre reste aussi très simple et passe par deux étapes : l'ajout du nouveau worker au cluster puis la redistribution des shards.

```
postgres=# SELECT * from citus_add_node('citus-worker4', 5432);
 citus_add_node
----------------
             54
(1 row)

postgres=# SELECT rebalance_table_shards('donnees_pour_analyse');
NOTICE:  Moving shard 102579 from citus-worker2:5432 to citus-worker4:5432 ...
NOTICE:  Moving shard 102580 from citus-worker3:5432 to citus-worker4:5432 ...
...
NOTICE:  Moving shard 102594 from citus-worker2:5432 to citus-worker4:5432 ...
NOTICE:  Moving shard 102593 from citus-worker1:5432 to citus-worker4:5432 ...
 rebalance_table_shards
------------------------

(1 row)
```

La suppression d'un nœud worker est également très simple. Cela peut être réalisé sans qu'il ne soit utile d'arrêter le cluster. L'opération nécessite de déplacer les shards vers les autres nœuds du cluster puis éventuellement d'enlever ce nœud de la liste des nœuds actifs.

```
postgres=# SELECT * from citus_drain_node('citus-worker2', 5432);
NOTICE:  Moving shard 102580 from citus-worker2:5432 to citus-worker1:5432 ...
NOTICE:  Moving shard 102583 from citus-worker2:5432 to citus-worker3:5432 ...
...
NOTICE:  Moving shard 102609 from citus-worker2:5432 to citus-worker4:5432 ...
NOTICE:  Moving shard 102610 from citus-worker2:5432 to citus-worker1:5432 ...
 citus_drain_node
------------------

(1 row)

postgres=# SELECT * from citus_get_active_worker_nodes();
   node_name   | node_port
---------------+-----------
 citus-worker3 |      5432
 citus-worker2 |      5432
 citus-worker1 |      5432
 citus-worker4 |      5432
```

```
(4 rows)

postgres=# SELECT citus_remove_node('citus-worker2', 5432);
 citus_remove_node
-------------------

(1 row)

postgres=# SELECT * from citus_get_active_worker_nodes();
   node_name    | node_port
---------------+-----------
 citus-worker3 |      5432
 citus-worker1 |      5432
 citus-worker4 |      5432
(3 rows)
```

L'ajout d'un nouveau nœud coordinateur en mode standby est là aussi une opération simple. La procédure est identique à celle présentée dans le chapitre relatif à la préparation des nœuds. Il n'y a pas de modifications particulières de paramétrage à prévoir sur le nœud master. Comme il est possible de le voir, ces processus ont été réalisés manuellement. Cela dit, rien n'interdit de les automatiser afin d'ajuster les ressources allouées aux besoins réels.

7.2 Tolérance à la panne

Dans quelle mesure le système accepte la perte d'un ou de plusieurs nœuds ? On considère que le système répond toujours au besoin s'il est en mesure de fournir des résultats aux clients. La tolérance à la panne est testée sur l'architecture de base présentée en figure 2.1. L'exécution d'un script de recherche des actualités dominantes (TestTolerancePanne.ipynb) permettra de vérifier si le système est toujours opérationnel et dans quelles conditions le service est rendu ou non.

	Workers OK	**1 worker KO**	**2 workers KO**
Coordinateurs OK	Fonctionnement nominal	Accès en mode dégradé	Non fonctionnel
Coord. Standby KO	Fonctionnement nominal	Accès en mode dégradé	Non fonctionnel
Coord. Master KO	Accès aux données OK	Accès en mode dégradé	Non fonctionnel
Coordinateurs KO	Non fonctionnel	Non fonctionnel	Non fonctionnel

- **Fonctionnement nominal** : le service est complètement fonctionnel.
- **Accès aux données OK** : le service est rendu à l'utilisateur final mais l'ajout de données n'est plus possible.
- **Accès en mode dégradé** : le service est rendu à l'utilisateur mais l'accès aux données est très dégradé (au temps nécessaire à l'éxécution de chaque requête s'ajoute un timeout de 2s lié à la non détection d'un worker).
- **Non fonctionnel** : le service est totalement indisponible à l'utilisateur.

En somme, tant qu'un nœud coordinateur est accessible, le service est accessible de manière transparente. Le système tolère la perte d'un nœud worker lorsque les shards ont un facteur de réplication égal à 2 mais très vite, les performances s'effondrent. Ce dernier constat laisse suggérer que des optimisations dans le paramétrage de Citus peuvent être opérées afin de réduire le timeout lorsqu'un nœud n'est plus accessible. Des optimisations peuvent éventuellement être réalisées dans les scripts d'analyse afin de limiter le nombre de requêtes nécessaires à l'obtention des résultats. Toutefois, la charge de la résolution du problème est alors transférée à l'utilisateur final, ce qui n'est pas forcement le meilleur choix. Enfin, cela laisse aussi suggérer que la méthode de réplication intégré dans Citus n'est pas optimale et que l'approche consistant à mettre en place la "Streaming replication" sur les workers est à envisager sérieusement (voir les préconisation à ce propos dans la documantation de Citus).

8 Conclusion

Ce projet aura été l'occasion d'implémenter une solution basée sur le couple PostgreSQL / Citus dans le cadre du stockage de données massives et distribuées. Le volume de données ne justifiait pas à lui seul la mise en place d'une solution distribuée mais les essais menés ont pu démontrer que l'architecture mise en place passait effectivement bien à l'échelle. Un autre avantage de cette approche est qu'elle rend possible la reprise de bases de données existantes de façon transparente pour l'utilisateur final, sans perte des propriétés des bases relationnelles (Citus n'est en effet qu'une extension à PostgreSQL).

Le pré-traitement des données a été rendu possible grâce à l'utilisation de modèles génériques pré-entraînés injectés dans Spark-NLP. Les résultats obtenus ne sont pas tous du même niveau. Si l'extraction des mots clés sur les titres et des vecteurs représentatifs des mots des titres ont donné des résultats globalement pertinents, l'extraction de sentiment s'est révélée moins efficace. Cela vient donc rappeler que les modèles existants ne sont pas nécessairement efficaces en toutes circonstances et qu'une grande partie de leur efficacité dépend du type de données en entrée. Il aurait aussi été éventuellement possible d'employer un modèle de détection de sentiments entraîné sur un corpus de "tweets" par exemple (les titres des flux RSS ont un nombre caractère moyen du même ordre de grandeur que celui des tweets). Quoi qu'il en soit, un regard critique sur ce point est donc impératif car il est risqué de baser une analyse sur des données brutes ou intermédiaires de mauvaise qualité.

La recherche d'actualités dominantes a permis d'expérienter plusieurs approches. Des approches visuellement intéressantes (nuages de mots et graphe dans Gephi) sur des périodes précises ont été présentées. Ces méthodes ne permettent toutefois pas un réel passage à l'échelle. Le choix a donc été fait de travailler sur une analyse de graphes basée sur les couples de mots clés et procéder à de la détection de communauté par un algorithme dédié (Louvain). La prise en compte des vecteurs représentatifs des titres a permis de faire émerger des actualités dominantes par l'aggrégation de communautés proches. Le critère de distance cosinus a permis d'objectiver les choix de regroupements. Ainsi, la combinaison de ces deux données issues de la phase de pré-traitement (mots clé et vecteurs) a aidé à répondre à la première partie de la problématique du projet.

Les différences de traitement de l'information entre les flux RSS se sont révélé être plus complexe à identifier. La mise en évidences des sujets d'intérêt a permis de montrer qu'il y avait des actualités dominantes plus traitées que d'autres selon le flux RSS observé. Toutefois, la prise en compte du taux de sentiment positif s'est révélée moins précise. En deça d'un écart significatif des taux de sentiments positifs entre deux flux RSS sur un ensemble d'actualité, il n'est pas possible d'exploiter cette information. Globalement, les données extraites lors de cette phase ont permis de répondre au moins partiellement à la seconde partie de la problématique.

Annexes

A Stack technique

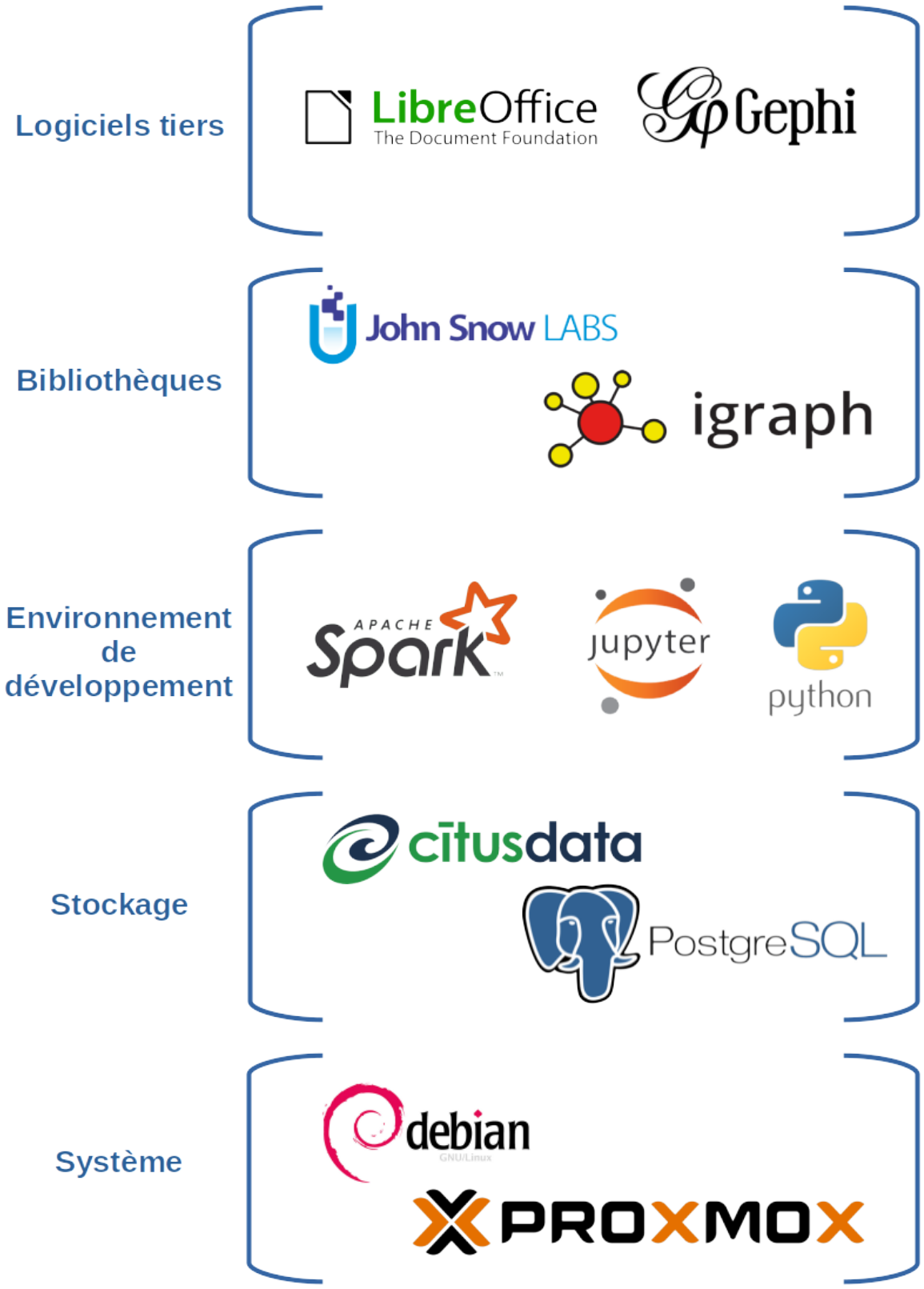

FIGURE A.1 – Outils mis en œuvre dans le cadre de ce projet

B Nuages de mots

FIGURE B.1 – Nuage de 100 mots - Janvier 2022

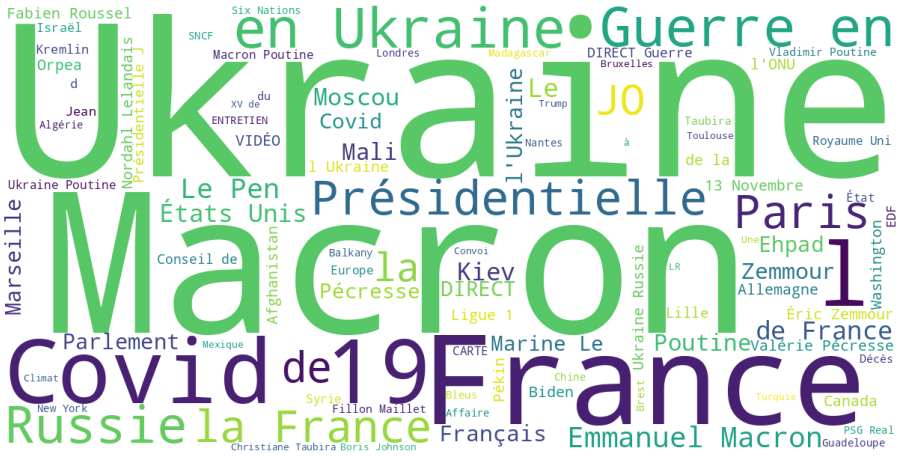

FIGURE B.2 – Nuage de 100 mots - Février 2022

FIGURE B.3 – Nuage de 100 mots - Mars 2022

C Communautées identifiées

annee	mois	member	mots_cles
2022	1	0	[Covid-19 , DIRECT, Olivier Véran, Procès, Gabriel Attal, Emmanuel Macron, Ligue 1, Dembélé]
2022	1	1	[Covid-19, Omicron, France]
2022	1	2	[Russie, Ukraine, Moscou, États-Unis, Washington]
2022	1	3	[Macron, Pécresse, Le Pen, Zemmour]
2022	1	4	[Anne Hidalgo, Présidentielle, Christiane Taubira]
2022	1	5	[Mbappé, PSG]
2022	1	6	[Al-Attiyah, Dakar, Loeb]
2022	1	7	[Djokovic, Tennis]
2022	1	8	[Ibiza, Jean-Michel Blanquer]
2022	2	0	[Poutine, Ukraine, Macron, Moscou, Kiev]
2022	2	1	[Russie, Guerre en Ukraine]
2022	2	2	[JO, Pékin]
2022	2	3	[Zemmour, Le Pen, Pécresse]
2022	2	4	[Maëlys, Nordahl Lelandais]
2022	2	5	[DIRECT, Guerre en Ukraine]
2022	2	6	[Ehpad, Orpea]
2022	3	0	[Russie, Ukraine, Guerre en Ukraine, France, Moscou, Zelensky, Marioupol, Londres, VIDÉO]
2022	3	1	[Poutine, Macron, Biden, Pécresse, Mélenchon, Le Pen, Zemmour]
2022	3	2	[Kiev, DIRECT, Guerre en Ukraine , Russes]
2022	3	3	[Emmanuel Macron, Présidentielle]
2022	3	4	[Colonna, Corse]
2022	3	5	[Deschamps, Giroud, Bleus]
2022	3	6	[CARTE, Covid-19]
2022	4	0	[Le Pen, Macron, Mélenchon]
2022	4	1	[Guerre en Ukraine, Kiev, Ukraine, Russie, Marioupol, Moscou, Boutcha, France, Zelensky]
2022	4	2	[Emmanuel Macron, Marine Le Pen, Présidentielle, VIDÉO, STORY]
2022	4	3	[Elon Musk, Twitter]
2022	4	4	[DIRECT, Présidentielle]
2022	4	5	[Ligue 1, PSG]
2022	5	0	[Finlande, Suède]
2022	5	1	[Mbappé, PSG]
2022	5	2	[Guerre en Ukraine, Russie, Kiev, Ukraine, Moscou, Azovstal]
2022	5	3	[Nadal, Roland-Garros, Djokovic, Alcaraz]
2022	5	4	[LFI, PS]
2022	5	5	[Elon Musk, Twitter]
2022	5	6	[Poutine, Macron, Scholz]
2022	6	0	[Législatives, Nupes, RN]
2022	6	1	[Guerre en Ukraine, Kiev, Russie, Ukraine, Sieverodonetsk, Macron]
2022	6	2	[Finlande, Suède, Turquie]
2022	6	3	[Mercato, PSG, Galtier]
2022	6	4	[Nadal, Roland-Garros]
2022	6	5	[Canicule , DIRECT]

D Similarité entre communautés adjacentes

communaute_A	nb_titres_A	mots_cles_A	communaute_B	nb_titres_B	mots_cles_B	similarite_cosinus
2022_2_1	210	Russie;Guerre en ...	2022_3_0	636	Russie;Guerre en ...	0.9830284
2022_2_3	87	Zemmour;Le Pen;Pé...	2022_3_4	62	Le Pen;Zemmour	0.9802742
2022_2_0	423	Poutine;Ukraine;M...	2022_3_1	696	Ukraine;Poutine;M...	0.97723234
2022_2_1	210	Russie;Guerre en ...	2022_3_1	696	Ukraine;Poutine;M...	0.9401461
2022_2_0	423	Poutine;Ukraine;M...	2022_3_2	243	Kiev;DIRECT;Guerr...	0.9264452
2022_2_0	423	Poutine;Ukraine;M...	2022_3_0	636	Russie;Guerre en ...	0.9249519
2022_2_1	210	Russie;Guerre en ...	2022_3_2	243	Kiev;DIRECT;Guerr...	0.91089916
2022_2_5	67	DIRECT;Guerre en ...	2022_3_2	243	Kiev;DIRECT;Guerr...	0.83410114
2022_2_5	67	DIRECT;Guerre en ...	2022_3_1	696	Ukraine;Poutine;M...	0.7842864
2022_2_5	67	DIRECT;Guerre en ...	2022_3_0	636	Russie;Guerre en ...	0.7804709
2022_2_5	67	DIRECT;Guerre en ...	2022_3_3	169	Emmanuel Macron;P...	0.57764804
2022_2_0	423	Poutine;Ukraine;M...	2022_3_3	169	Emmanuel Macron;P...	0.5156629
2022_2_3	87	Zemmour;Le Pen;Pé...	2022_3_3	169	Emmanuel Macron;P...	0.49381468
2022_2_2	86	JO;Pékin	2022_3_6	51	Deschamps;Giroud;...	0.484214
2022_2_4	33	Maëlys;Nordahl Le...	2022_3_5	68	Colonna;Corse	0.46377245
2022_2_2	86	JO;Pékin	2022_3_1	696	Ukraine;Poutine;M...	0.39020905
2022_2_5	67	DIRECT;Guerre en ...	2022_3_6	51	Deschamps;Giroud;...	0.37067437
2022_2_5	67	DIRECT;Guerre en ...	2022_3_7	25	CARTE;Covid-19	0.36829904
2022_2_3	87	Zemmour;Le Pen;Pé...	2022_3_5	68	Colonna;Corse	0.35918707
2022_2_2	86	JO;Pékin	2022_3_0	636	Russie;Guerre en ...	0.35781437
2022_2_1	210	Russie;Guerre en ...	2022_3_3	169	Emmanuel Macron;P...	0.33378902
2022_2_4	33	Maëlys;Nordahl Le...	2022_3_1	696	Ukraine;Poutine;M...	0.32848093
2022_2_2	86	JO;Pékin	2022_3_2	243	Kiev;DIRECT;Guerr...	0.32833084
2022_2_4	33	Maëlys;Nordahl Le...	2022_3_3	169	Emmanuel Macron;P...	0.32327545
2022_2_4	33	Maëlys;Nordahl Le...	2022_3_7	25	CARTE;Covid-19	0.31552142
2022_2_0	423	Poutine;Ukraine;M...	2022_3_6	51	Deschamps;Giroud;...	0.30026147
2022_2_1	210	Russie;Guerre en ...	2022_3_6	51	Deschamps;Giroud;...	0.29527667
2022_2_2	86	JO;Pékin	2022_3_3	169	Emmanuel Macron;P...	0.29447156
2022_2_0	423	Poutine;Ukraine;M...	2022_3_7	25	CARTE;Covid-19	0.2819051
2022_2_4	33	Maëlys;Nordahl Le...	2022_3_0	636	Russie;Guerre en ...	0.26491112
2022_2_6	49	Ehpad;Orpea	2022_3_1	696	Ukraine;Poutine;M...	0.2629843
2022_2_4	33	Maëlys;Nordahl Le...	2022_3_2	243	Kiev;DIRECT;Guerr...	0.2558578
2022_2_5	67	DIRECT;Guerre en ...	2022_3_5	68	Colonna;Corse	0.25396654
2022_2_6	49	Ehpad;Orpea	2022_3_0	636	Russie;Guerre en ...	0.2531
2022_2_6	49	Ehpad;Orpea	2022_3_3	169	Emmanuel Macron;P...	0.25037706
2022_2_4	33	Maëlys;Nordahl Le...	2022_3_4	62	Le Pen;Zemmour	0.24980427
2022_2_6	49	Ehpad;Orpea	2022_3_2	243	Kiev;DIRECT;Guerr...	0.24357249
2022_2_3	87	Zemmour;Le Pen;Pé...	2022_3_7	25	CARTE;Covid-19	0.24277303
2022_2_1	210	Russie;Guerre en ...	2022_3_7	25	CARTE;Covid-19	0.24159993
2022_2_2	86	JO;Pékin	2022_3_7	25	CARTE;Covid-19	0.23924696
2022_2_5	67	DIRECT;Guerre en ...	2022_3_4	62	Le Pen;Zemmour	0.23487024
2022_2_3	87	Zemmour;Le Pen;Pé...	2022_3_1	696	Ukraine;Poutine;M...	0.21720499
2022_2_4	33	Maëlys;Nordahl Le...	2022_3_6	51	Deschamps;Giroud;...	0.20277438
2022_2_6	49	Ehpad;Orpea	2022_3_6	51	Deschamps;Giroud;...	0.19811638
2022_2_6	49	Ehpad;Orpea	2022_3_7	25	CARTE;Covid-19	0.19122073
2022_2_0	423	Poutine;Ukraine;M...	2022_3_4	62	Le Pen;Zemmour	0.18773119
2022_2_6	49	Ehpad;Orpea	2022_3_4	62	Le Pen;Zemmour	0.16853927
2022_2_0	423	Poutine;Ukraine;M...	2022_3_5	68	Colonna;Corse	0.1667655
2022_2_3	87	Zemmour;Le Pen;Pé...	2022_3_2	243	Kiev;DIRECT;Guerr...	0.1363117
2022_2_2	86	JO;Pékin	2022_3_5	68	Colonna;Corse	0.13041048
2022_2_6	49	Ehpad;Orpea	2022_3_5	68	Colonna;Corse	0.11968931
2022_2_3	87	Zemmour;Le Pen;Pé...	2022_3_6	51	Deschamps;Giroud;...	0.11127678
2022_2_3	87	Zemmour;Le Pen;Pé...	2022_3_0	636	Russie;Guerre en ...	0.105463825
2022_2_1	210	Russie;Guerre en ...	2022_3_5	68	Colonna;Corse	0.0910217
2022_2_1	210	Russie;Guerre en ...	2022_3_4	62	Le Pen;Zemmour	0.08018912
2022_2_2	86	JO;Pékin	2022_3_4	62	Le Pen;Zemmour	0.04405083

E Groupes de communautés

```
Clustering with 285 elements and 60 clusters
[ 0] 2017_5_2, 2017_6_4, 2017_6_7, 2017_5_3, 2017_7_0, 2017_8_1, 2017_8_3,
     2017_9_4, 2017_9_8, 2017_10_3, 2017_10_4, 2017_11_4, 2017_11_5,
     2017_12_2, 2018_1_7, 2018_2_1, 2018_2_2, 2018_3_3, 2018_3_5, 2018_4_0,
     2018_4_7, 2018_5_6
[ 1] 2017_5_5, 2017_6_6
[ 2] 2017_6_1, 2017_7_1
[ 3] 2017_6_2, 2017_7_3
[ 4] 2017_7_4, 2017_8_7
[ 5] 2017_7_5, 2017_8_6
[ 6] 2017_8_0, 2017_9_7, 2017_10_0, 2017_11_8, 2017_12_7, 2018_1_1
[ 7] 2017_9_6, 2017_10_2
[ 8] 2017_10_5, 2017_11_2
[ 9] 2017_11_3, 2017_12_0, 2018_1_9
[10] 2017_12_3, 2018_1_6, 2018_2_0, 2018_3_0, 2018_4_2
[11] 2017_12_6, 2018_1_2
[12] 2018_3_7, 2018_4_6
[13] 2018_4_1, 2018_5_1, 2018_6_0, 2018_7_5
[14] 2018_4_4, 2018_5_3
[15] 2018_5_7, 2018_6_3, 2018_6_4, 2018_7_3
[16] 2018_6_1, 2018_7_1, 2018_7_4
[17] 2018_8_4, 2018_9_2, 2018_10_3, 2018_11_4
[18] 2018_10_2, 2018_11_3, 2018_12_3, 2019_1_3, 2019_2_8, 2019_3_3, 2019_4_6,
     2019_5_8, 2019_6_5, 2019_7_5, 2019_8_3, 2019_9_1, 2019_10_2
[19] 2018_10_5, 2018_11_2, 2018_12_1, 2019_1_2, 2019_2_4
[20] 2018_11_1, 2018_12_0, 2019_1_0, 2019_2_0, 2019_3_1, 2019_4_1, 2019_5_7
[21] 2018_11_6, 2018_12_6
[22] 2019_1_1, 2019_2_1, 2019_3_4, 2019_4_8, 2019_5_9
[23] 2019_1_6, 2019_2_3
[24] 2019_2_5, 2019_3_0, 2019_4_4
[25] 2019_4_3, 2019_5_3
[26] 2019_4_5, 2019_5_2, 2019_6_2, 2019_7_0
[27] 2019_5_1, 2019_6_0
[28] 2019_6_7, 2019_7_6
[29] 2019_7_1, 2019_8_2
[30] 2019_7_2, 2019_8_1, 2019_9_3, 2019_9_5, 2019_10_6, 2019_11_0, 2019_12_1,
     2020_1_5, 2020_2_1, 2020_2_6
[31] 2019_7_8, 2019_8_0
[32] 2019_9_0, 2019_10_1
[33] 2019_9_4, 2019_10_3, 2019_11_3
[34] 2019_10_5, 2019_11_1, 2019_10_8
[35] 2019_12_0, 2020_1_1
[36] 2020_1_0, 2020_2_0, 2020_1_2, 2020_3_0, 2020_4_0, 2020_4_1, 2020_4_2,
     2020_5_0, 2020_5_1, 2020_5_2, 2020_6_0, 2020_6_1, 2020_5_6, 2020_7_0,
     2020_7_1, 2020_8_2, 2020_8_5, 2020_9_0, 2020_9_1, 2020_9_5, 2020_10_0,
     2020_10_1, 2020_11_0, 2020_12_0, 2020_12_2, 2020_11_1, 2020_12_1,
     2021_1_1, 2021_1_0, 2021_2_0, 2021_3_0, 2021_3_2, 2021_4_0, 2021_4_1,
     2021_4_2, 2021_5_3, 2021_5_0, 2021_6_0, 2021_6_2, 2021_7_1, 2021_7_3,
     2021_7_9, 2021_6_5, 2021_7_4, 2021_8_1, 2021_8_3, 2021_7_2, 2021_8_0,
     2021_9_1, 2021_9_5, 2021_8_2, 2021_9_3, 2021_9_7, 2021_10_1, 2021_10_0,
     2021_10_3, 2021_10_2, 2021_11_2, 2021_11_7, 2021_11_0, 2021_11_1,
     2021_11_3, 2021_12_0, 2021_12_1, 2022_1_1, 2022_1_0
[37] 2020_1_8, 2020_2_2, 2020_3_1
[38] 2020_4_4, 2020_5_3
[39] 2020_5_4, 2020_6_2
[40] 2020_6_5, 2020_7_2, 2020_8_3, 2020_9_9, 2020_10_9
[41] 2020_8_1, 2020_9_4
[42] 2020_9_2, 2020_10_3
[43] 2020_9_3, 2020_10_5
[44] 2020_10_6, 2020_11_7
[45] 2021_1_2, 2021_2_4, 2021_3_5, 2021_3_1, 2021_4_4
[46] 2021_1_4, 2021_2_7
[47] 2021_3_4, 2021_4_5, 2021_4_3, 2021_5_6
[48] 2021_5_4, 2021_6_1
[49] 2021_5_7, 2021_6_3
[50] 2021_5_8, 2021_6_8, 2021_7_8
[51] 2021_5_9, 2021_6_7
[52] 2021_9_0, 2021_10_4
[53] 2021_9_9, 2021_10_5, 2021_11_8, 2021_12_4, 2022_1_3, 2022_2_3, 2022_3_4
[54] 2021_12_5, 2022_1_4, 2021_12_7
[55] 2022_1_2, 2022_2_0, 2022_2_1, 2022_2_5, 2022_3_0, 2022_3_1, 2022_3_2,
     2022_4_1, 2022_5_0, 2022_5_2, 2022_5_6, 2022_6_1, 2022_6_2
[56] 2022_3_3, 2022_4_0, 2022_4_2, 2022_4_4
[57] 2022_4_3, 2022_5_5
[58] 2022_4_5, 2022_5_1, 2022_6_3
[59] 2022_5_3, 2022_6_4, 2022_7_1
```

F Labels d'actualités #1

actualite	label_actu	date_mini	date_maxi	nb_titres
0	[Ligue, Football, PSG, DIRECT, Neymar, Mercato, Monaco]	2017-05-01	2018-05-26	2880
1	[Trump, FBI, Climat, États-Unis, Macron, Paris, Russie, Comey]	2017-05-01	2017-06-30	219
2	[Affaire Grégory, Murielle Bolle, Lambert, Jacob, Mans, Jacqueline Jacob, Grégory, Dijon, Marcel Jacob]	2017-06-14	2017-07-29	119
3	[Roland-Garros, Wimbledon, Tennis, Nadal, Mladenovic, Djokovic, Rafael Nadal]	2017-06-01	2017-07-24	202
4	[Venezuela, Maduro, Constituante, Caracas, Colombie, Trump, Nicolas Maduro, Macron]	2017-07-03	2017-08-31	154
5	[Macron, Paris, Trump, Marseille, Français, G20, Philippe]	2017-07-01	2017-08-31	677
6	[Catalogne, Barcelone, Espagne, Madrid, Attentats, Puigdemont, Carles Puigdemont]	2017-08-01	2018-01-31	522
7	[Paris, JO, Bonbonnes, Macron, Jeux, Bourse, Sentinelle, Berlin, Loup Bureau]	2017-09-01	2017-10-31	244
8	[Macron, Violences, Philippe, Guyane, Juppé, Wauquiez, Français]	2017-10-01	2017-11-30	364
9	[Trump, Jérusalem, États-Unis, Poutine, Macron, Israël, Palestiniens]	2017-11-01	2018-01-31	433
10	[Syrie, Macron, Ghouta, Russie, Trump, Douma, Erdogan]	2017-12-01	2018-04-30	824
11	[France, la France, Météo, Gall, Neige, Norvège, Macron]	2017-12-01	2018-01-31	148
12	[Brésil, Lula, Neymar, L'ex-président Lula]	2018-03-01	2018-04-28	58
13	[SNCF, TGV, TER, Grève SNCF, Montparnasse, Transilien, Pagaille]	2018-04-01	2018-07-31	455
14	[Gaza, Israël, Palestiniens, Jérusalem, Hamas, Palestinien, Cisjordanie]	2018-04-01	2018-05-31	154
15	[Trump, Donald Trump, Corée du Nord, États-Unis, Poutine, G7, Kim Jong-un]	2018-05-01	2018-07-31	463
16	[France, Coupe du monde 2018, Tour de France, Bleus, DIRECT, Mondial, Belgique]	2018-06-01	2018-07-31	749
17	[DIRECT, PSG, Ligue, Ligue 1, Football Leaks, Lyon, Monaco]	2018-08-04	2018-11-29	339
18	[Brexit, Boris Johnson, Theresa May, Londres, Royaume-Uni, Bruxelles, Britanniques, Parlement]	2018-10-01	2019-10-31	1030
19	[Macron, Trump, Grand débat, Gilets, Français, États-Unis, Philippe]	2018-10-01	2019-02-28	1073
20	[Gilets, Paris, France, DIRECT, Notre-Dame, Macron, Ligue]	2018-11-01	2019-05-31	3875
21	[Yémen, l'ONU, Hodeida, Sanaa, Suède, Houthis]	2018-11-01	2018-12-29	89
22	[Venezuela, Maduro, Guaido, Etats-Unis, Juan Guaido, Brésil, Colombie]	2019-01-03	2019-05-31	535
23	[Carlos Ghosn, Japon, Renault, Nissan, Ghosn]	2019-01-03	2019-02-21	74
24	[Algérie, Bouteflika, Alger, Algériens, Abdelaziz Bouteflika, Paris, Genève]	2019-02-02	2019-04-30	263
25	[Macron, Européennes, Europé, Français, RN, Grand débat, Corse]	2019-04-01	2019-05-31	672
26	[DIRECT, Tour de France, Ligue, PSG, Rennes, Coupe de France, Débarquement]	2019-04-01	2019-07-31	660
27	[Bac, Fiches, Fiche, Découvrez, Les]	2019-05-09	2019-06-25	118
28	[Wimbledon, Roland-Garros, Djokovic, Nadal, DIRECT, Federer, Thiem, Paire]	2019-06-01	2019-07-14	80
29	[Nantes, Steve, Disparition à, DIRECT, Ligue, Steve Caniço, Disparition de Steve]	2019-07-03	2019-08-31	171
30	[DIRECT, Ligue, PSG, France, Coronavirus, Neymar, Ligue 1]	2019-07-01	2020-02-29	1250
31	[Macron, G7, Amazonie, Emmanuel Macron, Brésil, Biarritz, Bolsonaro]	2019-07-01	2019-08-31	394
32	[Rouen, Lubrizol, Seveso, Incendie à, L’, É, DIRECT]	2019-09-07	2019-10-30	251
33	[Paris, Macron, Gilets, Municipales, Attaque, Extinction Rebellion, DIRECT, Tuerie à, Castaner]	2019-09-01	2019-11-30	722
34	[Bolivie, Chili, Morales, Evo Morales, Santiago, Mexique, Jeanine Añez, Cuba]	2019-10-02	2019-11-29	153
35	[SNCF, RATP, Grève, TGV, Grèves, DIRECT, Paris]	2019-12-01	2020-01-30	336
36	[Covid-19, Coronavirus, DIRECT, France, États-Unis, DIRECT - Covid-19, Paris]	2020-01-01	2022-01-31	36694
37	[Paris, Municipales, Coronavirus, Benjamin Griveaux, Agnès Buzyn, Anne Hidalgo, Buzyn]	2020-01-01	2020-03-31	578
38	[Dr Christophe Prudhomme, Le billet]	2020-04-01	2020-05-31	51
39	[Paris, Violences, Anne Hidalgo, Dati, Coronavirus, Hidalgo, Agnès Buzyn]	2020-05-01	2020-06-30	379
40	[PSG, Ligue des champions, Football, Neymar, Stade Rennais, Coupe de France, Ligue]	2020-06-01	2020-10-31	252
41	[Biélorussie, Loukachenko, Minsk, Poutine, Russie, Emmanuel Macron, Moscou, Tikhanovskaïa, Alexandre Loukachenko]	2020-08-07	2020-09-29	162
42	[Roland-Garros, Tennis, Djokovic, Nadal, Thiem, Novak Djokovic, Rome, Pliskova, Gaston, Wawrinka]	2020-09-07	2020-10-30	93
43	[Veolia, Suez, Rachat de Suez, Engie, Bruno Le Maire]	2020-09-02	2020-10-27	45
44	[Éthiopie, Tigré, Azerbaïdjan, Arménie, Nagorny Karabakh]	2020-10-02	2020-11-29	37
45	[États-Unis, Joe Biden, Trump, Donald Trump, Capitole, Biden, Johnson]	2021-01-01	2021-04-30	949
46	[Russie, Navalny, Poutine, Alexeï Navalny, Moscou, Covid-19, CEDH]	2021-01-05	2021-02-28	100
47	[Football, PSG, Stade Rennais, Mbappé, Benzema, Bayern, Football - Ligue des champions]	2021-03-01	2021-05-31	249
48	[Roland-Garros, Tennis, Federer, Nadal, Djokovic, Français, Wimbledon]	2021-05-09	2021-06-18	108
49	[Régionales, RN, Paca, LREM, LR, Hauts-de-France, Renaud Muselier]	2021-05-02	2021-06-30	309
50	[Emmanuel Macron, Macron, Polynésie, Rwanda, Covid-19, DIRECT, Drôme]	2021-05-01	2021-07-30	217
51	[Nantes, VIDÉO, Rennes, Policière, DIRECT, France, Tour de France]	2021-05-02	2021-06-30	91
52	[Macron, Marseille, Emmanuel Macron, Bernard Tapie, Présidentielle, Biden, DIRECT]	2021-09-01	2021-10-31	469
53	[Zemmour, Macron, Pécresse, Le Pen, Présidentielle, Mélenchon, LR, DIRECT]	2021-09-01	2022-03-31	566
54	[Présidentielle, Valérie Pécresse, Christiane Taubira, Anne Hidalgo, LR, DIRECT, Zemmour]	2021-12-01	2022-01-31	317
55	[Ukraine, Guerre en Ukraine, Russie, Macron, Kiev, France, Poutine]	2022-01-01	2022-06-30	3496
56	[Macron, Présidentielle, Emmanuel Macron, Marine Le Pen, Le Pen, DIRECT, Mélenchon]	2022-03-01	2022-04-30	897
57	[Twitter, Elon Musk, Musk]	2022-04-02	2022-05-28	43
58	[PSG, Mbappé, Ligue 1, Mercato, Galtier, Zidane, Real Madrid]	2022-04-02	2022-06-30	113
59	[Roland-Garros, Nadal, Wimbledon, Djokovic, Alcaraz, Kyrgios, Tennis]	2022-05-01	2022-07-10	133

G Labels d'actualités #2

actualite	label_actu	date_mini	date_maxi	nb_titres
2017_10_1	[Marseille, Attaque, Italie, Ligue]	2017-10-01	2017-10-31	68
2017_10_6	[NBA, Cleveland, Golden State, San Antonio, Boston]	2017-10-06	2017-10-31	32
2017_10_7	[Tennis, Nadal, Federer, Caroline Garcia, Masters]	2017-10-05	2017-10-31	63
2017_11_0	[Zimbabwe, Mugabe, Robert Mugabe, Emmerson Mnangagwa, Mnangagwa]	2017-11-14	2017-11-27	56
2017_11_1	[Liban, Saad Hariri, Emmanuel Macron, Paris, Macron]	2017-11-04	2017-11-30	46
2017_11_6	[Transat Jacques Vabre, Direct , Direct, Dick, Direct]	2017-11-04	2017-11-19	49
2017_11_7	[Rugby, de France, XV de France, France, Japon, Bleus]	2017-11-01	2017-11-29	59
2017_12_1	[Johnny Hallyday, Saint-Barthélemy, Emmanuel Macron, Paris, Champs-Élysées, [Direct], IMAGES]	2017-12-04	2017-12-30	135
2017_12_4	[TER, Lille, Millas, Ligue 1]	2017-12-02	2017-12-29	33
2017_12_5	[SNCF, Montparnasse, Panne, Transports, Trains]	2017-12-03	2017-12-30	54
2017_12_8	[Millas, Collision, TER, Accident, [Direct], SNCF]	2017-12-05	2017-12-30	73
2017_5_0	[Macron, Présidentielle, Emmanuel Macron, DIRECT, Le Pen]	2017-05-01	2017-05-31	654
2017_5_1	[Législatives, FN, Marine Le Pen, Valls, La France, PS]	2017-05-01	2017-05-31	325
2017_5_4	[Mélenchon, Marseille, Cazeneuve, Rémi Fraisse, Législatives]	2017-05-01	2017-05-31	63
2017_5_6	[NBA, Basket, Golden State, Houston, Boston, San Antonio, Cleveland]	2017-05-02	2017-05-30	33
2017_5_7	[Cyril Hanouna, Canular, CSA, Touche]	2017-05-19	2017-05-31	22
2017_6_0	[Législatives, Londres, DIRECT, Français, PS]	2017-06-01	2017-06-30	390
2017_6_3	[Theresa May, Royaume-Uni, Brexit, Corbyn]	2017-06-03	2017-06-29	49
2017_6_5	[Turquie, Mathias Depardon, Macron, Mercato]	2017-06-02	2017-06-24	30
2017_6_8	[Marche, La République, Législatives]	2017-06-01	2017-06-27	18
2017_7_2	[Tour de France, Froome, Bardet, DIRECT, Aru, Tour]	2017-07-01	2017-07-25	195
2017_7_6	[Simone Veil, Panthéon]	2017-07-01	2017-07-07	33
2017_7_7	[Rennes, LGV, Paris, TGV, Macron]	2017-07-01	2017-07-31	52
2017_8_10	[Tennis, ATP, Nadal, Montréal, Cincinnati]	2017-08-01	2017-08-30	34
2017_8_2	[Athlétisme, Mondiaux, Bolt, Vicaut, Kevin Mayer, Pierre-Ambroise Bosse]	2017-08-01	2017-08-31	86
2017_8_4	[Trump, Charlottesville, Corée du Nord, États-Unis, Tempête Harvey]	2017-08-01	2017-08-31	111
2017_8_5	[SNCF, Montparnasse]	2017-08-01	2017-08-21	29
2017_8_8	[Texas, Harvey, Louisiane, Tempête Harvey, Etats-Unis, L'ouragan Harvey]	2017-08-16	2017-08-31	36
2017_8_9	[Beauval, Zoo]	2017-08-01	2017-08-11	11
2017_9_0	[Irma, Saint-Martin, Ouragan Irma, Floride, Saint-Barth]	2017-09-04	2017-09-30	222
2017_9_1	[Rohingyas, Birmanie, Bangladesh, Aung San Suu Kyi, l'ONU, L'ONU]	2017-09-01	2017-09-29	49
2017_9_10	[Liliane Bettencourt, L'Oréal, Décès]	2017-09-01	2017-09-26	14
2017_9_11	[Wauquiez, LR, Pécresse]	2017-09-01	2017-09-30	34
2017_9_12	[Teddy Riner, Judo]	2017-09-01	2017-09-03	13
2017_9_13	[Trump, Corée du Nord, Pyongyang, Kim Jong-Un, Japon]	2017-09-01	2017-09-30	136
2017_9_2	[Turquie, Loup Bureau, France, Paris]	2017-09-01	2017-09-29	64
2017_9_3	[FN, Philippot, Le Pen, Marine Le Pen, Florian Philippot, Jean-Marie Le Pen]	2017-09-01	2017-09-29	50
2017_9_5	[DIRECT, Ouragan Maria, Guadeloupe, Martinique, Irma]	2017-09-03	2017-09-30	104
2017_9_9	[Mondiaux, Riner]	2017-09-02	2017-09-26	22
2018_10_0	[Brésil, Bolsonaro, Jair Bolsonaro, Lula, Haddad]	2018-10-02	2018-10-31	79
2018_10_1	[Trump, Khashoggi, Turquie, Macron, Saoudiens]	2018-10-01	2018-10-31	146
2018_10_4	[Corse, Collision, Méditerranée, Tempête Adrian, Alerte]	2018-10-06	2018-10-31	42
2018_10_6	[Clinton, Obama, CNN, États-Unis]	2018-10-24	2018-10-25	8
2018_11_0	[Marseille, Immeubles, Marseille , Gaudin, Marseille, Jean-Claude Gaudin, Logement, Montpellier]	2018-11-01	2018-11-30	131
2018_11_5	[Gabart, Joyon, DIRECT Route du Rhum, Route du rhum]	2018-11-06	2018-11-18	26
2018_11_7	[Pakistan, Bibi, Asia Bibi]	2018-11-01	2018-11-30	23
2018_12_2	[Bush, George HW, Washington]	2018-12-01	2018-12-20	20
2018_12_4	[Indonésie, Tsunami]	2018-12-04	2018-12-28	32
2018_12_5	[CGT, FO, Routiers]	2018-12-01	2018-12-18	24
2018_1_0	[Mathieu Gallet, CSA, Radio France]	2018-01-03	2018-01-31	31
2018_1_3	[Paris, la Seine, Bourse, RER C, Ritz]	2018-01-01	2018-01-31	131
2018_1_4	[Corse, Simeoni, Talamoni, Jacqueline Gourault, Gilles Simeoni]	2018-01-02	2018-01-25	47
2018_1_5	[Afghanistan, Attaque, Kaboul, Save the Children, Jalalabad, Daech]	2018-01-02	2018-01-31	21
2018_1_8	[Lactalis, Craon, Bruno Le Maire, Leclerc, Intermarché]	2018-01-03	2018-01-31	105
2018_2_3	[Nordahl Lelandais, Maëlys, Affaire Maëlys, RTT, Lelandais]	2018-02-01	2018-02-27	66
2018_2_4	[Jean-Marie Le Pen, FN, Front national]	2018-02-07	2018-02-28	28
2018_2_5	[TF1, Orange]	2018-02-01	2018-02-27	14
2018_2_7	[Macron, Corse, Syrie, Merkel, Poutine, Wauquiez]	2018-02-01	2018-02-28	132
2018_2_8	[Venezuela, Maduro]	2018-02-02	2018-02-27	16
2018_3_1	[Russie, Moscou, Londres, Ex-espion, Affaire Skripal]	2018-03-01	2018-03-31	200
2018_3_10	[TF1, Canal+, Canal, Orange, Nicolas Sarkozy]	2018-03-01	2018-03-29	38
2018_3_11	[Macron, Merkel, Affaire Skripal, Moscou, Philippe]	2018-03-01	2018-03-29	103
2018_3_2	[Attaques, Arnaud Beltrame, Radouane Lakdim, Trèbes, Aude]	2018-03-02	2018-03-29	104
2018_3_4	[France, 6 Nations, Bleus, Cardiff, Crunch]	2018-03-01	2018-03-31	76
2018_3_6	[Carlos, Attentat du Drugstore]	2018-03-05	2018-03-15	13
2018_3_8	[Paris, Meurtre, Bourse, Mireille Knoll, Londres]	2018-03-01	2018-03-31	145
2018_3_9	[Trump, États-Unis, Stormy Daniels, Macron, Donald Trump, Affaire Skripal, Pékin, Pyongyang, Commerce, Miami]	2018-03-01	2018-03-31	116
2018_4_3	[Trèbes, Super U, Attentats, Attaques]	2018-04-01	2018-04-16	16
2018_4_5	[Cyclisme, Tour de Bretagne]	2018-04-01	2018-04-30	21
2018_4_8	[Emmanuel Macron, Donald Trump, Syrie, TF1, DIRECT]	2018-04-01	2018-04-30	89
2018_5_0	[Paris, Attaque, Bourse, Migrants, Collomb]	2018-05-01	2018-05-31	207
2018_5_2	[Harry, Meghan, Meghan Markle, DIRECT]	2018-05-12	2018-05-22	26
2018_5_4	[Zad, Notre-Dame-des-Landes]	2018-05-06	2018-05-31	21
2018_5_5	[Italie, Giuseppe Conte, Mouvement 5]	2018-05-03	2018-05-31	51
2018_5_8	[RDC, Ebola]	2018-05-03	2018-05-28	19
2018_6_2	[Turquie, Erdogan, Tunisie, Méditerranée]	2018-06-03	2018-06-27	49
2018_6_5	[Macron, Migrants, Lifeline, Malte, G7]	2018-06-04	2018-06-30	219
2018_6_6	[Montpellier, Castres, Top, Toulon, de France]	2018-06-01	2018-06-25	16
2018_7_0	[Affaire Benalla, Macron, l'Assemblée, Congrès, Collomb]	2018-07-01	2018-07-31	274
2018_7_10	[Canada, Toronto, Québec]	2018-07-04	2018-07-30	26
2018_7_2	[Thaïlande, Grotte, Les, Phuket]	2018-07-02	2018-07-23	70

2018_7_6	[Japon, Aum]	2018-07-02	2018-07-29	53
2018_7_7	[Simone Veil, Panthéon]	2018-07-01	2018-07-28	14
2018_7_8	[Brexit, Theresa May, Royaume-Uni, Boris Johnson, Londres]	2018-07-03	2018-07-31	39
2018_7_9	[Nantes, Breil, CRS, Jeune, «La]	2018-07-01	2018-07-31	77
2018_8_0	[Booba, Kaaris, Orly, Bagarre, Orly Booba]	2018-08-01	2018-08-23	31
2018_8_1	[Venezuela, Maduro]	2018-08-01	2018-08-31	42
2018_8_2	[Marseille, TGV, McDonald's, Ligue]	2018-08-02	2018-08-31	43
2018_8_3	[Israël, Gaza, Hamas]	2018-08-01	2018-08-29	34
2018_8_5	[Indonésie, Lombok]	2018-08-05	2018-08-28	42
2018_9_0	[Benalla, Sénat, Macron, Kavanaugh, Affaire Benalla, Poitiers]	2018-09-04	2018-09-28	67
2018_9_1	[Martinique, Guadeloupe, Tempête Isaac, Isaac, Macron]	2018-09-12	2018-09-29	15
2019_10_0	[Syrie, Turquie, Trump, France, Kurdes]	2019-10-01	2019-10-31	484
2019_10_10	[SNCF, TGV Atlantique, TER, Toulon, Bretagne, Marseille, TGV]	2019-10-02	2019-10-31	105
2019_10_4	[Japon, Hagibis, DIRECT, Samoa, Coupe du monde de rugby]	2019-10-05	2019-10-31	39
2019_10_7	[Lilas, Lycéen, Seine-Saint-Denis]	2019-10-04	2019-10-13	14
2019_10_9	[L’, Assemblé, Lubrizol, Sé, Ligonnè, France]	2019-10-01	2019-10-31	41
2019_11_2	[Brésil, Lula, Bolsonaro, Marée]	2019-11-01	2019-11-18	36
2019_11_4	[Trump, Destitution, Congrès, États-Unis, Donald Trump]	2019-11-01	2019-11-28	77
2019_11_5	[Français, Haïti, Mexique]	2019-11-01	2019-11-30	46
2019_11_6	[Michel Fourniret, Disparition d'Estelle Mouzin, Monique Olivier]	2019-11-13	2019-11-29	16
2019_12_2	[Boris Johnson, Royaume-Uni, Brexit, Johnson, Jeremy Corbyn, Grande-Bretagne, Londres]	2019-12-01	2019-12-30	66
2019_12_3	[COP25, Climat, Madrid]	2019-12-01	2019-12-27	40
2019_12_4	[Willy Bardon, Affaire Kulik, Affaire Kulik]	2019-12-06	2019-12-12	16
2019_12_5	[Fabien, Tempê, Météo]	2019-12-21	2019-12-26	20
2019_12_6	[Liban, Carlos Ghosn, Japon, Beyrouth]	2019-12-02	2019-12-31	39
2019_12_7	[Mogadiscio, Somalie]	2019-12-10	2019-12-30	11
2019_12_8	[Algérie, Tebboune, Abdelaziz Djerad]	2019-12-03	2019-12-30	51
2019_1_4	[Mondial 2019, Bleus, Handball, Russie, Mondial]	2019-01-02	2019-01-28	47
2019_1_5	[Brésil, Rupture, Bolsonaro, Jair Bolsonaro]	2019-01-01	2019-01-31	46
2019_2_2	[PSG, Ligue, DIRECT, Mbappé, Manchester United]	2019-02-01	2019-02-28	79
2019_2_6	[Alexandre Benalla, Vincent Crase, Mediapart]	2019-02-17	2019-02-26	18
2019_2_7	[Benalla, Crase]	2019-02-05	2019-02-28	38
2019_3_2	[Européennes, LREM, PS, Nathalie Loiseau, Raphaël Glucksmann]	2019-03-01	2019-03-31	112
2019_3_5	[Crash, Boeing, Ethiopie, Airlines, 737 MAX]	2019-03-10	2019-03-29	63
2019_3_6	[Mozambique, Zimbabwe, Cyclone Idai, Cyclone, Idai]	2019-03-05	2019-03-29	27
2019_3_7	[Trump, Golan, États-Unis, Sénat, Kim]	2019-03-01	2019-03-31	76
2019_4_0	[Sri Lanka, Attentats, Colombo, Pâques, Zahran Hashim]	2019-04-21	2019-04-30	107
2019_4_10	[Kim Jong-un, Poutine, Corée du Nord, Vladivostok, Trump]	2019-04-02	2019-04-26	28
2019_4_11	[Ukraine, Zelensky, Volodymyr Zelensky]	2019-04-01	2019-04-27	23
2019_4_2	[Libye, Tripoli, Haftar, l'ONU, maréchal Haftar, Paris]	2019-04-03	2019-04-28	66
2019_4_7	[Carlos Ghosn, Japon, Tokyo, Renault, Nissan]	2019-04-01	2019-04-30	69
2019_4_9	[Tarn-et-Garonne, Chute]	2019-04-18	2019-04-23	13
2019_5_0	[Lyon, Explosion, Lyon, Attaque à, DIRECT]	2019-05-03	2019-05-31	100
2019_5_10	[Yémen, Armes, Houthis]	2019-05-02	2019-05-29	25
2019_5_5	[Israël, Gaza, Palestinien]	2019-05-02	2019-05-30	36
2019_5_6	[L’, Cannes, É, Tour Eiffel, Balkany]	2019-05-08	2019-05-28	54
2019_6_1	[Olonne, Sables-d’, Sables-d'Olonne, Macron, SNSM]	2019-06-07	2019-06-21	27
2019_6_10	[Trump, Macron, États-Unis, G20, Poutine, Marseille, Kim, Elton John]	2019-06-01	2019-06-30	172
2019_6_3	[Renault, Nissan, Fiat Chrysler]	2019-06-03	2019-06-27	25
2019_6_4	[France, Canicule, Gard, Paris, Allemagne, la France]	2019-06-01	2019-06-30	205
2019_6_6	[Belfort, General Electric, GE, Bruno Le Maire]	2019-06-02	2019-06-30	24
2019_6_8	[-sur-Sarthe, Condé]	2019-06-11	2019-06-13	10
2019_6_9	[Sea-Watch, Lampedusa, Migrants]	2019-06-01	2019-06-30	15
2019_7_10	[Manche, Franky Zapata, la Manche, Traversé]	2019-07-10	2019-07-30	17
2019_7_11	[Barry, Louisiane, États-Unis, Tempête]	2019-07-11	2019-07-14	11
2019_7_3	[Rugy, Assemblé, Franç, L’, Haine, Dî, LREM]	2019-07-01	2019-07-31	81
2019_7_4	[Paris, LREM, Municipales, Marseille, Benjamin Griveaux]	2019-07-01	2019-07-31	204
2019_7_7	[Afghanistan, Kaboul]	2019-07-01	2019-07-31	31
2019_7_9	[Greta Thunberg, Climat]	2019-07-01	2019-07-30	26
2019_8_10	[Strasbourg, Ligue Europa, DIRECT]	2019-08-01	2019-08-31	23
2019_8_11	[Norvège, Fusillade, Attaque]	2019-08-05	2019-08-31	21
2019_8_12	[Espagne, Grande Canarie, Français, Benidorm]	2019-08-01	2019-08-30	39
2019_8_4	[États-Unis, Trump, Donald Trump, G7, la Chine, Groenland, Paso, Etats-Unis]	2019-08-01	2019-08-31	220
2019_8_5	[Cachemire, Pakistan, Inde, L'Inde]	2019-08-02	2019-08-30	47
2019_8_6	[New York, Greta Thunberg, Climat]	2019-08-10	2019-08-30	29
2019_8_7	[Italie, Simon Gautier, Matteo Salvini, M5S, Français, Disparition]	2019-08-04	2019-08-30	72
2019_8_8	[Russie, Moscou, G8, Kremlin, Poutine, Canada, L'opposition, France]	2019-08-01	2019-08-30	75
2019_8_9	[Migrants, Arms, Viking, Lampedusa]	2019-08-02	2019-08-30	29
2019_9_10	[Israël, Hezbollah, Gantz, Netanyahu, Cisjordanie]	2019-09-01	2019-09-27	46
2019_9_2	[Bahamas, Dorian, Ouragan Dorian, L'ouragan Dorian, États-Unis]	2019-09-01	2019-09-15	62
2019_9_6	[Jacques Chirac, Invalides, Français, Marine Le Pen, Bernadette Chirac]	2019-09-26	2019-09-30	115
2019_9_7	[US Open, Nadal, Medvedev, Monfils]	2019-09-01	2019-09-09	20
2019_9_8	[Ferrand, Mutuelles de Bretagne, Richard Ferrand]	2019-09-11	2019-09-13	16
2019_9_9	[Aigle Azur, Air France, Dubreuil, Crash du Rio-Paris, Airbus]	2019-09-02	2019-09-30	38
2020_10_10	[Turquie, Méditerranée, la France, Karabakh, France]	2020-10-01	2020-10-30	48
2020_10_11	[Nigeria, Lagos]	2020-10-11	2020-10-31	15
2020_10_2	[Alpes-Maritimes, Tempête Alex, Intempéries, Morbihan, DIRECT]	2020-10-01	2020-10-31	117
2020_10_4	[Mali, Sophie Pétronin, France, Bamako, Macron]	2020-10-01	2020-10-28	43
2020_10_8	[Affaire Estelle Mouzin, Michel Fourniret, Ardennes, Monique Olivier]	2020-10-01	2020-10-30	10
2020_11_4	[Vienne, Autriche]	2020-11-02	2020-11-06	20
2020_11_5	[Football, Maradona, Diego Maradona, Giroud, Naples, Dembélé, Deschamps, Messi, Isco]	2020-11-01	2020-11-30	76
2020_11_6	[Paris, Producteur, DIRECT, Paris , France, Gérald Darmanin, Londres, Covid-19]	2020-11-01	2020-11-30	169
2020_11_8	[Maurice Genevoix, Panthéon, Macron]	2020-11-09	2020-11-11	13
2020_12_3	[Marseille, Benoît Payan, Michèle Rubirola, Ligue, Printemps marseillais?, Nîmes, Rubirola, Lille, Football - Ligue 1]	2020-12-01	2020-12-30	64
2020_1_10	[Carlos Ghosn, Japon, Nissan, Liban, Coronavirus]	2020-01-01	2020-01-31	83
2020_1_11	[Brest, Coup, Lorient]	2020-01-02	2020-01-30	28
2020_1_12	[Libye, Berlin, Paris, Erdogan, Parlement, Alger, Londres, Moscou, Poutine, Tripoli, Turquie]	2020-01-02	2020-01-29	56
2020_1_3	[Iran, Boeing, Téhéran, Crash, Donald Trump]	2020-01-01	2020-01-29	120
2020_1_6	[Harry, Meghan, Canada, Meghan Markle]	2020-01-02	2020-01-28	19
2020_1_7	[Trump, États-Unis, Soleimani, Sénat, Iran]	2020-01-02	2020-01-31	169
2020_1_9	[Davos, Greta Thunberg, Donald Trump, Trump]	2020-01-09	2020-01-31	24
2020_2_3	[Retraites , Assemblé, L’, LFI]	2020-02-03	2020-02-29	29

2020_2_4	[Piotr Pavlenski, Affaire Griveaux, Alexandra de Taddeo, Juan Branco, Paris]	2020-02-14	2020-02-23	28
2020_2_5	[Angers, Vanille, Alerte, Brest, Ligue 1, DIRECT]	2020-02-01	2020-02-29	29
2020_2_7	[César, Roman Polanski]	2020-02-09	2020-02-29	24
2020_2_8	[Violences, Didier Gailhaguet, Gailhaguet, Sarah Abitbol, Maracineanu, Tariq Ramadan, Beyer]	2020-02-01	2020-02-27	56
2020_2_9	[Essonne, LBD]	2020-02-19	2020-02-28	12
2020_3_2	[Syrie, Erdogan, Idleb, Turquie, Poutine]	2020-03-01	2020-03-31	46
2020_3_3	[Biathlon, Nove Mesto, DIRECT, Johannes Boe]	2020-03-05	2020-03-14	11
2020_4_3	[Philippe, É, Coronavirus, DIRECT, Macron]	2020-04-01	2020-04-29	39
2020_4_5	[Google, Apple]	2020-04-06	2020-04-27	17
2020_5_5	[Déconfinement, Les, Coronavirus, DIRECT, REPORTAGE, Français, Ouest-France, Paris]	2020-05-01	2020-05-31	165
2020_6_3	[Municipales, Marseille, Lyon, LR, Paris]	2020-06-01	2020-06-30	187
2020_6_4	[Macron, Convention, Français, Poutine, la France, Emmanuel Macron, Philippe, Merkel]	2020-06-01	2020-06-30	114
2020_7_3	[Sainte-Sophie, Turquie, Erdogan]	2020-07-01	2020-07-31	42
2020_7_4	[Jean Castex, Matignon, DIRECT, Philippe, Français, Édouard Philippe]	2020-07-01	2020-07-29	101
2020_7_5	[Nokia, Lannion, Paris, Sanofi]	2020-07-01	2020-07-27	19
2020_7_6	[Washington, Pékin, Chine, Hong Kong, Hongkong, États-Unis, Coronavirus]	2020-07-01	2020-07-31	63
2020_8_0	[Beyrouth, Liban, Explosions, Macron, Français]	2020-08-04	2020-08-31	249
2020_8_7	[Maurice, Marée]	2020-08-08	2020-08-29	23
2020_9_10	[Liban, Beyrouth, Macron, Hezbollah, REPORTAGE, Emmanuel Macron]	2020-09-01	2020-09-29	50
2020_9_6	[Lubrizol, Rouen]	2020-09-10	2020-09-30	21
2020_9_7	[Lesbos, Moria, Grèce]	2020-09-07	2020-09-27	41
2021_10_6	[Châteaubriant, Pierre-Louis Basse, Mémoire]	2021-10-07	2021-10-23	16
2021_10_7	[Yémen, Marib, Houthis]	2021-10-03	2021-10-30	18
2021_11_10	[Mireille Knoll, Yacine Mihoub, Meurtre]	2021-11-02	2021-11-17	11
2021_11_11	[Formule, Hamilton, Verstappen, Bottas, Formule 1]	2021-11-06	2021-11-21	27
2021_11_12	[Tesla, Elon Musk, Twitter]	2021-11-02	2021-11-20	10
2021_11_13	[Bleus, Kazakhstan, France, Rugby, Benzema, All Blacks, Qatar]	2021-11-03	2021-11-28	41
2021_11_14	[Mayenne, Joggeuse, Sarthe, Mayenne , Disparition]	2021-11-08	2021-11-26	37
2021_11_15	[Transat Jacques Vabre, Ultim, DIRECT, Cammas, Pot-au-Noir]	2021-11-01	2021-11-28	67
2021_11_16	[Joséphine Baker, Panthéon]	2021-11-16	2021-11-30	18
2021_11_17	[Mali, Sophie Pétronin, Mali]	2021-11-01	2021-11-27	23
2021_11_4	[Cannes, Policiers]	2021-11-08	2021-11-17	22
2021_11_5	[Guadeloupe, Martinique, DIRECT, DIRECT - Covid-19, Covid-19, Sébastien Lecornu]	2021-11-03	2021-11-30	98
2021_11_6	[Thomas Pesquet, Terre, l'ISS, DIRECT VIDÉO]	2021-11-06	2021-11-18	25
2021_11_9	[Rugby, All Blacks, XV de France, Bleus, Nouvelle-Zélande]	2021-11-02	2021-11-30	53
2021_12_10	[Bleues, Handball, Mondial]	2021-12-02	2021-12-26	25
2021_12_2	[Formule, Hamilton, Verstappen, F1, Toto Wolff, Grand Prix d'Arabie Saoudite, GP d'Arabie Saoudite]	2021-12-03	2021-12-21	40
2021_12_3	[Philippines, Rai]	2021-12-04	2021-12-27	13
2021_12_6	[Chili, Gabriel Boric, Boric]	2021-12-02	2021-12-24	30
2021_12_8	[Allemagne, Olaf Scholz, Bundestag, Moscou]	2021-12-01	2021-12-26	37
2021_12_9	[États-Unis, Joe Biden, Omicron, Biden, France, Covid-19, Russie]	2021-12-01	2021-12-30	133
2021_1_3	[EXCLUSIF, Arnaud Démare, Dominic Thiem]	2021-01-01	2021-01-29	18
2021_1_5	[Carrefour, Couche-Tard, Paris]	2021-01-08	2021-01-30	20
2021_1_6	[PSG, Mauricio Pochettino, Mbappé, Leonardo, Suivez, Football]	2021-01-01	2021-01-29	40
2021_1_7	[Rave, Ille-et-Vilaine, Rennes, Lieuron]	2021-01-01	2021-01-23	21
2021_2_1	[Mars, Perseverance, Nasa]	2021-02-05	2021-02-28	28
2021_2_2	[Dunkerque, Olivier Véran, DIRECT, Covid-19, Nice]	2021-02-04	2021-02-27	48
2021_2_3	[Veolia, Suez]	2021-02-07	2021-02-26	12
2021_2_5	[Italie, Mario Draghi, Conte]	2021-02-02	2021-02-27	34
2021_2_6	[Darmanin, Le Pen]	2021-02-02	2021-02-24	18
2021_3_3	[Futuna, Wallis, Covid-19]	2021-03-07	2021-03-21	10
2021_3_6	[Basket, Samuel Nadeau, NBA, Real Madrid]	2021-03-09	2021-03-25	11
2021_3_7	[Pfizer, Moderna, Covid-19, AstraZeneca]	2021-03-01	2021-03-31	21
2021_3_8	[Lyon, Violences, DIRECT, PSG, Football - Ligue 1, Suisse, Ligue, Beauvais, Football - PSG]	2021-03-01	2021-03-31	106
2021_3_9	[Meghan, Harry, Elizabeth II]	2021-03-07	2021-03-09	10
2021_5_1	[Israël, Gaza, Hamas, Netanyahou, Netanyahu, Jérusalem, Palestiniens, Conflit]	2021-05-02	2021-05-31	120
2021_5_10	[Manchester City, Chelsea, Ligue des champions, PSG, Mbappé]	2021-05-01	2021-05-30	33
2021_5_2	[Cévennes, Double, Plantiers]	2021-05-11	2021-05-20	33
2021_5_5	[Avignon, Policier, Éric Masson]	2021-05-05	2021-05-29	29
2021_6_6	[Biden, Poutine, États-Unis, Genève, Kremlin]	2021-06-01	2021-06-30	47
2021_7_0	[Off, Avignon]	2021-07-04	2021-07-27	80
2021_7_5	[Brésil, Bolsonaro]	2021-07-01	2021-07-31	20
2021_7_6	[Floride, Immeuble, Elsa]	2021-07-01	2021-07-27	15
2021_8_4	[Joséphine Baker, Panthéon]	2021-08-22	2021-08-23	12
2021_9_10	[Verstappen, Hamilton, DIRECT, Formule 1 - GP, Russie]	2021-09-05	2021-09-26	13
2021_9_4	[Yannick Jadot, Sandrine Rousseau, Gérald Darmanin, Présidentielle]	2021-09-06	2021-09-30	33
2021_9_6	[Présidentielle, Anne Hidalgo, Marine Le Pen, Éric Zemmour, LR, Valérie Pécresse, Emmanuel Macron, PS]	2021-09-01	2021-09-30	127
2021_9_8	[Jean-Paul Belmondo, Invalides, DIRECT - Mort]	2021-09-06	2021-09-28	35
2022_1_5	[PSG, Mbappé, Ligue 1, Messi, Coupe de France]	2022-01-02	2022-01-31	36
2022_1_6	[Dakar, Loeb, Al-Attiyah, Sébastien Loeb, Audi]	2022-01-01	2022-01-31	30
2022_1_7	[Djokovic, Tennis, Melbourne, Australie, Open d'Australie, Monfils, Novak Djokovic, Dubaï]	2022-01-04	2022-01-30	51
2022_1_8	[Jean-Michel Blanquer, Ibiza, DIRECT]	2022-01-02	2022-01-26	25
2022_2_2	[JO, Pékin, la France, Fillon Maillet, Guillaume Cizeron, Gabriella Papadakis]	2022-02-01	2022-02-21	74
2022_2_4	[Nordahl Lelandais, Maëlys, Lelandais, Procès Maëlys]	2022-02-01	2022-02-18	24
2022_2_6	[Ehpad, Orpea, Korian, Maltraitance]	2022-02-01	2022-02-22	42
2022_3_5	[Colonna, Corse, Castex, Yvan Colonna]	2022-03-03	2022-03-31	60
2022_3_6	[Bleus, Giroud, Deschamps, Mbappé, Nkunku]	2022-03-05	2022-03-30	37
2022_3_7	[Covid-19 , CARTE, CARTES]	2022-03-01	2022-03-30	18
2022_5_4	[PS, LFI, PCF, DIRECT, EELV]	2022-05-01	2022-05-22	38
2022_6_0	[Législatives, Nupes, RN, DIRECT, Macron]	2022-06-01	2022-06-30	228
2022_6_5	[DIRECT, Canicule , Élisabeth Borne, Procès du 13-Novembre , RN, Législatives]	2022-06-07	2022-06-30	63
2022_7_0	[Tour de France, Pogacar, Covid-19, Wout Van Aert, Vingegaard, Van Aert, Étape]	2022-07-01	2022-07-14	72

H Statistiques des flux en fonction des sentiments

summary	nb_mots	nb_char	nb_char_maj	nb_chiffres	source	id	sentiment
count	84910	84910	84910	84910	Ouest-France	24	Negatif
count	55679	55679	55679	55679	Ouest-France	24	Positif
count	27167	27167	27167	27167	Le Telegramme	25	Negatif
count	15459	15459	15459	15459	Le Telegramme	25	Positif
count	11075	11075	11075	11075	Mediapart	30	Negatif
count	7429	7429	7429	7429	Mediapart	30	Positif
count	22564	22564	22564	22564	L'Humanité	34	Negatif
count	19862	19862	19862	19862	L'Humanité	34	Positif
count	97514	97514	97514	97514	Le figaro	50	Negatif
count	71471	71471	71471	71471	Le figaro	50	Positif
mean	12.266705923919444	81.72334236250147	3.8879048404192673	0.7605935696619951	Ouest-France	24	Negatif
mean	12.097882505073725	79.10661111011333	4.562815424127589	0.8570735824996857	Ouest-France	24	Positif
mean	10.48643574925461	75.56340413001068	3.141568815106563	0.7416718813266094	Le Telegramme	25	Negatif
mean	10.135196325764927	74.44064945986158	3.1576427970761367	0.9455980335079889	Le Telegramme	25	Positif
mean	10.251918735891648	69.52659142211219	2.5283069977426638	0.19214446952595937	Mediapart	30	Negatif
mean	9.383497105936195	63.469376766725	2.543545564678961	0.20393054246870373	Mediapart	30	Positif
mean	9.428691721326006	64.66335756071618	2.7875819890090408	0.29551497961354367	L'Humanité	34	Negatif
mean	9.070284966267243	61.52109555935958	2.8928607390997887	0.33843520290001006	L'Humanité	34	Positif
mean	11.062903788174006	71.6304530631499	2.7162048526365443	0.7248907849129356	Le figaro	50	Negatif
mean	10.584083054665529	67.64673783772439	3.075695037147934	0.8075163352968336	Le figaro	50	Positif

source	summary	taux_positifs__nb_mots	taux_positifs__nb_char	taux_positifs__nb_char_maj	taux_positifs__nb_chiffres
Ouest-France	count	0.39604094203671697	0.39604094203671697	0.39604094203671697	0.39604094203671697
Le Telegramme	count	0.3626659785107681	0.3626659785107681	0.3626659785107681	0.3626659785107681
Mediapart	count	0.4014807609165586	0.4014807609165586	0.4014807609165586	0.4014807609165586
L'Humanité	count	0.468156319238203	0.468156319238203	0.468156319238203	0.468156319238203
Le figaro	count	0.42294286475130927	0.42294286475130927	0.42294286475130927	0.42294286475130927

source	summary	taux_positifs__nb_mots	taux_positifs__nb_char	taux_positifs__nb_char_maj	taux_positifs__nb_chiffres
Ouest-France	mean	0.4965354756691719	0.49186491323323767	0.5399321337460288	0.5298207244638568
Le Telegramme	mean	0.49148371423241605	0.49625758556759125	0.5012758725185932	0.5604308031537948
Mediapart	mean	0.47788634483346365	0.47722782600899605	0.5015022683278693	0.5148785868658929
L'Humanité	mean	0.4903127951045223	0.4875489336298567	0.5092667733071043	0.5338514164601166
Le figaro	mean	0.4889402451947518	0.4856986086535131	0.531033874147708	0.526959399135586

I Part des fluxRSS dans les actualités

link	id_feed	taux_nombre_titres
http://www.ouest-france.fr/rss.xml	24	34.03
http://www.letelegramme.fr/france/rss.xml	25	10.32
http://www.mediapart.fr/articles/feed	30	4.48
http://www.humanite.fr/rss/actu.rss	34	10.27
http://www.lefigaro.fr/rss/figaro_flash-actu.xml	50	40.9

actualite	id_feed	label	date_mini	date_maxi	nb_titres_actu	nb_titres_actu_fluxRSS	taux_FluxRSS_Actu
36	24	Covid-19;Coronavi...	2020-01-01	2022-01-31	36694	19849	54.09
36	25	Covid-19;Coronavi...	2020-01-01	2022-01-31	36694	4332	11.81
36	30	Covid-19;Coronavi...	2020-01-01	2022-01-31	36694	503	1.37
36	34	Covid-19;Coronavi...	2020-01-01	2022-01-31	36694	1599	4.36
36	50	Covid-19;Coronavi...	2020-01-01	2022-01-31	36694	10411	28.37
20	24	Gilets;Paris;Fran...	2018-11-01	2019-05-31	3875	1842	47.54
20	25	Gilets;Paris;Fran...	2018-11-01	2019-05-31	3875	890	22.97
20	30	Gilets;Paris;Fran...	2018-11-01	2019-05-31	3875	76	1.96
20	34	Gilets;Paris;Fran...	2018-11-01	2019-05-31	3875	86	2.22
20	50	Gilets;Paris;Fran...	2018-11-01	2019-05-31	3875	981	25.32
55	24	Ukraine;Guerre en...	2022-01-01	2022-06-30	3496	542	15.5
55	25	Ukraine;Guerre en...	2022-01-01	2022-06-30	3496	180	5.15
55	30	Ukraine;Guerre en...	2022-01-01	2022-06-30	3496	203	5.81
55	34	Ukraine;Guerre en...	2022-01-01	2022-06-30	3496	327	9.35
55	50	Ukraine;Guerre en...	2022-01-01	2022-06-30	3496	2244	64.19
0	24	Ligue;Football;PS...	2017-05-01	2018-05-26	2880	2164	75.14
0	25	Ligue;Football;PS...	2017-05-01	2018-05-26	2880	48	1.67
0	30	Ligue;Football;PS...	2017-05-01	2018-05-26	2880	6	0.21
0	34	Ligue;Football;PS...	2017-05-01	2018-05-26	2880	20	0.69
0	50	Ligue;Football;PS...	2017-05-01	2018-05-26	2880	642	22.29
30	24	DIRECT;Ligue;PSG;...	2019-07-01	2020-02-29	1250	992	79.36
30	25	DIRECT;Ligue;PSG;...	2019-07-01	2020-02-29	1250	32	2.56
30	30	DIRECT;Ligue;PSG;...	2019-07-01	2020-02-29	1250	3	0.24
30	34	DIRECT;Ligue;PSG;...	2019-07-01	2020-02-29	1250	4	0.32
30	50	DIRECT;Ligue;PSG;...	2019-07-01	2020-02-29	1250	219	17.52
19	24	Macron;Trump;Gran...	2018-10-01	2019-02-28	1073	314	29.26
19	25	Macron;Trump;Gran...	2018-10-01	2019-02-28	1073	246	22.93
19	30	Macron;Trump;Gran...	2018-10-01	2019-02-28	1073	54	5.03
19	34	Macron;Trump;Gran...	2018-10-01	2019-02-28	1073	47	4.38
19	50	Macron;Trump;Gran...	2018-10-01	2019-02-28	1073	412	38.4
18	24	Brexit;Boris John...	2018-10-01	2019-10-31	1030	461	44.76
18	25	Brexit;Boris John...	2018-10-01	2019-10-31	1030	24	2.33
18	30	Brexit;Boris John...	2018-10-01	2019-10-31	1030	42	4.08
18	34	Brexit;Boris John...	2018-10-01	2019-10-31	1030	17	1.65
18	50	Brexit;Boris John...	2018-10-01	2019-10-31	1030	486	47.18
45	24	États-Unis;Joe Bi...	2021-01-01	2021-04-30	949	374	39.41
45	25	États-Unis;Joe Bi...	2021-01-01	2021-04-30	949	25	2.63
45	30	États-Unis;Joe Bi...	2021-01-01	2021-04-30	949	16	1.69
45	34	États-Unis;Joe Bi...	2021-01-01	2021-04-30	949	91	9.59
45	50	États-Unis;Joe Bi...	2021-01-01	2021-04-30	949	443	46.68
56	24	Macron;Présidenti...	2022-03-01	2022-04-30	897	343	38.24
56	25	Macron;Présidenti...	2022-03-01	2022-04-30	897	142	15.83
56	30	Macron;Présidenti...	2022-03-01	2022-04-30	897	86	9.59
56	34	Macron;Présidenti...	2022-03-01	2022-04-30	897	100	11.15
56	50	Macron;Présidenti...	2022-03-01	2022-04-30	897	226	25.2
10	24	Syrie;Macron;Ghou...	2017-12-01	2018-04-30	824	263	31.92
10	25	Syrie;Macron;Ghou...	2017-12-01	2018-04-30	824	46	5.58
10	30	Syrie;Macron;Ghou...	2017-12-01	2018-04-30	824	46	5.58
10	34	Syrie;Macron;Ghou...	2017-12-01	2018-04-30	824	11	1.33
10	50	Syrie;Macron;Ghou...	2017-12-01	2018-04-30	824	458	55.58

only showing top 50 rows

Bibliographie

[1] *Apache Spark™ - Unified Engine for Large-Scale Data Analytics*. URL : https://spark.apache.org/.

[2] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In : *Journal of Statistical Mechanics : Theory and Experiment* 2008.10 (oct. 2008), P10008. DOI : 10.1088/1742-5468/2008/10/p10008. URL : https://doi.org/10.1088/1742-5468/2008/10/p10008.

[3] *Citus Data | Distributed Postgres. At Any Scale*. URL : https://www.citusdata.com/.

[4] *Debian – Le Système d'exploitation Universel*. URL : https://www.debian.org/.

[5] *FAQ Citus - shard replication*. URL : https://docs.citusdata.com/en/v10.1/faq/faq.html#how-does-citus-handle-failure-of-a-worker-node.

[6] *Gephi - The Open Graph Viz Platform*. URL : https://gephi.org/.

[7] PostgreSQL Global Development Group. *PostgreSQL*. PostgreSQL. 2022-09-05T18 :55 :42.220960. URL : https://www.postgresql.org/.

[8] *Igraph – Network Analysis Software*. URL : https://igraph.org/.

[9] *Jeu de données à l'issue de la phase de pré-traitement*. URL : https://nextcloud-ext.sujets-libres.fr/index.php/s/qN8z3fWqycz57ks.

[10] *John Snow Labs - Spark NLP*. URL : https://nlp.johnsnowlabs.com/.

[11] Éditions Larousse. *Définitions : actualité, actualités - Dictionnaire de français Larousse*. URL : https://www.larousse.fr/dictionnaires/francais/actualit%C3%A9/956.

[12] Éditions Larousse. *Définitions : dominant, dominante - Dictionnaire de français Larousse*. URL : https://www.larousse.fr/dictionnaires/francais/dominant/26377.

[13] *Project Jupyter*. URL : https://jupyter.org.

[14] *Proxmox VE - Virtualization Management Platform*. URL : https://www.proxmox.com/en/proxmox-ve.