Master 2 Machine Learning for Data Science

Projet : Business Intelligence

Réalisé par:

Nadia RADOUANI – 21911973 FI

Enseignante: Mme Rafika BOUTALBI

Année universitaire: 2020-2021

Table des matières

Table des figures

I. Introduction

1. Objectif

Ce projet consiste à traiter et analyser un fichier texte contenant des informations sur des articles scientifiques parus dans des revues et conférences.

Le but est d'extraire les informations sous différents critères, réaliser un clustering à l'aide de l'algorithme de Louvain, puis un consensus entre les partitions découvertes, et enfin une visualisation sur Qlik sense dans le but de faire une analyse descriptive de notre base et une analyse des clusters obtenus.

2. Données

Les informations sur les articles comprennent comprennent : Le titre de l'article, le ou les auteurs, l'année de publication, nom de la revue (ou de la conférence), les citations entre articles et le résumé de l'article.

```
#*Improved Channel Routing by Via Minimization and Shifting.
#@Chung-Kuan Cheng
David N. Deutsch
#t1988
#cDAC
#index131751
#%133716
#%133521
#%134343
#!Channel routing area improvement by means of via minimization
and via shifting in two dimensions (compaction) is readily
achievable. Routing feature area can be minimized by wire
straightening. The implementation of algorithms for each of
these procedures has produced a solution for Deutsch's Difficult
Example
 the standard channel routing benchmark
 that is more than 5% smaller than the best result published
heretofore. Suggestions for possible future work are also given.
```

Figure 1 - Aperçu de mes données

II. Prétraitement des données

Un traitement de texte a été réalisé afin d'extraire les informations séparément sous forme d'un dataframe.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 596278 entries, 0 to 596277
Data columns (total 9 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   Title           596278 non-null  object
 1   authors         596278 non-null  object
 2   Venue           596278 non-null  object
 3   Year            596278 non-null  object
 4   ID              596278 non-null  object
 5   ListeCitations  596278 non-null  object
 6   NBRCitations    596278 non-null  int64
 7   abstract        596278 non-null  object
 8   NBRAuthors      596278 non-null  int64
dtypes: int64(2), object(7)
memory usage: 40.9+ MB
```

Figure 2 - Informations sur ma data

Un filtrage sur les revues : SIGMOD Conference, STOC et SIGIR, a été effectué en prenant de chaque revue 500 articles, construisant ainsi une base de 1500 articles sur laquelle nous avons travaillé le projet. Il était aussi nécessaire d'effectuer un nettoyage des valeurs des manquantes.

III. Construction des matrices

1. Matrice documents termes

- **Titres**

Pour cette matrice, nous avons utilisé le champ *Title* de notre dataframe. Nous avons appliqué *CountVectorizer* sous python, et avons limité le nombre des termes à 1000, et pour pouvoir faire des associations, nous avons ajouté une colonne supplémentaire contenant les indexes des articles.

(888, 1000)

ID	abstract	access	accessing	acm	action	active	ad	adapting	adaptive	addition	aggregate	aggr
599006	0	0	0	0	0	0	0	0	0	0	0	
598910	0	0	0	0	0	0	0	0	0	0	0	
599628	0	0	0	0	0	0	0	0	0	0	0	
598447	0	0	0	0	0	0	0	0	0	0	0	
598933	0	0	0	0	0	0	0	0	0	0	1	

5 rows × 1000 columns

Figure 3 - Aperçu de la matrice Doc-terms pour les titres

- **Abstract**

Même traitement que nous avons fait précédemment pour obtenir la matrice doc-terms pour les titres, a été effectué sur le champ *Abstract* de notre dataframe afin d'obtenir la matrice doc-terms pour les résumés.

(888, 1500)

ID	ability	able	abstract	acceptable	access	accessed	according	account	accuracy	accurate	achi
599006	0	0	0	0	0	0	0	0	0	0	
598910	0	0	0	0	0	0	0	0	0	0	
599628	0	0	0	0	3	0	0	0	0	0	
598447	0	0	0	0	0	0	0	0	0	0	
598933	0	0	0	0	0	0	0	0	0	0	

5 rows × 1500 columns

Figure 4 - Aperçu de la matrice Doc-terms pour les abstracts

2. Matrice documents auteurs

Cette matrice contient la liste des auteurs avec tous les indexes des articles.

ID	raghav kaushik	philip bohannon	jeffrey f. naughton	henry f. korth	tomasz imielinski	witold lipski jr.	arnon rosenthal	david s. reiner	charles d. cranor	theodore johnson	c spats
599006	1	1	1	1	0	0	0	0	0	0	
598910	0	0	0	0	1	1	0	0	0	0	
599628	0	0	0	0	0	0	1	1	0	0	
598447	0	0	0	0	0	0	0	0	1	1	
598933	0	0	0	0	0	0	0	0	0	0	

5 rows × 1732 columns

Figure 5 - Aperçu de la matrice Doc-Authors

Par ailleurs, nous avons crée une dataframe contenant la liste des auteurs, et nous avons attribué à chaque auteur un *id_authors*.

(1732, 2)

	authors	id_authors
0	raghav kaushik	0
1	philip bohannon	1
2	jeffrey f. naughton	2
3	henry f. korth	3
4	tomasz imielinski	4
5	witold lipski jr.	5
6	arnon rosenthal	6
7	david s. reiner	7
8	charles d. cranor	8
9	theodore johnson	9

Figure 6 - Aperçu de la dataframe Authors

3. Matrice document document

La matrice document document est une matrice binaire pour deux documents i et j si i est cité dans j.

ID	1060189	642616	622597	304160	300784	623077	2294	642352	642069	643719	597570	627554	1058
599006	1	1	1	1	1	1	1	1	1	1	1	1	
598910	0	0	0	0	0	0	0	0	0	0	0	0	
599628	0	0	0	0	0	0	0	0	0	0	0	0	
598447	0	0	0	0	0	0	0	0	0	0	0	0	
598933	0	0	0	0	0	0	0	0	0	0	0	0	

5 rows × 6498 columns

Figure 7 - Aperçu de la matrice Doc-Doc

IV. Clustering

Dans cette partie, nous avons fait du clustering sur les matrices construites précédemment, nous avons utilisé l'algorithme de Louvain. Le but est de déterminer des clusters des différents termes utilisés dans les titres et les abstracts, des clusters des auteurs et enfin des clusters des citations (documents).

1. Titres

Pour les titres, nous avons obtenu 13 clusters des termes utilisés avec une modularité de 0.46.

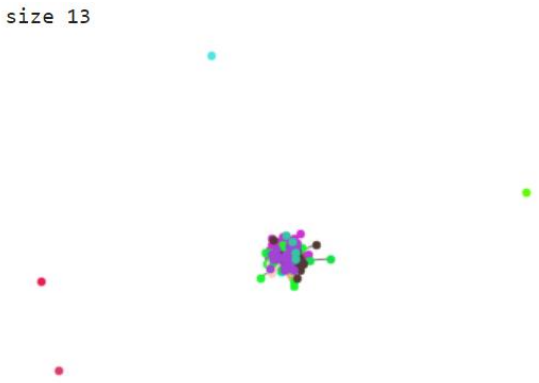

Figure 8 - Graphe Co-terms pour les titres

2. Abstract

L'algorithme de Louvain nous a permis d'avoir quatre clusters pour les abstracts avec une modularité de 0.18.

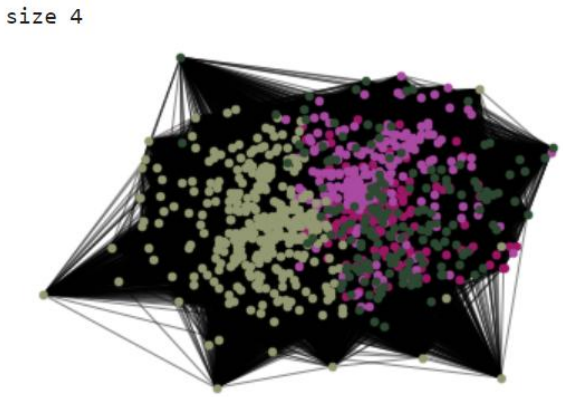

Figure 9 - Graphe Co-terms pour les abstracts

3. Auteurs

Pour les auteurs, nous avons obtenu 447 clusters, avec une modularité de 0.97.

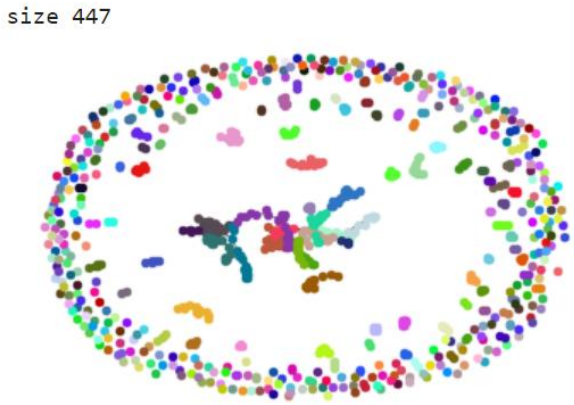

Figure 10 - Graphe Co-Authors

4. Citations

Enfin pour les citations, nous avons obtenu 208 clusters avec une modularité de 0.82.

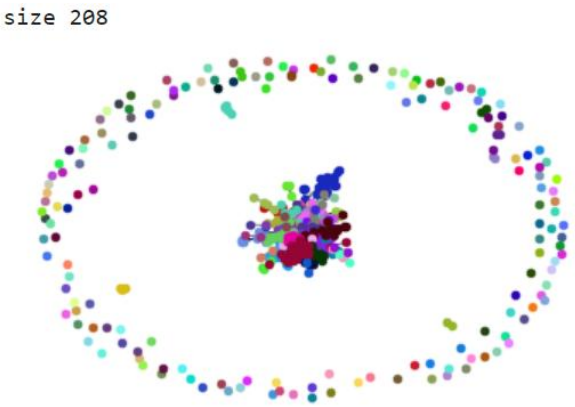

Figure 11 – Graphe Citations

V. Consensus

Dans cette partie, nous avons réalisé un consensus entre les partitions découvertes par les clustering précédents en utilisant le package Cluster_Ensembles sous python, nous avons limité le nombre de clusters à 16.

```
consensus_clustering_labels [ 9  0  9  4  1  0 11  2  3  0  0  8  3  0  0  3  2  4  9  5  9 11 15  1
  2  8  1  9  3  2  4  9  0 11  3  1  8  0  3  0  2  9 10  2  2  1  0 14
  1  5  2  1  1  1  1  0  2  4  3  4  2  8  1  0  1  0  4  4  9  1  9  3
  3  0  2  1  3  9  1 14  9  0 11  3  2  9  2  0  0  8  2  1  3  4  3  8
 11  8  6 14  5  0  1  2  0  2  2  0  2 11  2  0  8  4 11  2 13  4  3  3
 11  8  1  2  9  2  2  1  1  3  3  2  1 10  1  4  8  0  2  0  1  9  8  4
  9  8  9  8  7  2  3  4  1  4  0  6  1  9  8  8  9 10  3  9  8  0  2  0
  1  9  3  3  1  2 14  1 11  0  9  2  2  2  4  0  2  0  3  8  2  4  1  2
  8  9  8 11  3 11  1  1  3  0  0  0  1 15 15  8  0  8  9 11  9  6  9  0
 10  0  0  3  9 11  7  4  8  2  3  2  4  1 11  0  9  9  4  1  8  1  4  4
  4  0  3  9  4  3 14  3  9  9  5  8 10  1  1 14 11  0  3 11  4  2  0  9
  1  1  2  6  0  2  1  2  3  4  2  4  0  0  1 10  1  1 11  5  0  1 11  2
 11 11  0  1  0  8  2  9  1  4 11  5 11  0 11  0  9 10  0  4  1 10  4  9
```

Figure 12 - Aperçu du consensus

VI. Visualisation

1. Modèle de données

Nous avons importé les données sur Qlik Sense, ensuite nous avons crée un modèle des données où il y a une association entre les différentes dataframe par l'ID des articles.

Figure 13 - Modèle de données

2. Tableaux de bord

Nous avons construit deux tableaux de bord : un premier pour l'analyse descriptive de nos dataframe et un deuxième pour l'analyse des clusters.

Ci-dessous quelques graphes de ces tableaux de bord :

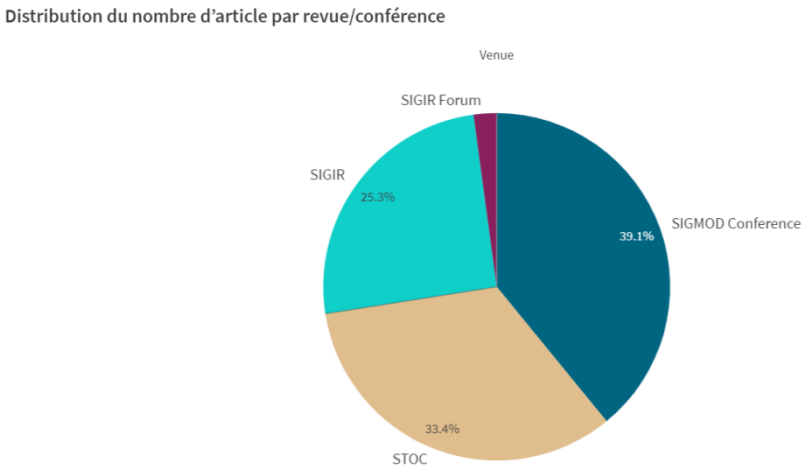

Figure 14 - Distribution du nombre d'articles par revue

Les articles appartiennent majoritairement à la revue SIGMOD Conference.

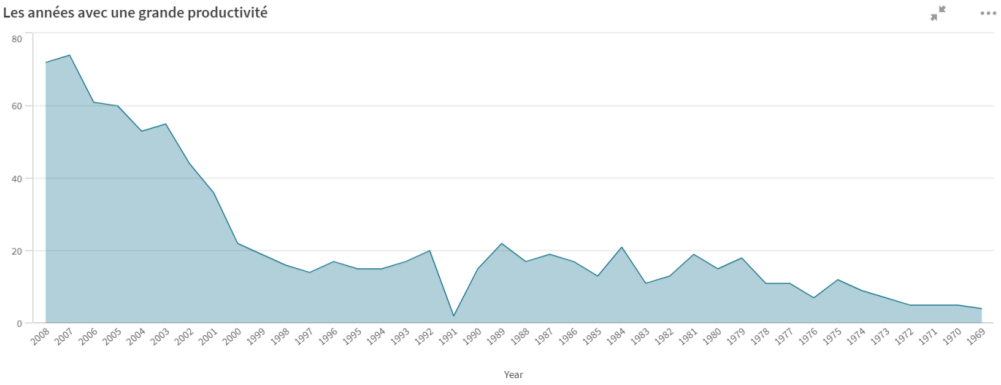

Figure 15 - Les années avec une grande productivité

L'année avec une grande productivité est 2008.

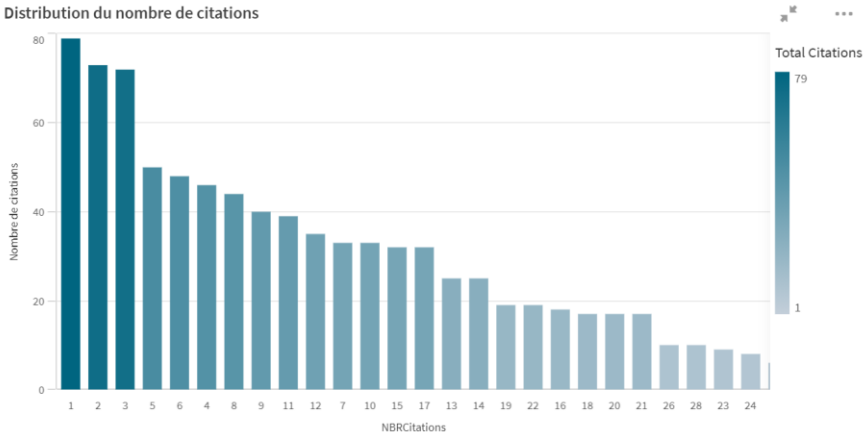

Figure 16 - Distribution du nombre de citations

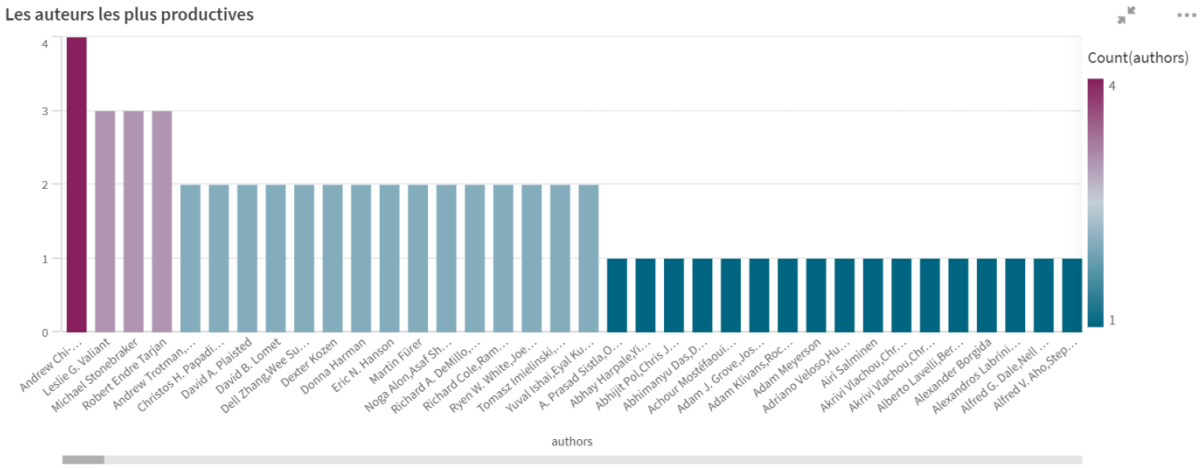

Figure 17 - Les auteurs les plus productifs

L'auteur le plus productif est Andrew.

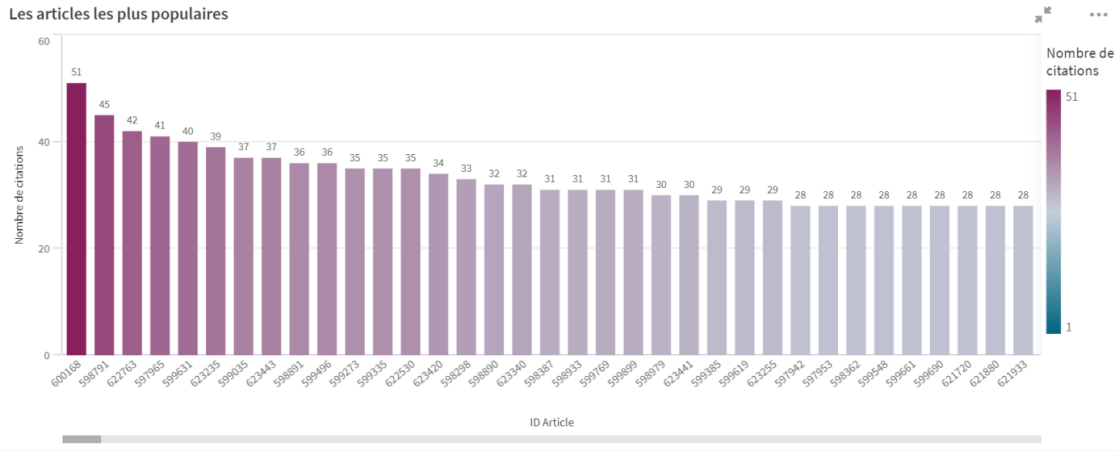

Figure 18 - Les articles les plus populaires

L'article le plus populaire est celui avec l'ID 600168.

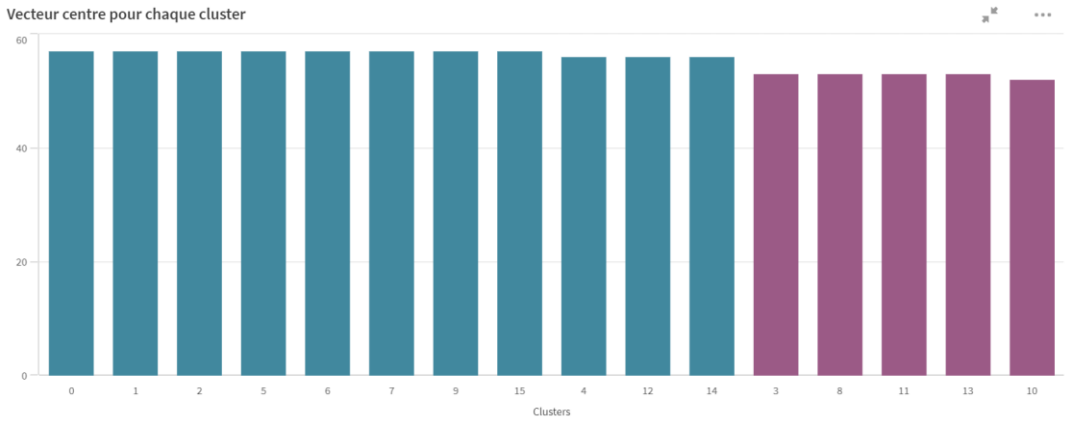

Figure 19 - Vecteur centre pour chaque cluster

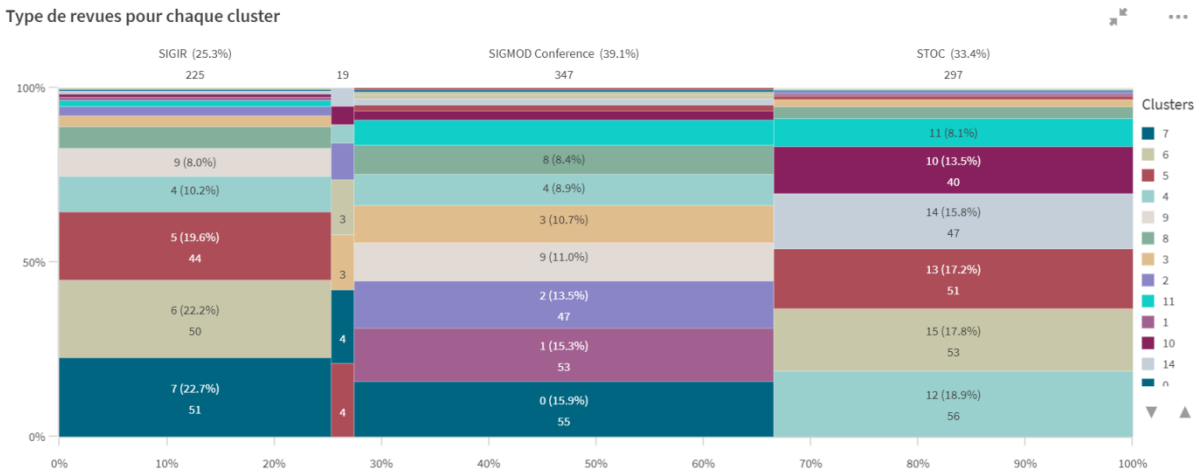

Figure 20 - Type de revue pour chaque cluster

VII. Conclusion

Ce projet nous a donné la possibilité de découvrir l'utilité de Qlik sense et de visualiser quelques résultats du clustering et faire une analyse descriptive de nos données.