Social Media Sentiment Analysis

Premier Partie :

Lecture et Normalisation des Données :

Table Reader : Lit les données à partir d'un fichier ou d'une source de données.

Column Filter : Permet de sélectionner ou d'exclure certaines colonnes des données.

- Les colonnes sélectionnée sont
- (Node.id (String)
- Authority.score (Number (double))
- Hub.score (Number (double))
- Good.Bad.Rating (Number (double))
- Good.Bad.Rating.binned (String)

Normalizations: **meta-nodes** Applique des normalisations sur certaines métriques comme le "Auth Score" et le "Hub Score" ainsi que l'"Attitude".

- **méta-Node normalize AuthScore :** Java Snippet Applique un effet de saturation aux valeurs d'AuthScore.

Si AuthScore > 0.3, alors la valeur est fixée à 1 (ScentCone, TubeSteak, TripMaster Monkey).
Math Formula :
Normalise les valeurs d'AuthScore de la plage [0, 0.3] à [0, 1].
en utilisant une formule telle que (AuthScore - min(AuthScore)) / (max(AuthScore) - min(AuthScore))

- **méta-Node normalize Hub Score** : même principe avec un effet de saturation de 0.6
- **Metanode: normalize attitude :**

Java Snippet :Applique un effet de saturation aux valeurs.
Saturation effect for attitude > 66", ce qui signifie que toutes les valeurs d'attitude supérieures à 66 seront ramenées à une valeur :

```
Double a = $Good.Bad.Rating$;
if(a > 66.0)  a =66.0;
if( a < -66.0)  a = -66.0;
return a;
```

Row Splitter Un sous-ensemble contient des lignes avec des attitudes positives (attitude > 0) et l'autre des attitudes négatives (attitude <= 0).

Math Formula (Positive attitude mapped to [0.5, 1]):

Math Formula (Negative attitude mapped to [0, 0.5]):

Concatenate : Combine les deux sous-ensembles de données traités

- **Numeric**

diviser une variable numérique en catégories ou "bins" basées sur des plages de valeurs. Chaque configuration montre comment les différentes métriques sont binées.

Authority.score:
low: Les valeurs de Authority.score inférieures à 0.4 sont classées comme low.
medium: Les valeurs de Authority.score entre 0.4 et 0.6 sont classées comme medium.
high: Les valeurs de Authority.score supérieures à 0.6 sont classées comme high.

Hub.score:
low: Les valeurs de Hub.score inférieures à 0.4 sont classées comme low.
medium: Les valeurs de Hub.score entre 0.4 et 0.6 sont classées comme medium.

high: Les valeurs de Hub.score supérieures à 0.6 sont classées comme high.

Good.Bad.Rating:
negative: Les valeurs de Good.Bad.Rating inférieures à 0.4 sont classées comme negative.
neutral: Les valeurs de Good.Bad.Rating entre 0.4 et 0.7 sont classées comme neutral.
positive: Les valeurs de Good.Bad.Rating supérieures à 0.7 sont classées comme positive.

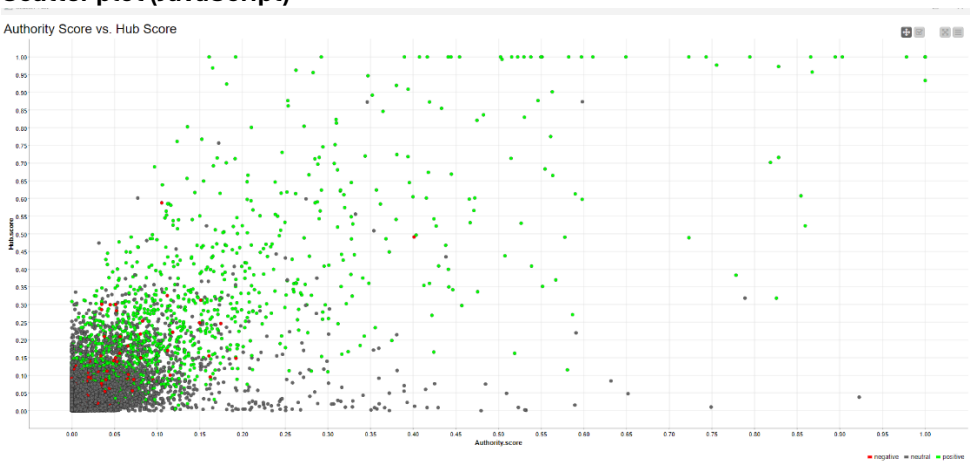
======➔ créer une nouvelle colonne dans le jeu de données avec les catégories correspondantes pour chaque enregistrement basé sur la valeur de la métrique spécifiée

Deuxième Partie :

Color Manager

attribuer des couleurs aux données en fonction de la valeur de la colonne sélectionnée, qui est Good.Bad.Rating_binned

Scatter plot (JavaScript)



Chaque point représente un individu ou une entité avec ses deux métriques correspondantes.
Couleurs des points :
Vert : Indique une attitude positive.
Rouge : Indique une attitude négative.
Gris : Indique une attitude neutre.

Concentration des points : La majorité des points se concentrent dans le coin inférieur gauche du graphique, indiquant que la plupart des individus ont des scores d'autorité et des scores de hub faibles.
Plus on s'éloigne du coin inférieur gauche, moins les points sont concentrés, indiquant qu'il y a moins d'individus ayant des scores élevés pour ces deux métriques.

Equal Size Sampling

créer des sous-ensembles de données de taille égale à partir d'un jeu de données plus grand. sélectionne le colonne nominale, **Good.Bad.Rating_binned,** le groupe de chaque ligne de données.
Il est configuré pour utiliser un échantillonnage exact (Use exact sampling), ce qui signifie qu'il va tenter de créer des échantillons de taille exactement égale pour chaque valeur unique dans la colonne nominale sélectionnée.
L'option "Enable static seed" est cochée avec une valeur spécifique. Cela garantit que les résultats de l'échantillonnage sont reproductibles. Si on exécute à nouveau le nœud avec la même graine (seed), on obtient le même échantillon.

k-Means

réaliser un clustering, qui est une méthode d'analyse non supervisée destinée à regrouper des objets en k clusters en se basant sur leurs caractéristiques.

Les objets dans le même groupe (ou cluster) sont plus similaires entre eux qu'aux objets des autres groupes.
les propriétés du nœud k-Means :
Nombre de clusters:
Vous avez défini le nombre de **clusters à 10**, ce qui signifie que l'algorithme va essayer de **regrouper les données en 10 clusters distincts**
initialisation des centroïdes:
Premières lignes (First k rows): Les centroïdes initiaux sont choisis comme étant les premières k lignes de l'ensemble de données
Nombre d'itérations:
Max. number of iterations: Définit à 99 le nombre maximum d'itérations pour l'algorithme k-Means. L'algorithme s'arrête soit lorsque les centroïdes ne changent plus (ou changent très peu), soit lorsque le nombre maximum d'itérations est atteint.
Sélection des colonnes:
Les colonnes à inclure dans l'analyse k-Means sont sélectionnées ici.
Include: **Authority.score, Hub.score, et Good.Bad.Rating** l'algorithme k-Means utilisera ces trois dimensions pour trouver des clusters.

Extract features from K-Means model :
ce nœud métier est désormais redondant car sue cet exemple K-Means produit déjà le centre du cluster à la deuxième sortie

Cluster Assigner (Attributeur de Clusters)
prendre de nouveaux points de données et de les assigner à un ensemble existant de clusters ou de prototypes. Ces prototypes sont typiquement le résultat d'un algorithme de clustering tel que k-means
Ports d'entrée :
Port 0 : Ce port doit recevoir le modèle de prototype
Port 1 : Ce port est pour la table de données réelle contenant les nouvelles données ou les données non étiquetées qui seront classées en fonction du prototype
Ports de sortie :
Port 0 : La sortie est les données d'entrée avec une attribution supplémentaire aux prototypes de cluster. cela signifie que chaque ligne dans la table de données originale aura maintenant une étiquette de cluster qui lui est attribuée.

Row Filter:

Ces nœuds filtrent les lignes de données, probablement sur la base d'une condition ou d'un critère spécifique (par exemple, sélectionner uniquement les données appartenant à Cluster_3 ou Cluster_1).

Reporting

Data to Report (BIRT):
Ces nœuds préparent les données pour le Reporting en utilisant BIRT (Business Intelligence and Reporting Tools),

séparation des données en clusters, suivie d'un filtrage et d'un tri basé sur un score d'autorité, et enfin, la préparation des données triées pour la génération de rapports. Les nœuds sont connectés de manière que les données soient traitées de manière séquentielle, avec des branches distinctes pour chaque **cluster spécifié (Cluster_3 et Cluster_1).**

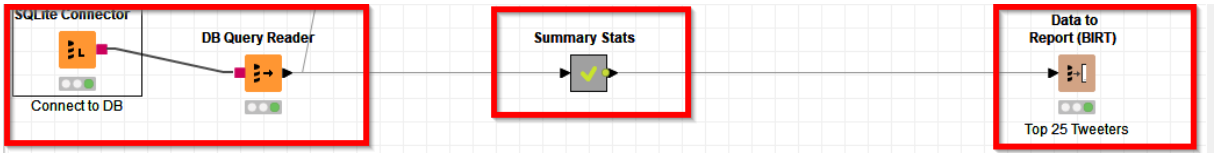
Authority.score et Hub.score sont des mesures de l'importance d'un utilisateur dans le réseau, basées sur des mesures comme le nombre de suiveurs ou la fréquence des interactions.
Good.Bad.Rating est une évaluation attribuée à chaque utilisateur. Les méthodes exactes pour déterminer pourraient être basées sur l'analyse du sentiment, la qualité du contenu, ou d'autres mesures de performance.
Good.Bad.Rating.binned et Hub.score_binned sont des versions catégorisées de ces scores, simplifiant les évaluations continues en groupes discrets comme 'low', 'neutral', et 'positive'.
Cluster indique le groupe auquel chaque utilisateur a été assigné par l'algorithme k-Means.
ce workflow semble être utilisé pour analyser les données des utilisateurs de médias sociaux, en classant les utilisateurs en groupes basés sur leurs comportements et interactions, et en évaluant leur influence et qualité.
Ces informations pourraient être utilisées pour identifier des influenceurs clés, pour cibler des publicités ou des campagnes, ou pour comprendre la structure de la communauté au sein du réseau social.

le cluster 3
le score d'autorité le plus élevé (0.659) et le score de hub le plus élevé (0.806), ce qui le rend significatif dans l'analyse des données de médias sociaux. Cela peut signifier que les utilisateurs dans le cluster 3 sont considérés comme très influents ou centraux dans le réseau social analysé.
Les utilisateurs avec une forte influence ou une forte activité.
les utilisateurs du cluster 3 pourraient être ciblés pour des stratégies de marketing en raison de leur influence élevée.
Cluster 3 a le **GoodBadRating** le plus élevé de 1, une taille de 6 membres, le plus haut AuthorityScore de 0.659, et le plus haut HubScore de 0.806.
En se basant sur le GoodBadRating, choisir le cluster 3 semble logique car il a le score le plus élevé, ce qui suggère que les membres de ce cluster sont évalués très positivement selon les critères définis pour cette mesure

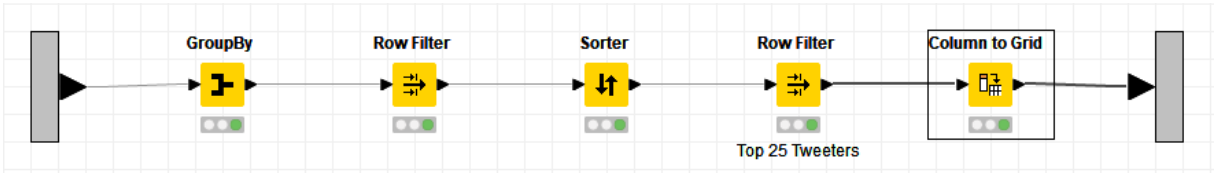
Cluster 1 avec le score le plus bas peut être étudié pour comprendre les caractéristiques ou les comportements qui contribuent à une évaluation négative.
Ceci est utile pour identifier les domaines d'amélioration ou les risques potentiels dans une communauté ou un réseau social (utilisateurs peu engageants, le contenu de mauvaise qualité, ou les comportements indésirables).
En comparant les données extrêmes, nous pouvons souvent obtenir des informations sur les facteurs qui influencent le plus les résultats.
Cela permet de déterminer quelles caractéristiques sont associées aux évaluations les plus hautes et les plus basses et peut être utilisé pour informer les décisions stratégiques, les interventions ciblées, ou les campagnes de communication dans le cadre des médias sociaux.

Twitter Analysis

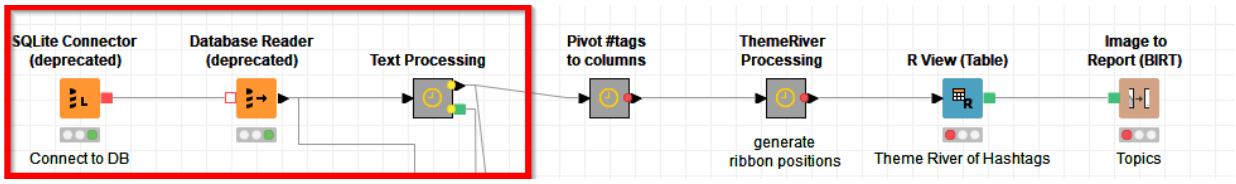
Premier Plan



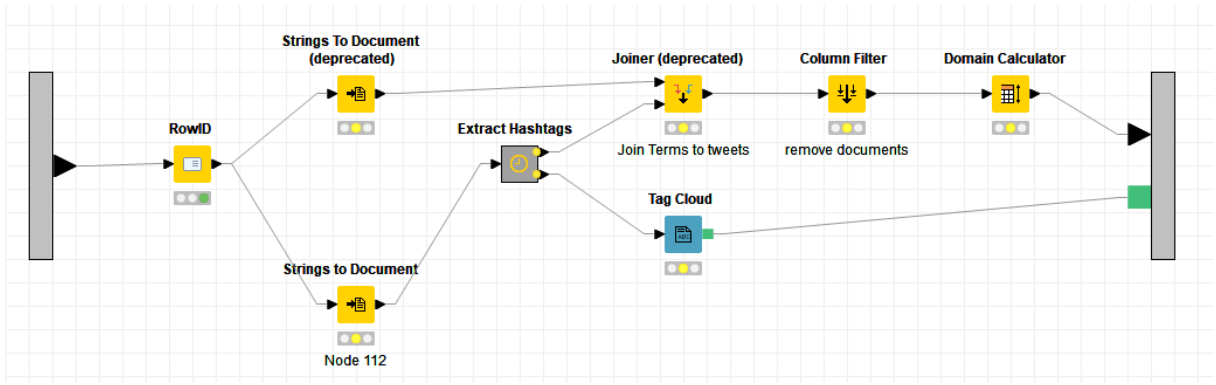
SQLite Connector : Ce nœud est utilisé pour établir une connexion à une base de données SQLite.
DB Query Reader : Une fois la connexion établie, ce nœud exécute une requête SQL sur la base de données.
Summary Stats : Ce nœud calcule des statistiques récapitulatives pour les données récupérées par le nœud DB Query Reader.

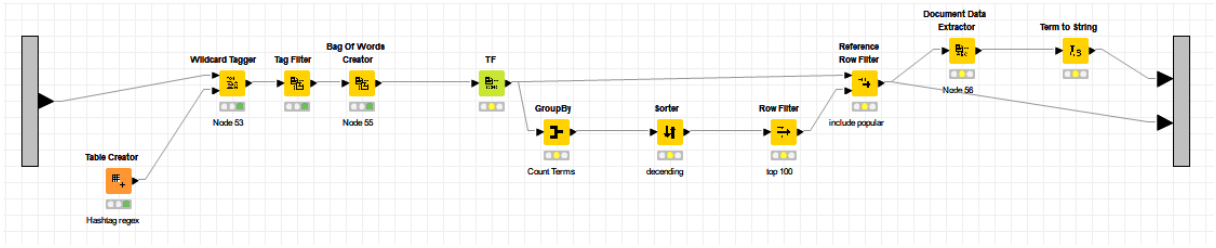


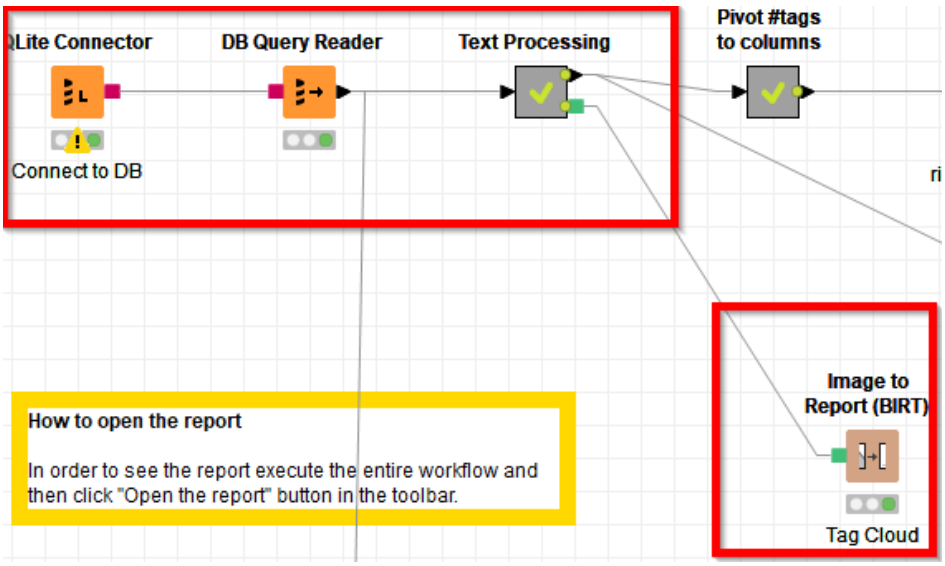
Data to Report (BIRT) : Ce nœud est utilisé pour préparer les données pour un rapport. BIRT a la fin on obtient le Top 25 Tweeters
Les données traitées à travers ce workflow ont subi une série d'opérations de transformation et d'agrégation méthodiques pour aboutir à une liste concise des utilisateurs les plus actifs sur Twitter en fonction du nombre de tweets. Cette liste est essentielle pour les analyses d'influence et de portée sur les réseaux sociaux.
Étape d'Agrégation (GroupBy) : L'agrégation initiale a servi à compiler les activités de tweet par utilisateur, fournissant une mesure quantitative de l'engagement de chaque compte. Cette étape est cruciale car elle établit la base pour identifier les utilisateurs qui dominent la conversation sur le sujet analysé.
Étape de Filtrage (Row Filter) : L'exclusion des utilisateurs contenant le terme spécifique "knime" a permis de purifier le jeu de données des entrées non pertinentes pour l'analyse en cours. Cela illustre l'importance de nettoyer les données pour une interprétation précise des résultats.
Étape de Tri (Sorter) : Le classement des utilisateurs par nombre décroissant de tweets est une pratique analytique standard qui met en lumière les utilisateurs les plus influents en tête de liste, facilitant ainsi l'identification rapide des principaux acteurs.
Deuxième Filtrage (Row Filter pour Top 25 Tweeters) : La sélection des 25 premières lignes post-tri cible l'élite des utilisateurs par activité de tweet. C'est une étape déterminante pour concentrer l'analyse sur les utilisateurs clés sans être submergé par le volume de données.
Réorganisation des Données (Column to Grid) : La répartition finale des données sélectionnées dans une grille de 5 colonnes était une décision de visualisation conçue pour présenter les données de manière structurée et accessible.
Le résultat est une table nette et organisée qui affiche les 25 utilisateurs les plus actifs, distribués sur 5 lignes. Chaque ligne présente un sous-ensemble de cette élite, aligné côte à côte pour une comparaison et une évaluation rapides.



Metanode TextProcessing







Le workflow commence avec une connexion à une base de données, indiquée par le noeud "Connect to DB" et le "Twitter API Connector", qui semblent être utilisés pour extraire des données
Les informations extraites sont ensuite traitées par une série
Connexion et Extraction des Données:
La base de données est connectée et les tweets sont extraits. Cette étape constitue la fondation de l'analyse, fournissant les données brutes pour les traitements subséquents.
Traitement de Texte (Text Processing, Strings To Document) :
Les tweets sont convertis en documents pour permettre l'analyse textuelle. Cette conversion est une étape essentielle pour l'application des techniques de traitement du langage naturel (NLP).
Extraction de Hashtags et Agrégation (Extract Hashtags, GroupBy) :
Les hashtags sont extraits et groupés pour calculer leur fréquence d'apparition. Ces informations sont cruciales pour identifier les sujets dominants et les tendances au sein des conversations Twitter.
Filtrage et Triage des Termes (Sorter, Row Filter) :
Les termes sont triés par ordre décroissant de fréquence et filtrés pour ne retenir que les plus pertinents, souvent les 15 premiers. Ce processus affine l'ensemble de données pour se concentrer sur les termes les plus influents.
Visualisation (Tag Cloud, Image to Report (BIRT)) :
La visualisation finale sous forme de nuée de mots met en évidence visuellement les hashtags prédominants. Les termes de grande fréquence apparaissent plus grands, ce qui permet une interprétation rapide et intuitive des thèmes les plus discutés.
Le résultat final, une image de nuée de mots intégrée dans un rapport BIRT, permet une analyse visuelle directe des tendances des hashtags sur Twitter, offrant une compréhension immédiate des points de discussion clés. Cette image peut être utilisée pour informer les décisions en matière de stratégie de con
Les différents utilisateurs sont représentés par des icônes, et les liens entre eux peuvent représenter des relations telles que des mentions, des retweets ou des réponses.

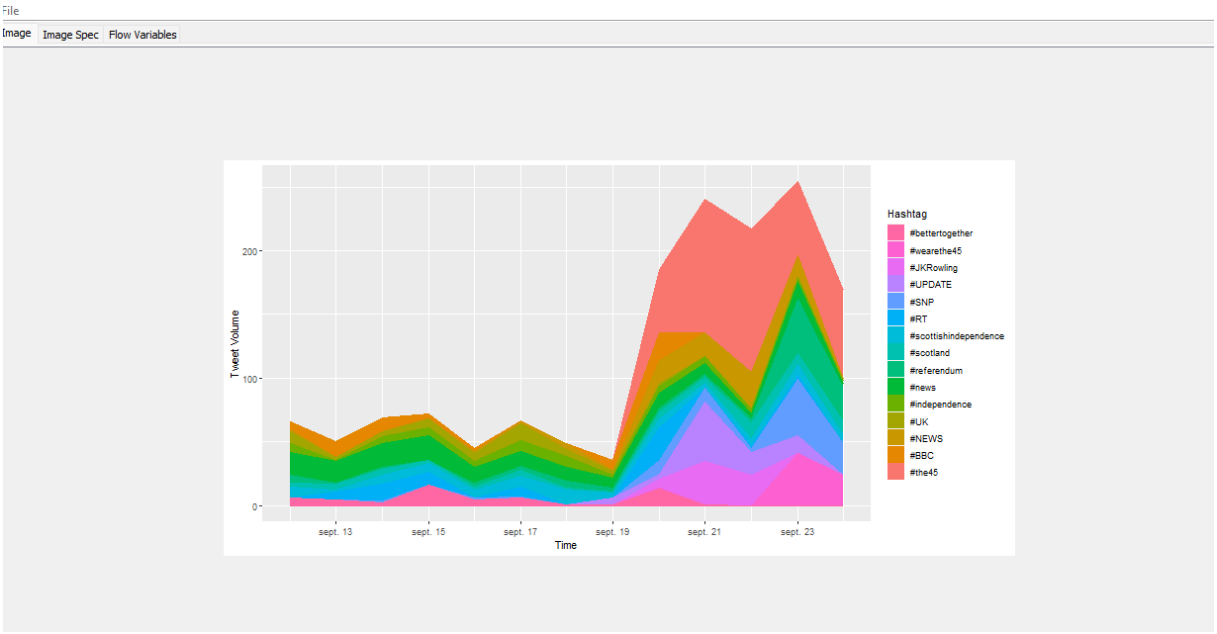
La méthodologie appliquée dans ce workflow KNIME démontre une approche rigoureuse et structurée, essentielle pour délivrer des insights actionnables à partir de données sociales volumineuses et complexes.
La finalité est de comprendre les dynamiques de communication et d'influence au sein Twitter, ce qui pourrait être pertinent pour des analyses de marché, de l'opinion publique ou de la diffusion de l'information.

Bag Of Words Creator : Dans le modèle BoW, le texte est transformé en une collection de mots indépendants sans prendre en compte l'ordre ou la grammaire.

TF

pour calculer la fréquence à laquelle chaque terme apparaît dans chaque document
Fréquence Relative : C'est la fréquence d'un terme divisée par le nombre total de termes dans le document, ce qui donne une mesure normalisée de la fréquence.
ce nœud convertit les données traitées en un format de chaîne de texte.

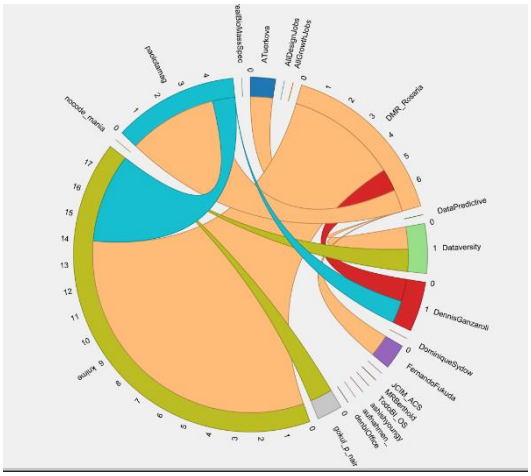
Row Filter & Reference Row Filter : Ces nœuds filtrent les lignes selon des critères spécifiques.
Document Data Extractor & Term to String : Ces nœuds convertissent les données de document en chaînes de caractères pour préparer les hashtags à une étape de visualisation ou d'exportation.

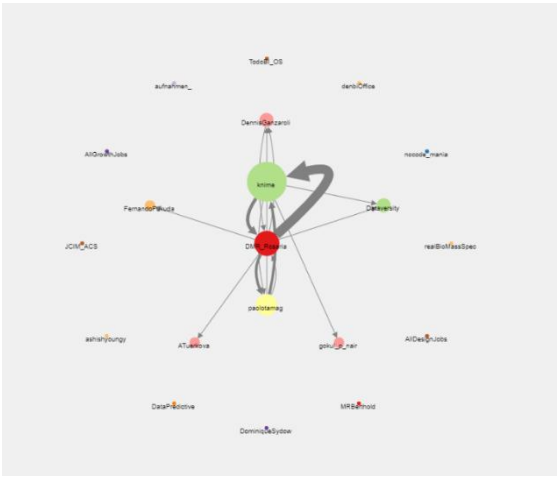


L'image finale représente un graphique qui est une forme de visualisation des données temporelles pour montrer le volume de mentions de différents hashtags sur Twitter au fil du temps.
Ce type de graphique est utile pour observer les tendances et la dynamique des discussions sur les réseaux sociaux.
Variabilité Temporelle : La largeur des bandes de couleur varie avec le temps, ce qui indique les fluctuations dans la fréquence des hashtags.
Un pic dans la largeur suggère une discussion accrue autour du sujet marqué par le hashtag à ce moment précis.

Comparaison des Hashtags : Les différentes couleurs représentent les différents hashtags. Cela permet de comparer facilement le volume de mentions entre eux au fil du temps.
Événements Clés : Des pics ou des changements notables dans la largeur des bandes peuvent coïncider avec des événements spécifiques, indiquant des moments où un sujet a gagné en popularité.
Interactions entre les Sujets : L'empilement des bandes montre comment certains sujets sont discutés en tandem ou si certains hashtags tendent à être mentionnés ensemble.
Reconnaissance des Tendances : Ce graphique peut être utilisé pour identifier les moments où certains sujets ont dominé la conversation sur Twitter. Ceci est utile pour les analyses de marché, la surveillance de la réputation, et la compréhension de l'engagement du public.
Planification de Contenu : Pour les créateurs de contenu et les marketeurs, comprendre quels hashtags attirent l'attention à quels moments peut aider à planifier des campagnes ou des publications stratégiques.
Analyse de l'Impact des Événements : En alignant les pics avec des événements du monde réel, on peut mesurer l'impact de ces événements sur les discussions en ligne.
Optimisation des Hashtags : Pour ceux qui cherchent à optimiser leur utilisation des hashtags, cette visualisation peut indiquer quels hashtags sont susceptibles de maximiser la visibilité à des moments donnés.
Pour réduire les conclusions, on pourrait dire que l'image est un outil puissant pour visualiser l'évolution de la popularité des hashtags sur Twitter, permettant aux utilisateurs de détecter rapidement les changements dans le comportement du public et d'ajuster leurs stratégies en conséquence.

03_Visualizing_Twitter_Network_with_a_Chord_Diagram





le flux de travail final a pour l'objectif soit de visualiser les interactions entre les utilisateurs de Twitter, plus précisément, comment les utilisateurs se retweetent les uns les autres

Accès aux données

Table Reader: Ce nœud lit les données depuis une source externe.

Transformation

Count Retweets: Ce nœud calcule le nombre de retweets entre chaque paire d'utilisateurs. Cela permet de comprendre non seulement combien de fois un utilisateur a été retweeté, mais aussi qui retweete qui.

Create matrix input for chord plot: Ce nœud crée une matrice d'adjacence pondérée par le nombre de retweets, qui est nécessaire pour générer le diagramme

Visualisation

Generic JavaScript ViewUtilise un script JavaScript pour créer une vue interactive représentant le réseau de retweets sous forme de diagramme en accordéon. Ce type de visualisation permet d'illustrer les interactions entre les utilisateurs de manière claire et esthétique.

Prétraitement : Arêtes du réseau

Row Filter & Rule-based Row Filter: Ces nœuds sont utilisés pour filtrer les données. Le configuré pour éliminer les lignes sans retweets, et le second pour supprimer les auto-retweets

GroupBy: t utilisé pour consolider les données et trouver les arêtes qui représentent les retweets entre les utilisateurs.

Construction de la liste des utilisateurs

Sorter & Row Filter: Ces nœuds sont utilisés pour trier les utilisateurs par le nombre de fois qu'ils ont été retweetés et pour ne garder que les 20 utilisateurs les plus retweetés.

Column Filter & Cross Joiner & Joiner: Ces nœuds construisent toutes les combinaisons possibles d'utilisateurs et ajoutent des informations sur la fréquence des retweets entre eux.

Missing Value & Pivot & RowID: Ces nœuds traitent les valeurs manquantes et pivotent les données pour créer une matrice pour les utilisateurs retweetés.

Diagramme en accordéon

La visualisation finale montre comment les utilisateurs interagissent entre eux en termes de retweets. Chaque segment de l'extérieur du cercle représente un utilisateur, et les arcs à l'intérieur relient les utilisateurs entre eux, avec l'épaisseur de l'arc représentant le volume de retweets. Cette visualisation permet d'identifier rapidement les utilisateurs les plus influents et les interactions clés au sein du réseau. Les arcs les plus larges représentent un grand nombre de retweets entre les utilisateurs

knime apparaît à la fois comme un utilisateur qui retweete et comme celui dont les tweets sont retweetés. Par exemple, le premier enregistrement montre que le tweet de "knime" a été retweeté 14 fois par DMR_Rosaria.

DMR_Rosaria a retweeté paolotamag 4 fois.

knime a retweeté paolotamag 4 fois également.

paolotamag a également une présence notable dans ce réseau d'interactions sur Twitter, avec des tweets qui ont été retweetés par au moins deux utilisateurs différents à plusieurs reprises

DennisGanzaroli a retweeté un tweet de DMR_Rosaria
DennisGanzaroli a également retweeté un tweet de knime

Conclusions
Ce flux de travail fournit une visualisation puissante qui résume les interactions complexes au sein d'un réseau social. Les conclusions tirées d'une telle analyse pourraient être utilisées pour identifier les influenceurs clés, comprendre les dynamiques de propagation de l'information, et potentiellement pour cibler des campagnes marketing ou de communication plus efficacement en fonction de ces interactions.