**Titre RNCP 39586**

**INGENIEUR EN SCIENCE DES DONNÉES SPECIALISÉ EN** **APPRENTISSAGE AUTOMATIQUE**

**Bloc 2 :**

**ANALYSE, ORGANISER ET VALORISER DES DONNÉES**

Une image contenant dessin humoristique, clipart, conception

Le contenu généré par l’IA peut être incorrect.

**Sujet :**

**Développement d’un outil d’analyse automatisé des commentaires YouTube**

**Auteur :**

**FURTADO LEAL Carla**

Sommaire

[I. Analyse des données 3](#_Toc206437389)

[A. Analyse du besoin 3](#_Toc206437390)

[B. Présentation d’un plan d’analyse 3](#_Toc206437391)

[C. Présentation des requêtes et des résultats sous forme de dashboard 4](#_Toc206437392)

[D. Méthodologie des tests statistiques 4](#_Toc206437393)

[II. Visualisation des données, interprétation et communication des résultats 5](#_Toc206437394)

[A. Visualisation des résultats de l’analyse 5](#_Toc206437395)

[B. Présentation de recommandations 5](#_Toc206437396)

[III. Support et accompagnement des utilisateurs 6](#_Toc206437397)

[A. Support de formation 6](#_Toc206437398)

[B. Documentation technique 6](#_Toc206437399)

# Analyse des données

## Analyse du besoin

*Analyser les besoins métier et les enjeux exprimés par un commanditaire en réalisant des entretiens exploratoires et en récupérant les informations stratégiques nécessaires afin de cadrer le travail d’analyse des données à produire*

*L’analyse du besoin permet d’identifier :*

- Les enjeux et la problématique : développer une méthode / un processus d’analyse qui nécessite le minimum d’intervention possible, le plus automatique possible et ne demande que des ajustements ponctuels. Ce processus doit permettre de tirer de la valeur des commentaires (analyse de sentiment, extractions des sujets récurrents)

- Le contexte : présenter les solutions existantes qui permettent de l’analyse de commentaires (ex : brand25, awario, commentmod, youtube comment finder,

- L’environnement :

- Les contraintes (ex : délai, logistique, coût, technique, réglementaire) :

Créer une infrastructure sécuriser qui permet de collecter, transformer et analyse les données de manière automatique.

## Présentation d’un plan d’analyse

En claire, quelle valeur/informations peut-on tirer des commentaires YouTube, utile pour l’orientation stratégique du client. Soit, en d’autres termes : comment optimiser et appliquer l’analyse de sentiment et le topic modeling de manière automatique en garantissant un certain niveau de performance. Nous classons l’exploitation de ces données en 2 niveaux d’analyse :

* **Niveau 1 : analyse de surface**

Pour ce premier niveau de l’analyse nous exploitons le nombre de commentaires, la date de publication, le nombre de likes associés et la longueur des commentaires. Ces métriques nous permettent d’analyser l’engagement des abonnées dans le temps avec l’évolution du nombre de commentaires et de connaitre le contenu du commentaire avec le plus de like. La longueur du commentaire permet d’afficher le commentaire le plus long. Cela permet de donner un aperçu de l’engagement et les prémisses des résultats du deuxième niveau. Cette analyse de premier niveau se présente sous la forme d’un dashboard. Cette analyse utilise principalement les méta données associées et calculées à partir des commentaires.

* **Niveau 2 : analyse de fond**

L’analyse de fond exploite directement les commentaires pour en qualifier le contenu et son impact de 2 façons :

* + La polarité du sentiment exprimé

L’analyse de sentiment permet de classer les commentaires en 2 ou 3 catégories, positif, négatif et neutre. Ainsi, cet axe d’analyse permet qualifier l’impact du contenu sur les spectateurs. Les résultats seront également présentés sous forme de tableau de bord. Associées au méta données ils permettrons répondre aux questions :

* Il y a -t-il de commentaires positifs, négatifs ou neutre
* A quel catégorie les commentaires les plus liké appartiennent-ils ?
  + Les principaux sujets abordés

Le topic modeling est un type d’apprentissage non supervisé qui se base un ensemble de documents pour en déterminer les principaux sujets. Ainsi les résultats dépendent de l’ensemble. Cette analyse permettra de savoir ce qui est globalement dit dans les commentaires, les principaux sujets abordés. Cet axe d’analyse permet d’identifier les leviers de cet impact. Associées au méta données il sera possible de dire :

* A quel(s) sujet(s) se rapportent les commentaires avec le plus de likes
* Les sujets abordés dans les commentaires classés négatives, positifs, neutres et dans l’ensemble

Étant donnée la nature du modèle, il n’y a pas d’entrainement préalable. De plus, les résultats ne sont pas stockés dans la base de données.

## Présentation des requêtes et des résultats sous forme de dashboard

*Réaliser des requêtes et des calculs en utilisant des outils de dashboarding, des tableurs, des requêtes SQL ou scripts Python afin de produire une analyse des données préalablement collectées. Plusieurs techniques d’analyse sont présentées. (ex : requêtes SQL, Jupyter Notebook, Tableurs, Dashboard) Elles permettent d’obtenir rapidement des résultats justes au regard de la problématique traitée.*

* Les requêtes faites avec pymongo pour obtenir les indicateurs
* Comparaison des data set analyse exploratoire
* Graphique fréquence des mots
* Nuage de mots

Notre outil d’analyse est alimenté par des données stockées dans des bases de données MongoDB compartimentées par chaine YouTube. Ainsi, les analyses menées sont indépendantes les unes des autres mais les méthodes sont identiques.

Par conséquent l’exploitation de ses données se fait via des requêtes Pymongo et des scripts python (Jupyter notebook pendant la phase de développement) via des bibliothèques comme pandas et plotly. Les requêtes Pymongo permettent d’accéder aux données stockées dans les collections pour calculer les indicateurs clés permettant de répondre à la problématique. Les instructions python utilisant les fonctions de plotly et pandas permettent de présenter ces indicateurs sous formes de graphiques pour alimenter le dashboard final. Ce dashboard sera développé sur streamlit \*.

En suivant le plan d’analyse, les indicateurs calculés pour l’analyse de surface sont :

* Graphique de l’évolution du nombre de commentaires au cour du temps
* Nuage des mots, bigramm
* Fréquence des mots

## Méthodologie des tests statistiques

*Élaborer des modèles statistiques et des tests d’hypothèses en modélisant des relations entre les variables, en évaluant la pertinence des résultats des simulations afin de valider ou réfuter des hypothèses.*

*La méthodologie présentée comporte :*

*- La formulation d’une ou plusieurs hypothèses*

*- Le test statistique associé*

*- L’interprétation des résultats*

*Les résultats obtenus par la méthodologie présentée permettent de valider ou réfuter l’hypothèse initiale.*

Nos données ne sont pas étiquetées pour l’analyse de sentiments. Dans ce cas il existe 2 options :

* Utilisé un modèle pré entrainé
* Utiliser des données similaires et entrainer un modèle classique puis l’appliquer à nos données.

Les modèles pré-entrainés sont généralement très performants à condition qu’ils ont été entrainés sur des données similaires, notamment en termes de langue. Or très souvent il s’agit de données en anglais, ce qui qui n’est pas idéal dans notre cas. De plus, l’utilisation de ce type de modèle nous fait perdre en explicabilité, à l’inverse des modèles de classifications classiques. Leur pouvoir explicatif est généralement très grand mais ils demandent une quantité de données important pour être performant.

Afin d’obtenir les meilleures performances nous avons :

* Testés différents modèles prés entrainés sur un large je de données de données labelisées
* Entrainé plusieurs modèles classiques sur le même jeu de données

Cependant cela nécessite de trouver des données suffisamment approchantes de nos commentaires You tube. Cela signifie :

* Langue
* Longueur des commentaires
* Vocabulaire

Nous avons trouvé un ensemble de commentaires Youtube obtenus à partir de l’API, rédigés en anglais et traduits en français \*. Nous avons donc testé si les distributions de la longueur des commentaires étaient significativement différentes grâce au test de **Mann–Whitney U (Wilcoxon rank-sum test). *Explication du test.***

# Visualisation des données, interprétation et communication des résultats

## Visualisation des résultats de l’analyse

*Représenter les données en choisissant les modèles de représentation les plus adaptés (ex : histogramme, Heat map, nuage de points) et en utilisant des outils de représentation adaptés (ex : Office, power BI) afin de permettre la compréhension et l’exploitation des données par le public visé.*

*Le choix des outils de mise en forme et des représentations est justifié avec les bénéfices attendus (ex : lisibilité, facilité d’utilisation) La mise en forme (ex : couleurs, légendes, échelles, graphiques) permet de communiquer les données au public ciblé avec clarté et justesse. Les choix de mise en forme prennent en compte les spécificités des personnes en situation de handicap le cas échéant (ex : formes, contrastes, choix des couleurs)*

Présentation de streamlit, des tableau de bord en justifiant les choix des graphique, des types de visualisation avec les textes, couleurs, organisation …

## Présentation de recommandations

Présenter des recommandations, en préparant son discours et des arguments, en structurant son analyse sur les données représentées afin d’aider les décideurs à établir leurs stratégies. La présentation des recommandations est structurée, synthétique et argumentée. Les recommandations permettent d’éclairer le commanditaire pour l’aider dans sa prise de décision.

# Support et accompagnement des utilisateurs

## Support de formation

*Former les utilisateurs à l’utilisation des données et des outils de visualisation en analysant le besoin de montée en compétences et en élaborant des supports de formation et de sensibilisation adaptés afin de permettre aux utilisateurs de maitriser l’exploitation des données.*

*L’enjeu et le sujet de la formation ou de la sensibilisation sont présentés. Le support est adapté au sujet et permet de monter en compétences le public visé (ex : présentation power point, newsletter, mail de bonnes pratiques…)*

Support de formation à l’utilisation de streamlit

Mise en garde sur comment utiliser les données (commentaires youtubes et méta données associés saut pseudo) car ce sont des données sensibles, expliquer les enjeux liés

Former sur le traitement du texte en vu d’analyse

Ce support de formation, présente les règles et mode d’utilisation de l’application « YOU REVIEW ». Ce document s’adresse à tous les personnes autorisées à utiliser l’application. C’est-à-dire toute personnes ayant un chaîne YouTube et possédant un identifiant et mot de passe permettant d’utiliser l’application.

Si vous utilisez cette application vous vous engagez à respecter les règles et les conditions d’utilisation de l’outil et des données de You Tube.

*Les données*

*L’outil*

*Les fonctionnalités*

## Documentation technique

*Rédiger la documentation technique d'utilisation du système d’analyse de données en identifiant le public concerné, en détaillant le fonctionnement du système d’analyse de données afin d'assurer la traçabilité et la transmission aux utilisateurs.*

*La documentation technique comporte :*

*- La description des sources de données (ex : origine, périmètre)*

*- La description des méthodes de calculs*

*- La description technique et fonctionnelle des indicateurs*

*La documentation permet la compréhension, la transmission et la reproductibilité de l’analyse de données.*

Data scientist / data engineer

* Sources des données : api youtube + kaggle (périmètre analyse textuelle)
* Méthodes de calculs : vectorisation, analyse de sentiment, topic modeling
* Description des indicateurs :
* Architecture du projet
* Architecture de sécurité
* Détail des techno : docker, prefect, streamlit