UFR des Mathématiques et Informatique

Master 2 MLDS - FA

Projet en Business Intelligence

Réalisé par:

Lilia HARIRECHE

2020 - 2021

Sommaire

I. Introduction

Dans le cadre de notre M2 MLDS, ce projet en Business Intelligence nous a été affecté.

La “Business Intelligence” désigne un ensemble de méthodes, de moyens et d'outils informatiques utilisés pour piloter une entreprise et aider à la prise de décision : tableaux de bord, rapports analytiques et prospectifs.

L’objectif de ce projet est de réaliser un tableau de bord contenant une analyse descriptive détaillée des bases de données créées à partir d’un fichier texte qui contient les informations de plusieurs articles.

Pour cela, nous allons présenter les différentes étapes faites durant ce projet, plus exactement les différents traitements de notre jeu de données, la création des data frames, la construction des différents graphes, ainsi que la réalisation d’ un consensus entre les partitions découvertes par le clustering.

II. Présentation et traitement du jeu de données

Dans ce projet, nous avons utilisé un fichier texte contenant des articles. Chaque article contient les informations suivantes: Id, titre, résumé, citations, auteurs, année, venue.

Ces informations sont des données non structurées, elles sont représentées sans format prédéfini et contiennent des caractères spéciaux comme le montre la figure ci-dessous.

```
#*Formal models for expert finding in enterprise corpora.
#@Krisztian Balog,Leif Azzopardi,Maarten de Rijke
#t2006
#cSIGIR
#index594377
#%595386
#%362694
#%772628
#%595551
#%26506
#%594777
#%935966
#%121844
#%596024
#%95047
#%595671
#!Searching an organization's document repositories for experts provides

#*Latent Semantic Indexing is an Optimal Special Case of Multidimensional
#@Brian T. Bartell,Garrison W. Cottrell,Richard K. Belew
#t1992
#cSIGIR
#index594378
#%771904
```

Pour avoir des données structurées que nous pourrions exploiter pour la suite du projet, nous avons d’abord effectué des traitements sur les différentes informations de chaque article de ce fichier texte. Et en raison des différents problèmes rencontrés lors du chargement des dataframes sous l’outil Qlik Sense, nous avons choisi de ne traiter que 3000 articles. Nous avons ainsi obtenu le data frame (***df_articles***) suivant:

Venue	Year	Authors	Title	Id	Abstract	ListCitation	NbrAuthors
SIGIR	2006	krisztian balog,leif	formal model for ex	594377	searching organiz	595386	3
SIGIR	1992	brian t. bartell,garr	latent semantic inde	594378	latent semantic i	362694	3
SIGIR	2000	rie kubota ando	latent semantic-spa	594379	we present nove	772628	1
SIGIR	1994	brian t. bartell,garr	automatic combinat	594380	searching online	595551	3
SIGIR	1988	christine barthes,p	planning in an expe	594381	in poster, presen	26506	2
SIGIR	1999	peter g. anick,sures	the paraphrase sear	594382	a data organizatic	594777	2

III. Création des matrices et des dataframes correspondants

Après avoir obtenu notre jeu de données (df_articles), nous avons créé trois matrices: co-terme title, co-terme abstract, et document-auteur que nous avons, respectivement, nommé dans notre code: doc_term_titre, doc_term_abstract et doc_auteur. Ainsi nous avons obtenu nos trois data frames: df_doc_term_titre, df_doc_term_abstract et df_doc_auteur ci-dessous.

Id	1	2	3	4	5
594377	0	0	0	0	0
594378	0	0	0	0	0
594379	0	0	0	0	0
594380	0	0	0	0	0

df_doc_term_titre

Id	1	2	3	4
594377	0	0	0	0
594378	0	0	0	0
594379	0	0	0	0
594380	0	0	0	0
594381	0	0	0	0
594382	0	0	0	0
594383	0	0	0	0
594384	0	2	0	0
594385	1	0	0	0

df_doc_term_abstract

Id	1	2	3	4
594377	1.0	1.0	1.0	1.0
594378	1.0	1.0	1.0	1.0
594379	1.0	1.0	1.0	1.0
594380	1.0	1.0	1.0	1.0
594381	1.0	1.0	1.0	1.0

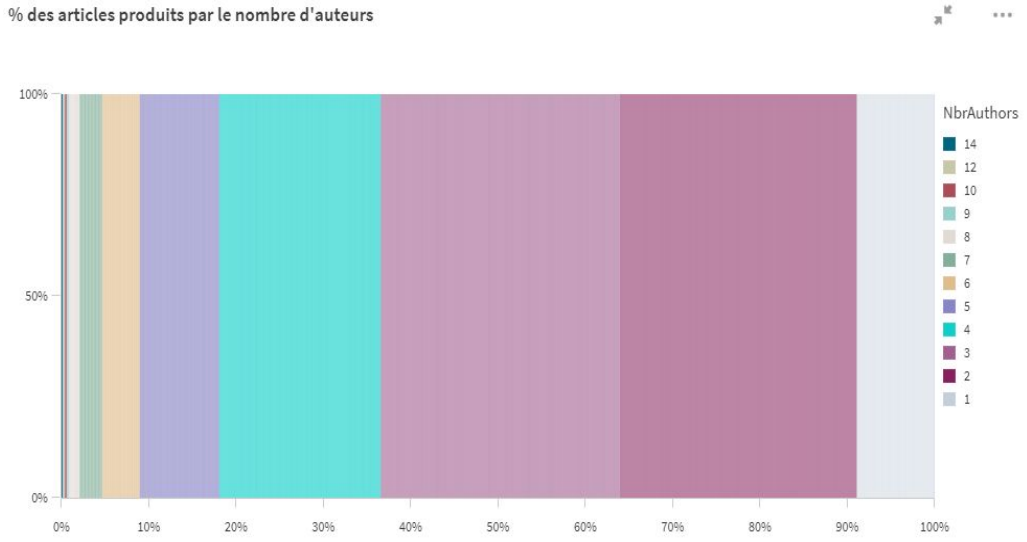
df_doc_auteur

IV. Analyse descriptive détaillée des bases de données créées

Afin d'effectuer une analyse des différents jeux de données obtenus précédemment, nous avons réalisé un tableau de bord contenant une analyse descriptive détaillée.
Pour cela, nous avons opté pour l'utilisation de ***Qlik sense***, une plateforme d'analyse de données de bout en bout qui sert de référence à une nouvelle génération d'analytics.

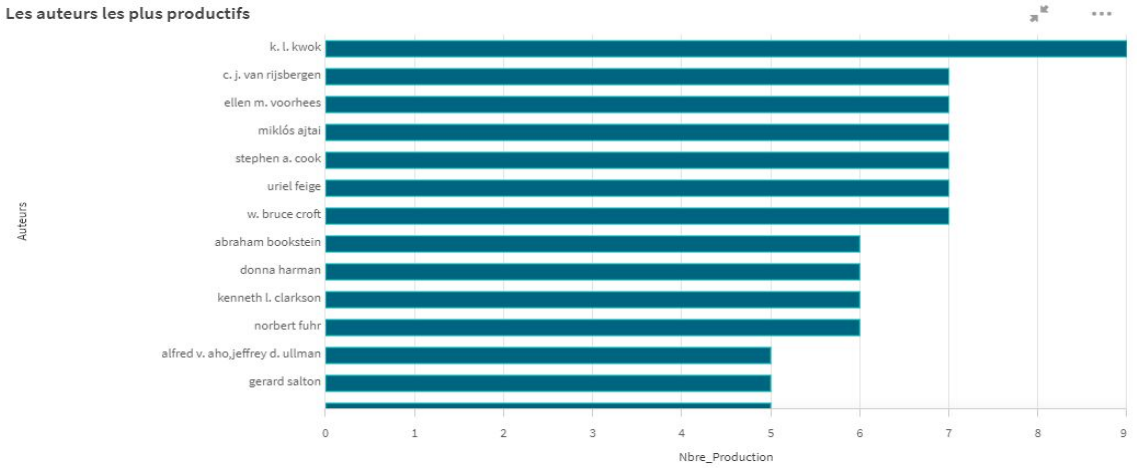
(Les résultats obtenus sont dans le fichier PDF accompagné avec ce rapport.)

- **Quelques exemples des résultats obtenus:**



Nous avons représenté dans cette analyse le nombre d'article en % produits par le nombre d'auteurs, par exemple on peut dire que:

- *Environ 25% des articles ont été produits par 2 auteurs.*
- *Environ 10 % des articles ont été produits par un seul auteur.*



Dans cette analyse, nous avons visualisé le nombre de production des auteurs.

- *L'auteur le plus productif est K. l.Kwok, avec 9 productions d'articles.*

V. Réalisation d'un clustering à l'aide de l'algorithme de Louvain et un consensus entre les partitions

La méthode de Louvain est un algorithme hiérarchique d'extraction de communautés applicable à de grands réseaux.
Elle permet d'effectuer le partitionnement d'un réseau en optimisant ***la modularité***, une valeur comprise entre -1 et 1 qui mesure la densité d'arêtes à l'intérieur des communautés comparée à celle des arêtes reliant les communautés entre elles. L'optimisation de la modularité conduit théoriquement au meilleur partitionnement possible.

Nous avons aussi réalisé un consensus entre les partitions découvertes par les clustering obtenus avec l'algorithme de Louvain.

(Les résultats obtenus sont dans le fichier PDF accompagné avec ce rapport.)

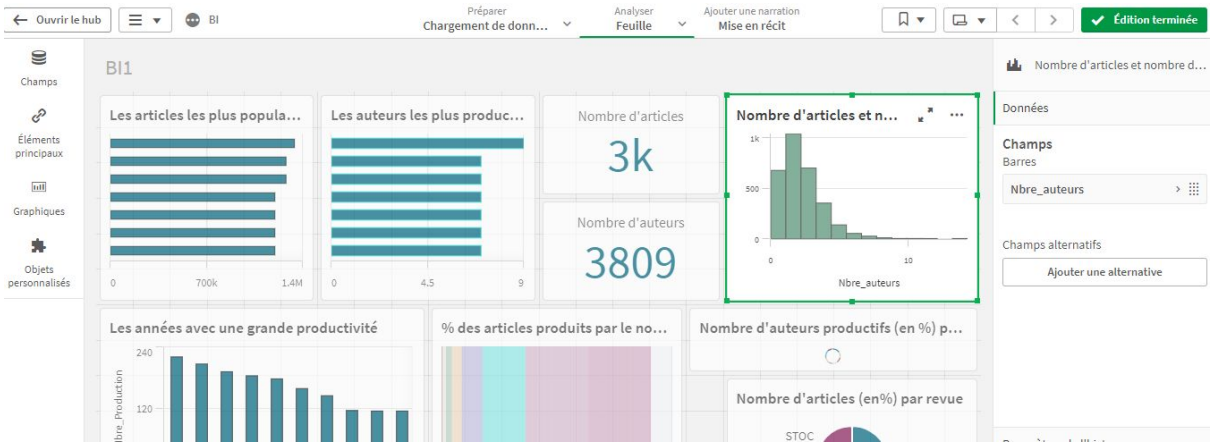
VI. Les outils utilisés

1. Langage de programmation "Python":

Python est l'un des langages de programmation les plus utilisés en Data Science et Machine Learning.
Nous avons décidé de l'utiliser par rapport aux différentes bibliothèques et packages proposés par ce langage, comme: *"matplotlib", "wordcloud", "textblob", "sklearn.feature_extraction.text".*

2. Qlik Sense:

Afin d'élaborer un tableau de bord et effectuer une visualisation des données, nous avons décidé d'utiliser la plateforme Qlik Sense.



La plateforme de Qlik Sense

Grâce à son intelligence artificielle sophistiquée et sa plateforme de cloud computing haute performance, nous pouvons prendre de meilleures décisions au quotidien.

La plateforme permet de combiner et charger les données, créer des visualisations intelligentes, puis créer des applications d'analyse complètes, accélérées par la génération de suggestions et l'automatisation.

VII. Conclusion

Durant ce projet nous avons pu effectuer des traitements sur un fichier texte afin de construire nos data frames utilisées dans notre étude.

Nous avons réalisé un clustering à l'aide de l'algorithme de Louvain, aussi un consensus des partitions découvertes, ainsi des graphes correspondants à nos résultats.

Afin de visualiser nos différents résultats, nous avons opté pour l'utilisation de la plateforme Qlik sense, ce qui nous a permis de nous familiariser avec cette dernière.

Malgré les différents problèmes rencontrés lors du traitement du fichier texte ou encore le chargement des données sur l'outil Qlik Sense, vu le grand volume des données et l'incapacité des machines pour effectuer les différents traitements, nous avons quand même pu réaliser notre étude en réduisant le nombre d'articles étudiés.

Bibliographie

Informatique décisionnelle:
https://www.futura-sciences.com/tech/definitions/informatique-informatique-decisionnelle-15057/

Qlik Sense:
https://www.qlik.com/fr-fr/products/qlik-sense#:~:text=Qlik%20Sense%20est%20une%20plateforme%20d'analyse%20de%20donn%C3%A9es%20compl%C3%A8te,enti%C3%A8rement%20vos%20solutions%20d'analyse.

https://www.youtube.com/watch?v=zs24DVVIALU

Méthode de Louvain:
https://fr.wikipedia.org/wiki/M%C3%A9thode_de_Louvain
https://neo4j.com/docs/graph-algorithms/current/algorithms/louvain/