Projet : Business Intelligence

Nous disposons d'un fichier texte contenant des informations sur des articles scientifiques parus dans des revues et conférences. Ces informations comprennent :

- Le titre de l'article,
- Le ou les auteurs,
- L'années de publication,
- Nom de la revue (ou de la conférence)
- Les citations entre articles

Partie 1:

1) Télécharger le jeu de données (fichier DBLP_Subset.csv sur le lien suivant http://up5.fr/yMZbV). Le jeu de données est un fichier texte :

```
#* --- paperTitle
#@ --- Authors
#t ---- Year
#c  --- publication venue
#index 00---- index id of this paper
#% ---- the id of references of this paper (there are multiple lines, with each indicating a reference)
#! --- Abstract

The following is an example:

#*Information geometry of U-Boost and Bregman divergence
#@Noboru Murata,Takashi Takenouchi,Takafumi Kanamori,Shinto Eguchi
#t2004
#cNeural Computation
#index436405
#%94584
#%282290
#%605546
#%620759
#%564877
#%564235
#%594837
#%479177
#%586607
#!We aim at an extension of AdaBoost to U-Boost, in the paradigm to build a stronger classification machine from a set of weak learning machines.
```

Chaque information est introduite par un caractère spécifique comme dans l'exemple ci-dessus.

2) Charger le fichier texte sous R
3) Réaliser les traitements de données (traitement de texte) afin d'extraire les informations séparément sous forme d'un dataframe (figure ci-dessous).
4) Construire la matrice Documents-termes pour les titres des articles (utiliser le package *tm*). Ajouter une colonne 'Id' représentant l'identifiant du document à la matrice Doc-termes. Cette colonne nous permettra de relier ce tableau au dataframe contenant les informations des articles.
5) Créer un dataframe représentant les auteurs et attribuer à chaque auteur un identifiant. Ajouter une nouvelle variable 'id_auteur' au dataframe contenant les informations des articles.

Ces opérations sont nécessaires afin de pouvoir lier ces informations au niveau de l'outil « Qlik Sense ».

	Venue	Year	Authers	Title	NbrAuther	Id	ListCitations	NbrCitaitons
1	#cSIAM Journal of Applied Mathematics	2005	,Christian Schmeiser,Yasmin Dolak	The Keller-Segel Model with Logistic Sensitivity Func...	2	1057374		0
2	#cSIAM Journal of Applied Mathematics	1977	,David S. Johnson,M. R. Garey	The Rectilinear Steiner Tree Problem in NP Complete.	2	1057375		0
3	#cSIAM Journal of Applied Mathematics	2006	,F. O'Doherty,James P. Gleeson	Non-Lorentzian Spectral Lineshapes Near a Hopf Bif...	2	1057376		0
4	#cSIAM Journal of Applied Mathematics	1969	,Ronald L. Graham	Bounds on Multiprocessing Timing Anomalies.	1	1057377		0
5	#cSIAM Journal of Applied Mathematics	1995	,Blake Temple,Eli Isaacson	Convergence of the 2×2 Godunov Method f...	2	1057378		0
6	#cSIAM Journal of Applied Mathematics	1995	,Antonin Chambolle	Image Segmentation by Variational Methods: Mumf...	1	1057379		0
7	#cSIAM Journal of Applied Mathematics	1995	,Vijay K. Samalam,Thomas M. Chen	Time-Dependent Behavior of Fluid Buffer Models wit...	2	1057380		0
8	#cSIAM Journal of Applied Mathematics	1995	,David A. Edwards	Constant Front Speed in Weakly Diffusive Non-Fickia...	1	1057381		0
9	#cSIAM Journal of Applied Mathematics	1995	,David R. Kassoy,Meng Wang	Nonlinear Oscillations in a Resonant Gas Column: A...	2	1057382		0
10	#cSIAM Journal of Applied Mathematics	1995	,Eric Henderson,James Vesenka,Richard Miller	Tip Reconstruction for the Atomic Force Microscope.	3	1057383		0
11	#cSIAM Journal of Applied Mathematics	1995	,Robert T. Tranquillo,Richard B. Dickinson	Transport Equations and Indices for Random and Bia...	2	1057384		0
12	#cSIAM Journal of Applied Mathematics	1995	,Michael Landman,Nancy Kopell	Spatial Structure of the Focusing Singularity of the ...	2	1057385		0
13	#cSIAM Journal of Applied Mathematics	1995	,P. A. Martin,C. J. Luke	Fluid-Solid Interaction: Acoustic Scattering by a Smo...	2	1057386		0
14	#cSIAM Journal of Applied Mathematics	1995	,Zbigniew Galias	On a Discrete-Time Nonlinear System Associated wit...	1	1057387		0
15	#cSIAM Journal of Applied Mathematics	1995	,Michael Vogelius,Fadil Santosa	Erratum to the Paper: First-Order Corrections to the ...	2	1057388		0
16	#cSIAM Journal of Applied Mathematics	1995	,Robert Almgren	Crystalline Saffman-Taylor Fingers.	1	1057389		0
17	#cSIAM Journal of Applied Mathematics	1995	,Stanley Ocken	Recognizing Convergent Orbits of Discrete Dynamic...	1	1057390		0
18	#cSIAM Journal of Applied Mathematics	1995	,Tzy-Wei Huang,Sze-Bi Hsu	Global Stability for a Class of Predator-Prey Systems.	2	1057391		0
19	#cSIAM Journal of Applied Mathematics	1995	,Aslak Tveito,Nils Henrik Risebro,Helge Holden	Maximum Principles for a Class of Conservation Laws.	3	1057392		0
20	#cSIAM Journal of Applied Mathematics	1995	,R. H. Tew,John R. Ockendon,J. R. King,S. J. Chapman...	Stokes Phenomenon and Matched Asymptotic Expa...	5	1057393		0
21	#cSIAM Journal of Applied Mathematics	1995	,Amitabha Bose	Symmetric and Antisymmetric Pulses in Parallel Cou...	1	1057394		0
22	#cSIAM Journal of Applied Mathematics	1995	,Wolfgang Ring	Identification of a Core from Boundary Data.	1	1057395		0
23	#cSIAM Journal of Applied Mathematics	1995	,Brian T. R. Wetton,John M. Stockie	Stability Analysis for the Immersed Fiber Problem.	2	1057396		0
24	#cSIAM Journal of Applied Mathematics	1995	,Vladimir A. Sharafutdinov	On Emission Tomography of Inhomogeneous Media.	1	1057397		0
25	#cSIAM Journal of Applied Mathematics	2005	,George Christakos,Hwa-Lung Yu	Porous Media Upscaling in Terms of Mathematical E...	2	1057398		0
26	#cSIAM Journal of Applied Mathematics	1995	,E. A. Spiegel,L. N. Howard,N. J. Balmforth	Instability of Rapidly Rotating Polytropes.	3	1057399		0
27	#cSIAM Journal of Applied Mathematics	1995	,Phoebus Rosakis	An Equal Area Rule for Dissipative Kinetics of Propa...	1	1057400		0
28	#cSIAM Journal of Applied Mathematics	1995	,Thomas P. Witelski,Andrew B. White Jr.,Donald S. Co...	Shock Formation in a Multidimensional Viscoelastic ...	3	1057401		0

6) Une fois ces trois ‘dataframe’ créés :
    - Un ‘dataframe’ contenant les informations articles
    - Un ‘dataframe’ contenant la matrice documents-termes
    - Un ‘dataframe’ contenant les identifiants des auteurs

Générer les fichiers CSV pour ces tableaux de données et les charger sur une application Qlik Sense.

Partie 2:

7) Réaliser un tableau de bord contenant une analyse descriptive détaillée de la base de données créée :
    - Les auteurs les plus productives
    - Les articles les plus populaires
    - Les années avec une grande productivité
    - Distribution du nombre d’article par revue/conférence
    - Distribution du nombre de citation
    - Les termes les plus utilisés

Vous pouvez proposer une multitude d’analyses en croisant ces différentes variables !

Partie 3:

8) Sélectionner les trois ‘Venues’ suivante SIGIR, SIGMOD et STOC.
9) Réaliser un clustering à l’aide d’un k-means, enregistrer les résultats sous format CSV de sorte à pouvoir faire le lien avec les fichiers précédemment chargés.
10) Réaliser une détection de communautés sur les matrices Citations ainsi que la matrice Co-auteurs à l’aide de la fonction “fastgreedy.community” du package R “Igraph”.
11) Créer un second tableau de bord permettant d’analyser les résultats du clustering :
    - Les termes les plus fréquents pour chaque clusters
    - Le type de revues pour chaque cluster …
    - Le Vecteur centre (moyen) pour chaque cluster
    - Insérer les figures de graphes obtenus par la détection de communautés.

Partie 4:

12) Refaire la partie 3 en considérant toutes les “Venues”.

Toute autre proposition est la bienvenue (d’autres algorithmes de clustering, nuages de mots, ...)!

Le projet est à remettre le 10 Novembre 2019 (à 23h59) au plus tard. Un fichier compressé contenant, le code R, les feuilles en format PDF obtenu à partir de Qlick Sense ainsi qu’un rapport de 10 pages.

Bonus : Vous pouvez refaire le même traitement en utilisant les ‘Abstract’ au lieu des titres et créer un troisième tableau de bord pour analyser les résultats obtenus !

Important : Faire attention aux noms données aux différentes variables. Qlik Sense ne réalise de jointure, sauf si deux colonnes ont le même nom.