PARIS DESCARTES UNIVERSITY

BUISNESS INTELLIGENCE

Projet Traitement de texte

Étudiant :
Ilyes ZEMALACHE

Professeur :
Rafika Boutalbi

02 Novembre 2019

Table des matières

Résumé

Sur cette étude on va passer par plusieurs étapes, tout d'abord on va lire le fichier, et bien l'organiser dans un data Frame, ensuite faire le nettoyage du text, afin d'appliquer les fonctions nécessaires pour l'obtention de la matrice doc termes, on appliquera aussi un skmeans sur une partie de donnée, et le tout sera résumé dans un tableau de bord de l'outil Qlik Sense.

1 Introduction

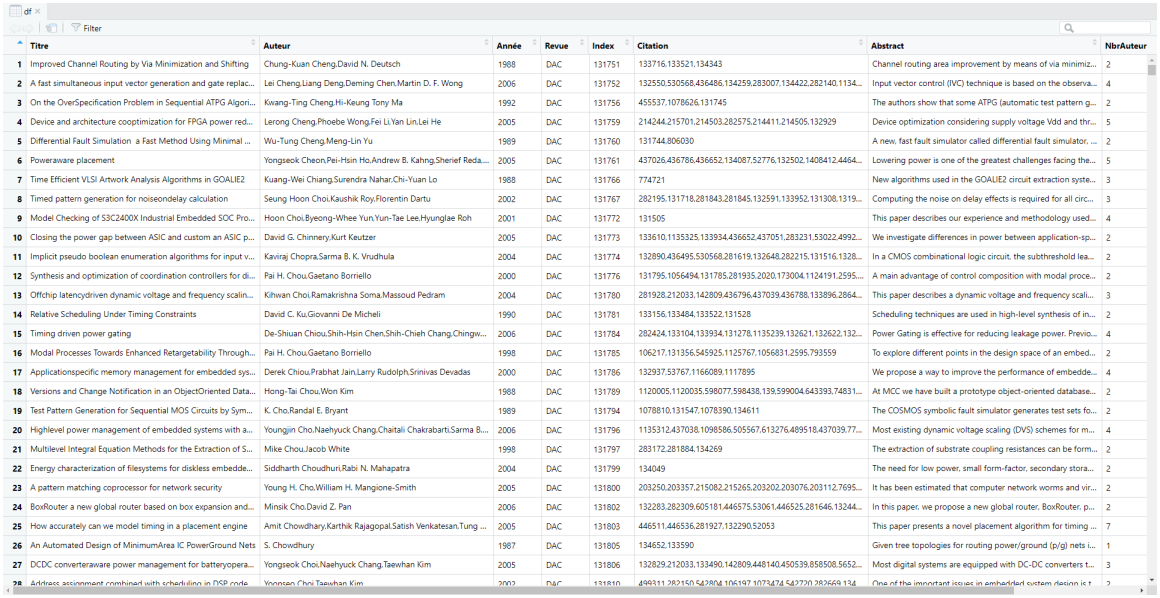
Nous disposons d'un fichier texte contenant des informations sur des articles scientifiques parus dans des revues et conférences. Ces informations comprennent :

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

Nous souhaitons faire une étude sur ce document, tout en passant par une série d'opérations.

2 Nettoyage

Grace à la fonction remplissage, qui traite ligne par ligne notre document on réussit à remplir notre data frame "df".
Ce qu'il nous reste à compléter maintenant, c'est l'exportation dans un fichier csv, juste avant de faire cela, on fait un petit nettoyage, qui se présente sur la suppression de toutes les lignes avec des abstract et/ou auteur qui ont une valeur "NA".



df

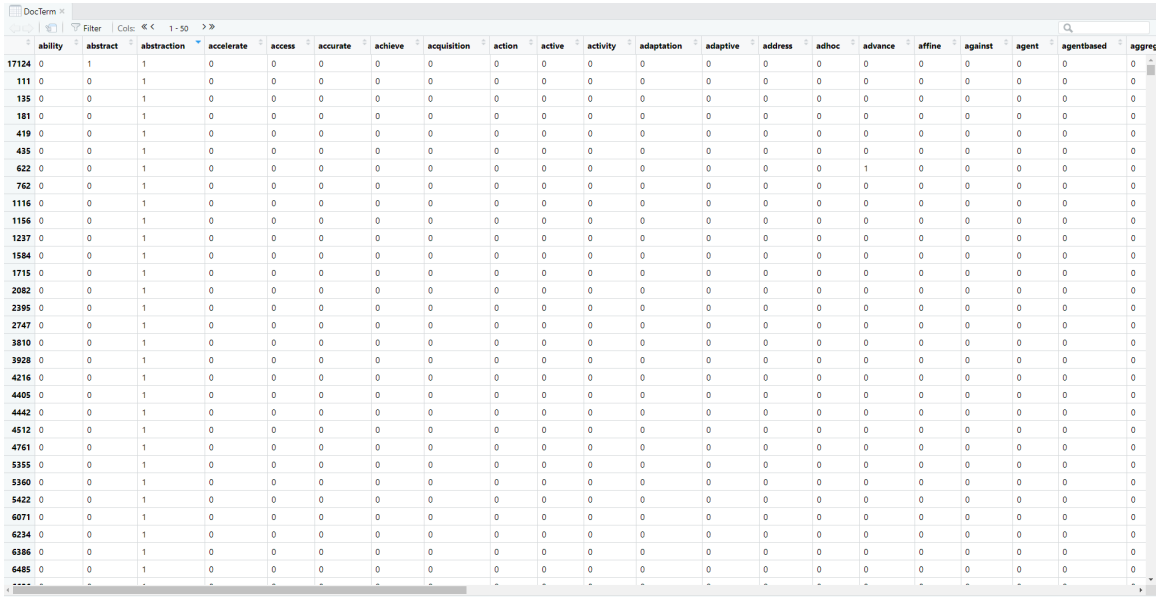
Filter

	Titre	Auteur	Année	Revue	Index	Citation	Abstract	NbrAuteur
1	Improved Channel Routing by Via Minimization and Shifting	Chung-Kuan Cheng,David N. Deutsch	1988	DAC	131751	133716,133521,134343	Channel routing area improvement by means of via minimiz...	2
2	A fast simultaneous input vector generation and gate replac...	Lei Cheng,Liang Deng,Deming Chen,Martin D. F. Wong	2006	DAC	131752	132550,530568,436486,134259,283007,134422,282140,1134...	Input vector control (IVC) technique is based on the observa...	4
3	On the OverSpecification Problem in Sequential ATPG Algori...	Kwang-Ting Cheng,Hi-Keung Tony Ma	1992	DAC	131756	455537,1078626,131745	The authors show that some ATPG (automatic test pattern g...	2
4	Device and architecture cooptimization for FPGA power red...	Lerong Cheng,Phoebe Wong,Fei Li,Yan Lin,Lei He	2005	DAC	131759	214244,215701,214503,282575,214411,214505,132929	Device optimization considering supply voltage Vdd and thr...	5
5	Differential Fault Simulation a Fast Method Using Minimal ...	Wu-Tung Cheng,Meng-Lin Yu	1989	DAC	131760	131744,806030	A new, fast fault simulator called differential fault simulator. ...	2
6	Poweraware placement	Yongseok Cheon,Pei-Hsin Ho,Andrew B. Kahng,Sherief Reda,...	2005	DAC	131761	437026,436786,436652,134087,52776,132502,1408412,4464...	Lowering power is one of the greatest challenges facing the...	5
7	Time Efficient VLSI Artwork Analysis Algorithms in GOALIE2	Kuang-Wei Chiang,Surendra Nahar,Chi-Yuan Lo	1988	DAC	131766	774721	New algorithms used in the GOALIE2 circuit extraction syste...	3
8	Timed pattern generation for noiseondelay calculation	Seung Hoon Choi,Kaushik Roy,Florentin Dartu	2002	DAC	131767	282195,131718,281843,281845,132591,133952,131308,1319...	Computing the noise on delay effects is required for all circ...	3
9	Model Checking of S3C2400X Industrial Embedded SOC Pro...	Hoon Choi,Byeong-Whee Yun,Yun-Tae Lee,Hyunglae Roh	2001	DAC	131772	131505	This paper describes our experience and methodology used...	4
10	Closing the power gap between ASIC and custom an ASIC p...	David G. Chinnery,Kurt Keutzer	2005	DAC	131773	133610,1135325,133934,436652,437051,283231,53022,4992...	We investigate differences in power between application-sp...	2
11	Implicit pseudo boolean enumeration algorithms for input v...	Kaviraj Chopra,Sarma B. K. Vrudhula	2004	DAC	131774	132890,436495,530568,281619,132648,282215,131516,1328...	In a CMOS combinational logic circuit, the subthreshold lea...	2
12	Synthesis and optimization of coordination controllers for di...	Pai H. Chou,Gaetano Borriello	2000	DAC	131776	131795,1056494,131785,281935,2020,173004,1124191,2595...	A main advantage of control composition with modal proce...	2
13	Offchip latencydriven dynamic voltage and frequency scalin...	Kihwan Choi,Ramakrishna Soma,Massoud Pedram	2004	DAC	131780	281928,212033,142809,436796,437039,436788,133896,2864...	This paper describes a dynamic voltage and frequency scali...	3
14	Relative Scheduling Under Timing Constraints	David C. Ku,Giovanni De Micheli	1990	DAC	131781	133156,133484,133522,131528	Scheduling techniques are used in high-level synthesis of in...	2
15	Timing driven power gating	De-Shiuan Chiou,Shih-Hsin Chen,Shih-Chieh Chang,Chingw...	2006	DAC	131784	282424,133104,133934,131278,1135239,132621,132622,132...	Power Gating is effective for reducing leakage power. Previo...	4
16	Modal Processes Towards Enhanced Retargetability Through...	Pai H. Chou,Gaetano Borriello	1998	DAC	131785	106217,131356,545925,1125767,1056831,2595,793559	To explore different points in the design space of an embed...	2
17	Applicationspecific memory management for embedded sys...	Derek Chiou,Prabhat Jain,Larry Rudolph,Srinivas Devadas	2000	DAC	131786	132937,53767,1166089,1117895	We propose a way to improve the performance of embedde...	4
18	Versions and Change Notification in an ObjectOriented Data...	Hong-Tai Chou,Won Kim	1988	DAC	131789	1120005,1120035,598077,598438,139,599004,643393,74831...	At MCC we have built a prototype object-oriented database...	2
19	Test Pattern Generation for Sequential MOS Circuits by Sym...	K. Cho,Randal E. Bryant	1989	DAC	131794	1078810,131547,1078390,134611	The COSMOS symbolic fault simulator generates test sets fo...	2
20	Highlevel power management of embedded systems with a...	Youngjin Cho,Naehyuck Chang,Chaitali Chakrabarti,Sarma B....	2006	DAC	131796	1135312,437038,1098586,505567,613276,489518,437039,77...	Most existing dynamic voltage scaling (DVS) schemes for m...	4
21	Multilevel Integral Equation Methods for the Extraction of S...	Mike Chou,Jacob White	1998	DAC	131797	283172,281884,134269	The extraction of substrate coupling resistances can be form...	2
22	Energy characterization of filesystems for diskless embedde...	Siddharth Choudhuri,Rabi N. Mahapatra	2004	DAC	131799	134049	The need for low power, small form-factor, secondary stora...	2
23	A pattern matching coprocessor for network security	Young H. Cho,William H. Mangione-Smith	2005	DAC	131800	203250,203357,215082,215265,203202,203076,203112,7695...	It has been estimated that computer network worms and vir...	2
24	BoxRouter a new global router based on box expansion and...	Minsik Cho,David Z. Pan	2006	DAC	131802	132283,282309,605181,446575,53061,446525,281646,13244...	In this paper, we propose a new global router, BoxRouter, p...	2
25	How accurately can we model timing in a placement engine	Amit Chowdhary,Karthik Rajagopal,Satish Venkatesan,Tung ...	2005	DAC	131803	446511,446536,281927,132290,52053	This paper presents a novel placement algorithm for timing ...	7
26	An Automated Design of MinimumArea IC PowerGround Nets	S. Chowdhury	1987	DAC	131805	134652,133590	Given tree topologies for routing power/ground (p/g) nets i...	1
27	DCDC converteraware power management for batteryopera...	Yongseok Choi,Naehyuck Chang,Taewhan Kim	2005	DAC	131806	132829,212033,133490,142809,448140,450539,858508,5652...	Most digital systems are equipped with DC-DC converters t...	3
28	Address assignment combined with scheduling in DSP code ...	Youngsoo Choi,Taewhan Kim	2002	DAC	131810	499311,282150,542804,106197,1073474,542720,282669,134...	One of the important issues in embedded system design is t...	2

FIGURE 1 – Data Frame

3 DocTerm

Grâce à la fonction "DocumentTermMatrix" du package tm sur R, on réalise la matrice document termes, on obtient un fichier de taille énorme c'est à dire + de 2.6 GO, ce qui nous oblige de retirer les termes les moins fréquents grâce à la fonction "removeSparseTerms", on se retrouve avec 1061 termes au lieu de 9000.



DocTerm

Filter | Cols: 1 - 50

	ability	abstract	abstraction	accelerate	access	accurate	achieve	acquisition	action	active	activity	adaptation	adaptive	address	adhoc	advance	affine	against	agent	agentbased	aggreg
17124	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
111	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
135	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
181	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
419	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
435	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
622	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
762	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1116	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1156	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1237	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1584	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1715	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2082	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2395	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2747	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3810	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3928	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4216	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4405	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4442	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4512	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4761	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5355	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5360	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5422	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6071	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6234	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6386	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6485	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FIGURE 2 – Doc Termes

4 Qlik Sense Interprétation

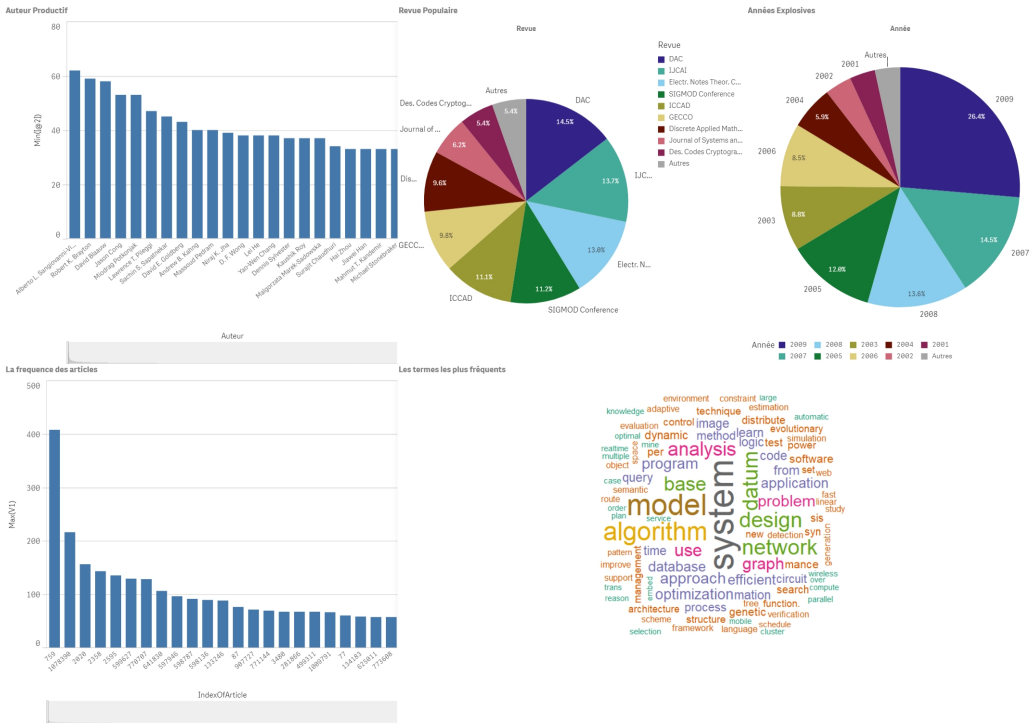


FIGURE 3 – Doc Termes

Tout d'abord, on s'est intéressé aux auteurs les plus productifs, ensuite on a fait le lien avec l'année la plus explosive qui est 2009 avec 26.4% d'articles, on voit aussi que les revues les plus populaires sont respectivement "DAC" , "IJCAI", "Electr Notes". L'article le plus cité c'était celui qui a comme index "759". On a renforcé notre interpretation avec wordCloud du R, et on a trouvé que les termes "System","Model","Algorithm" sont les plus fréquents.

5 Application de k-means

Sur cette section on s'est intéressé sur une petite partie de notre dataframe, et voila ce qu'on obtient :

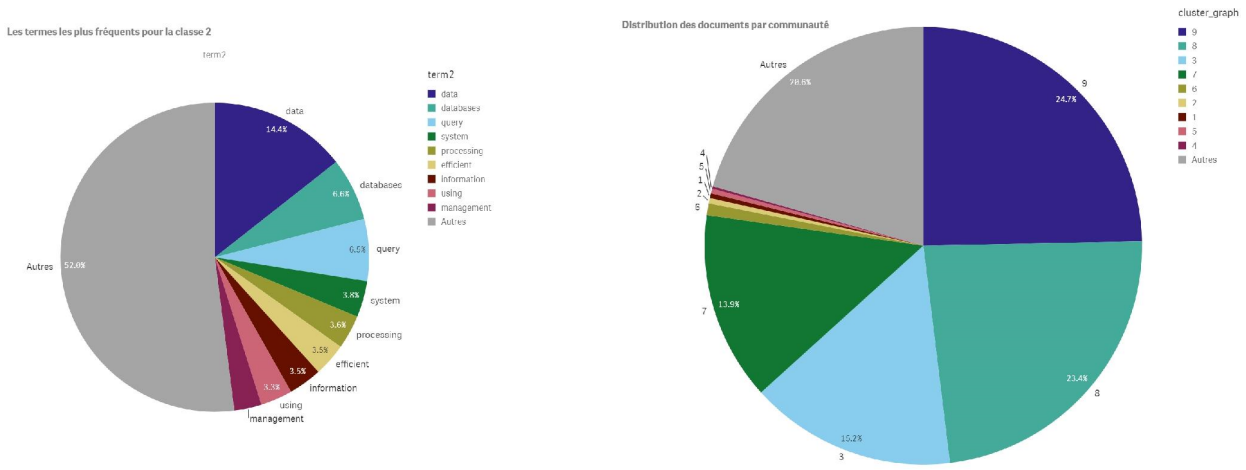


Figure 4 – Doc Termes

6 Conclusion

Enfin il faut savoir que plusieurs calculs ont été évité, à cause du volume très grand des formats CSV obtenu, comme par exemple la matrice co-Auteur qui pese 67GB, et citations.

On a essaye d'exploiter le maximum notre dataframe pour avoir des informations pertinentes .

Lien GitHub