**Social Media Sentiment Analysis**

**Une image contenant texte, capture d’écran, diagramme, nombre

Description générée automatiquement**

**Partie 1 :**

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.

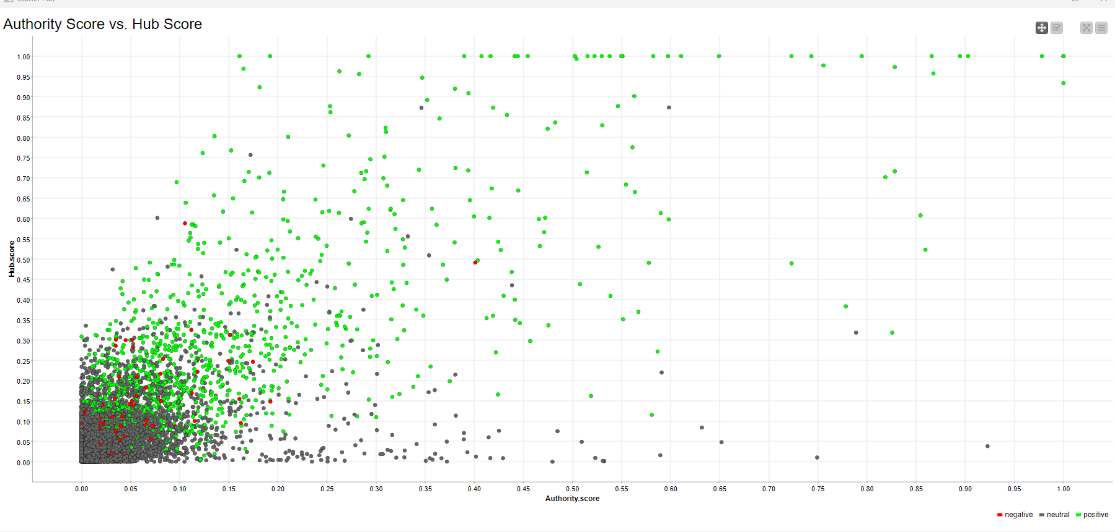
======🡺 créer une nouvelle colonne dans le jeu de données avec les catégories correspondantes pour chaque enregistrement basé sur la valeur de la métrique spécifiée

**Partie 2 :**

**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).

**Partie 3 :**

**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.

**Nœud Assignateur de Clusters :**

Une image contenant texte, capture d’écran, diagramme, Police

Description générée automatiquement

But : Le nœud assignateur de clusters est utilisé pour catégoriser

de nouveaux points de données à des clusters prédéfinis.

Ces clusters auraient été créés plus tôt en utilisant un algorithme de clustering k-means.

Fonctionnement :

**Modèle de Prototype (Port d'entrée 0)** : Ce port prend en entrée le modèle contenant les prototypes de clusters. Ces prototypes représentent les centres de clusters obtenus par un algorithme de clustering k-means .

**Données d'Entrée (Port d'entrée 1)** : Ce port est destiné à recevoir les nouvelles données qui seront assignées aux prototypes de clusters.

**Assignation** : Chaque point de donnée dans le tableau d'entrée est comparé aux prototypes de clusters, et est assigné au prototype le plus proche selon une métrique de distance spécifique (la distance euclidienne).

**Sortie :**

**Données Assignées (Port de sortie 0)** : Les données d'entrée avec une colonne supplémentaire indiquant à quel cluster prototype chaque point de donnée a été assigné.

Entré on a une table a 8 colonne , après l’exécution de ce nœud on aura une colonne de plus

Ce processus est utilisé les nouveaux points de données doivent être classés en catégories basées sur un modèle existant, sans avoir à exécuter à nouveau l'algorithme de clustering sur l'ensemble des données.

**Etapes d’analyse Social Media Sentiment :**

Ce workflow est conçu pour effectuer une classification non supervisée, ou "clustering", des utilisateurs de médias sociaux en utilisant l'algorithme k-means.

les étapes clés du workflow :

***Lecture et Normalisation des Données*** : Les données sont lues à l'aide d'un nœud "Table Reader", filtrées avec un "Column Filter", et normalisées.

***Gestion des Couleurs*** : Après la normalisation, un nœud "Color Manager" est utilisé pour attribuer des couleurs à des fins de visualisation basées sur le score d'attitude des utilisateurs (vert pour positif, rouge pour négatif et gris pour neutre).

***Échantillonnage et Clustering :***

Échantillonnage de Taille Égale : Cela indique que le workflow prend un nombre égal d'échantillons pour chaque classe, probablement pour équilibrer la distribution des classes.

***k-Means*** : Le nœud k-means est configuré avec K=10, ce qui signifie que l'algorithme tentera de trouver 10 groupes dans les données.

Visualisation :

***Scatter Plot (JavaScript) :*** Un graphique de dispersion est créé, probablement pour visualiser les clusters ou d'autres relations telles que le score d'autorité par rapport au score de hub.

***Extraction de Caractéristiques et Rapports :***

***Extrait les caractéristiques du modèle k-Means*** : Bien que la note indique que ce nœud est maintenant redondant.

***Section de Rapport :*** Il y a un processus de reporting qui inclut un nœud "GroupBy", un nœud "Joiner" obsolète, et deux instances du nœud "Report (BIRT)" pour créer des rapports.

***Attribution de Clusters :***

Cluster Assigner : Ce nœud attribue des clusters aux données.

Row Filter et Sorter : Filtre et trie les lignes selon certains critères (comme le numéro de cluster et le score d'autorité).

Remarque : Certains nœuds sont marqués comme obsolètes, ce qui signifie qu'ils sont dépassés et peuvent ne pas être pris en charge dans les futures versions de KNIME.

**Résumer :**

Le workflow KNIME présenté constitue un exemple sophistiqué d'application des techniques de machine learning pour l'analyse des réseaux sociaux, plus précisément par le biais du clustering basé sur l'algorithme k-means.

La première phase du processus consiste en la préparation des données, où les données extraites sont d'abord soumises à un filtrage de colonnes pour ne retenir que les variables pertinentes, puis normalisées pour garantir une uniformité des échelles de mesure. Cette étape cruciale permet de s'assurer que les distances calculées lors du clustering reflètent fidèlement les similarités entre les instances.

Postérieurement, le workflow met en œuvre un nœud "Color Manager" pour l'affectation visuelle, attribuant des couleurs distinctes en fonction des scores d'attitude — une métaphore colorimétrique rendant les résultats plus intuitifs pour l'analyse humaine.

L'étape suivante engage l'algorithme k-means, configuré ici pour identifier dix clusters distincts, une décision modélisée sur la prémisse d'une hétérogénéité présupposée au sein de la population des utilisateurs. Ce choix d'un nombre prédéterminé de clusters doit être évalué à l'aide de méthodes comme la méthode du coude ou des scores de silhouette pour valider la cohérence interne des groupes formés.

Enfin, le workflow converge vers une composante analytique et de reporting robuste. Les résultats du clustering sont explorés au moyen de visualisations JavaScript, offrant une perspective comparative des scores d'autorité et de hub, des métriques clés dans l'analyse des réseaux.

Les clusters sont alors assignés et triés, et un processus de reporting est amorcé avec l'intégration de BIRT (Business Intelligence and Reporting Tools), facilitant la génération de rapports détaillés et la diffusion des insights extraits.

Cette structuration méticuleuse du workflow non seulement met en lumière la capacité de KNIME à traiter et analyser des données complexes, mais démontre également la flexibilité et la puissance de la plateforme pour l'adaptation aux exigences spécifiques de l'analyse des réseaux sociaux.

**Twitter Analysis**

***Twitter Data Collection***

Une image contenant texte, capture d’écran, diagramme, ligne

Description générée automatiquement

Ici dans ce workflow : collecte et le stockage de données Twitter.

Ce processus systématique illustre comment KNIME peut être utilisé pour automatiser la collecte de données sociales en temps réel et pour mettre à jour une base de données avec de nouvelles informations.

***Twitter API Connector :*** Ce nœud sert de point d'entrée pour se connecter à l'API Twitter. L'utilisateur doit fournir des clés API et des jetons d'accès, qui peuvent être obtenus en enregistrant une application via le portail de développement Twitter. Ces informations d'identification sont essentielles pour authentifier les requêtes API et permettre l'extraction de données.

***Twitter Search :*** Après l'authentification, le nœud "Twitter Search" est utilisé pour rechercher et collecter des tweets en fonction de termes spécifiques fournis par l'utilisateur. Cela pourrait inclure des hashtags, des mots-clés ou des phrases spécifiques.

***SQLite Connector et DB Update :*** Les données collectées sont ensuite acheminées vers une base de données SQLite via le "SQLite Connector". Si la base de données contient déjà des enregistrements liés aux nouveaux tweets collectés, le nœud "DB Update" est utilisé pour mettre à jour les nombres de favoris et de retweets.

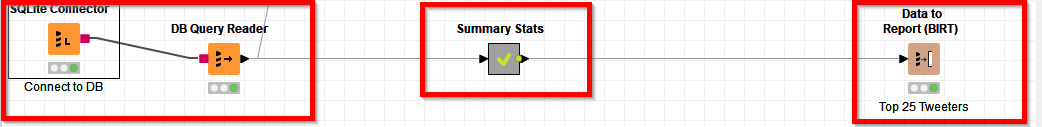
***Row Filter et Column Filter :*** Avant l'écriture dans la base de données, un "Row Filter" est appliqué pour supprimer les lignes déjà mises à jour, et un "Column Filter" pour éliminer les colonnes superflues, ne retenant que les données pertinentes pour le stockage.

***DB Writer :*** Enfin, les nouveaux tweets collectés et filtrés sont écrits dans la base de données SQLite à l'aide du nœud "DB Writer". Cela ajoute des informations fraîches à la base de données à chaque exécution du workflow.

Le workflow peut être exécuté périodiquement pour collecter des données longitudinales, ce qui est utile pour suivre l'évolution des tendances et des discussions sur Twitter au fil du temps. De plus, il peut être configuré pour s'exécuter automatiquement via le KNIME Server ou par ligne de commande avec l'exécuteur en lot de KNIME, permettant une collecte de données efficace et sans surveillance.

***Twitter Data 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.

Une image contenant texte, ligne, Police, nombre

Description générée automatiquement

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.

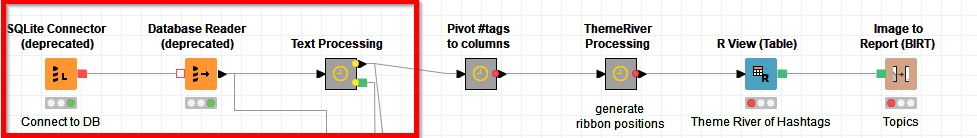
Étape de Filtrage (Row Filter) : L'exclusion des utilisateurs contenant le terme spécifique "knime" a permis de purifier le jeu de données des entrées non pertinentes pour l'analyse en cours. Cela illustre l'importance de nettoyer les données pour une interprétation précise des résultats.

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**

Une image contenant texte, capture d’écran, ligne, diagramme

Description générée automatiquement

Une image contenant ligne, Tracé, diagramme, texte

Description générée automatiquement

Une image contenant texte, capture d’écran, diagramme, ligne

Description générée automatiquement

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.

Une image contenant texte, capture d’écran, Tracé, Caractère coloré

Description générée automatiquement

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.

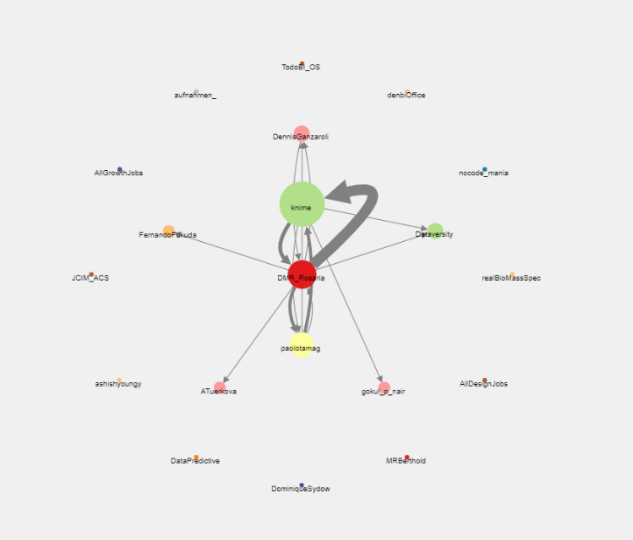
**Optimisation des Hashtags** : Pour ceux qui cherchent à optimiser leur utilisation des hashtags, cette visualisation peut indiquer quels hashtags sont susceptibles de maximiser la visibilité à des moments donnés.

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.



Une image contenant diagramme, cercle, Graphique, capture d’écran

Description générée automatiquement



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

Le workflow combine le traitement de texte, l'analyse des tendances des hashtags et l'analyse des réseaux pour examiner les interactions et l'influence au sein des données collectées.

Le processus débute par la collecte des tweets via l'API Twitter, où l'authentification est nécessaire à l'aide des clés API et des jetons d'accès. Le nœud "Twitter API Connector" est prévu à cet effet. Si les clés ne sont pas disponibles, le workflow peut utiliser un ensemble de données prédéfini.

Les données de tweets sont ensuite lues à partir d'une base de données via un "SQLite Connector" (marqué comme obsolète) et un "Database Reader" (également obsolète).

Une fois les données chargées, elles sont traitées par le nœud "Text Processing" pour extraire les hashtags et calculer les fréquences des termes. Ces informations sont utilisées pour créer un nuage de mots et pour suivre l'évolution des hashtags au fil du temps avec le traitement "ThemeRiver".

La partie centrale du workflow est l'analyse des réseaux, où les données de tweets sont analysées pour déterminer les utilisateurs influents et visualiser le réseau de retweets. Cela est réalisé par le nœud "Analyze Network" et visualisé localement avec "Network Viewer".

Enfin, les résultats de l'analyse sont préparés pour le reporting. Un "R View (Table)" pour les statistiques sommaires, "Theme River of Hashtags", "Influential Users", et "Retweet Network" sont convertis en images pour être inclus dans un rapport généré par BIRT (Business Intelligence and Reporting Tools).

Le workflow se termine avec la création de rapports qui incluent des visualisations comme le nuage de mots et la liste des 25 meilleurs twitteurs, fournissant une synthèse complète des tendances et de l'influence des utilisateurs pendant la période autour du référendum.

Le workflow est bien structuré pour permettre une réplication ou une extension future, avec des annotations qui orientent sur la manière d'éditer et de visualiser les rapports. Cependant, la mise à jour des nœuds obsolètes est nécessaire pour assurer une compatibilité à long terme avec les versions futures de KNIME.

***Bag of Words (BoW) :*** *Le modèle Bag of Words est une représentation simplifiée utilisée dans le traitement* du langage naturel et l'analyse de texte. Dans ce modèle, un texte (comme une phrase ou un document) est représenté comme le sac (ou ensemble) de ses mots, sans tenir compte de la grammaire et même de l'ordre des mots mais en maintenant la multiplicité. Les principales fonctionnalités du modèle BoW comprennent la construction d'un vocabulaire de mots connus (ou tokens) et l'utilisation de ce vocabulaire pour encoder des documents.

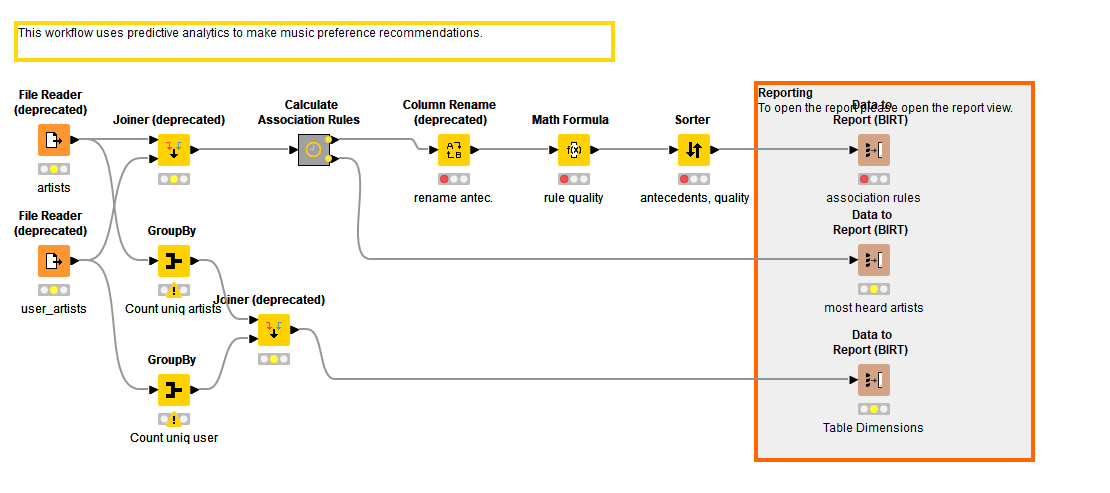
Le processus commence par la tokenisation et la suppression des mots fonctionnels (stop words). Les tokens restants sont alors souvent normalisés en termes de casse, et parfois en racine ou en lemme. Chaque mot unique est attribué à une caractéristique dans l'espace de vecteur, où chaque document est alors représenté par un vecteur dans ce même espace. La dimensionnalité de ce vecteur est égale au nombre de mots dans le vocabulaire. Cela permet une analyse quantitative par des algorithmes d'apprentissage automatique, bien que l'une des limites du modèle BoW soit son incapacité à capturer le contexte ou la sémantique des mots dans les documents, puisqu'il traite chaque mot indépendamment.

Term Frequency (TF) : Term Frequency est une mesure quantifiant l'importance d'un terme dans un document par rapport à un corpus. Dans le cadre du modèle BoW, le TF d'un terme est simplement le nombre de fois que le terme apparaît dans le document. Cette fréquence peut être utilisée telle quelle ou peut être pondérée par la longueur du document pour obtenir la fréquence relative du terme, afin d'éviter un biais vers les documents plus longs.

En plus de la simple fréquence des termes, des variantes plus sophistiquées, comme la fréquence de terme pondérée par l'inverse de la fréquence du document (TF-IDF), sont souvent utilisées pour réduire l'impact des mots qui apparaissent très fréquemment dans un corpus et qui sont donc moins informatifs. TF-IDF accorde plus de poids aux termes qui sont fréquents dans un document mais rares dans le corpus, ce qui permet de capturer l'importance relative des mots en tenant compte de leur répartition dans différents documents.

BoW et TF sont des composants fondamentaux dans les systèmes de traitement de texte qui transforment le texte en données numériques exploitables par des algorithmes d'apprentissage. Bien qu'efficaces pour de nombreuses tâches analytiques, leur utilisation est limitée dans des cas nécessitant une compréhension profonde de la sémantique, du contexte ou de la syntaxe, où des modèles plus avancés, tels que Word Embeddings ou des réseaux neuronaux profonds, pourraient être préférés.

**lastfm\_RecommendationsTwitter**



***File Reader:*** Il y a deux nœuds File Reader qui sont obsolètes.

***Joiner***

***GroupBy :*** Il y a deux nœuds GroupBy. Ces nœuds sont généralement utilisés pour regrouper des données selon plusieurs critères et effectuer des opérations telles que le comptage, la moyenne,, sur ces groupes.

***Calculate Association Rules :***

Ce nœud est utilisé pour calculer des règles d'association utiles pour les systèmes de recommandation.

***Column Rename :*** Ce nœud est utilisé pour renommer les colonnes de vos données.

***Math Formula :*** Ce nœud vous permet d'effectuer des calculs mathématiques sur les données. Il semble être utilisé ici pour calculer la qualité des règles.

***Sorter :*** Ce nœud trie les données en fonction de critères spécifiés.

***Reporting :*** Cette section implique la création de rapports en utilisant BIRT (Business Intelligence and Reporting Tools), un système de reporting open-source. Les nœuds "Data to Report" préparent les données pour le reporting, et les outils de reporting sont utilisés pour créer des représentations visuelles de l'analyse des données.

***Table Dimensions*** : Il s'agit probablement d'un nœud pour ajuster ou définir les dimensions des tableaux dans la phase de reporting.

Ce workflow analyse les préférences musicales et formuler des recommandations basées sur des méthodes analytiques prédictives.

Ce processus est articulé à travers une séquence d'opérations de traitement de données, chacune jouant un rôle distinct et crucial dans la transformation de données brutes en informations actionnables.

Acquisition de données : L'initiation du flux opérationnel se fait via deux instances de nœuds File Reader Ces nœuds servent à l'importation des ensembles de données.

Intégration de données : Le nœud Joiner est employé pour la fusion des jeux de données sur la base de clés communes, permettant ainsi une consolidation des informations relatives aux artistes et aux utilisateurs.

Agrégation : Les nœuds GroupBy interviennent deux fois dans le flux de travail pour regrouper les données selon des attributs spécifiques et effectuer des opérations de comptage distinctes, ce qui est essentiel pour résumer les interactions des utilisateurs avec différents artistes.

Analyse d'associations : Le nœud Calculate Association Rules sert à identifier et à générer des règles d'association. Ces règles dévoilent des relations intrinsèques entre les éléments du jeu de données, facilitant la création de systèmes de recommandation.

Transformation de données : La fonction du nœud Column Rename est de renommer les colonnes pour une cohérence et une interprétation optimisées des données ultérieures.

Calculs personnalisés : Le nœud Math Formula permet l'application de formules mathématiques sur les données.

Math Formula est utilisé pour calculer de nouvelles valeurs de colonnes ou manipuler des valeurs existantes en se basant sur des expressions mathématiques. Elle est configurée ici pour calculer une métrique nommée "qualité de règle". L'expression configurée dans ce nœud est :

$ItemSetSupport$ \* $RuleConfidence%$

qualité de règle" est calculée comme le produit de ItemSetSupport et de RuleConfidence

La qualité de règle calculée sera utilisée en aval, pour trier ou filtrer les recommandations avant de les présenter à l'utilisateur final ou de procéder à une analyse plus approfondie.

Tri et ordonnancement : L'utilisation du nœud Sorter est prévue pour l'ordonnancement des règles d'association basé sur le critère de qualité, favorisant une priorisation dans la présentation des recommandations.

Rapport et visualisation : La phase de rapport est gérée par un ensemble de nœuds Data to Report (BIRT), qui préparent les données pour une présentation structurée et une visualisation via le système de rapports BIRT. Cette étape est cruciale pour la communication efficace des découvertes analytiques et des recommandations.

Dimensionnement des données : Le nœud Table Dimensions paraît être configuré pour définir les dimensions des tableaux générés dans le cadre du reporting.

Pour conclure, ce workflow requiert une modernisation en remplaçant les nœuds obsolètes par leurs équivalents actuels pour assurer une efficacité opérationnelle maximale et une pérennité dans un environnement de données en constante évolution. Une telle mise à jour permettra d'accroître la fiabilité des résultats de recommandation et de reporting générés par ce processus analytique.