Projet Network Analysis for Information Retrieval

Hakim Ziani, Nassim Ettorche

Universite de Lyon 2

Abstract. Dans ce projet d'analyse de réseaux pour la récupération d'information, nous avons abordé de manière systématique le traitement et l'exploitation de données textuelles et structurelles issues de diverses sources. Initialement, nous avons concentré nos efforts sur l'acquisition et le nettoyage des données, assurant leur sauvegarde dans un format optimisé pour une réutilisation aisée, tout en réalisant une classification automatique supervisée et en évaluant des statistiques de base sur les documents et auteurs. Ensuite, nous avons structuré notre corpus en construisant un graphe basé sur des informations telles que les citations, permettant des analyses poussées via des mesures de centralité et de visualisation de la topologie des données. La mise en œuvre d'un moteur de recherche a permis la récupération d'articles via des mots-clés, exploitant des similarités sémantiques et structurales. L'ajout de clustering, à travers des méthodes comme le spectral, a enrichi notre approche en intégrant des classifications automatiques. Enfin, nous avons adressé le défi de la classification supervisée des nœuds du graphe, utilisant des caractéristiques textuelles et structurelles, et exploré les avantages comparatifs de ces représentations pour une récupération d'information précise et nuancée.

Keywords: Ai, Graph, Machine Learning

1 Exploration de la donnees

Nous effectuons un prétraitement sur l'ensemble des articles scientifiques présents dans notre base de données, qui comprend une variété d'informations liées à ces articles. Cette opération débute par le nettoyage du texte, consistant à éliminer les caractères non souhaités, les mots vides et les termes peu fréquents. Par la suite, nous procédons à la représentation des documents en quantifiant la fréquence d'apparition de chaque terme au sein de chaque article. Ainsi, chaque terme se transforme en une dimension au sein d'un espace vectoriel. Nous obtenons la répartition suivante :

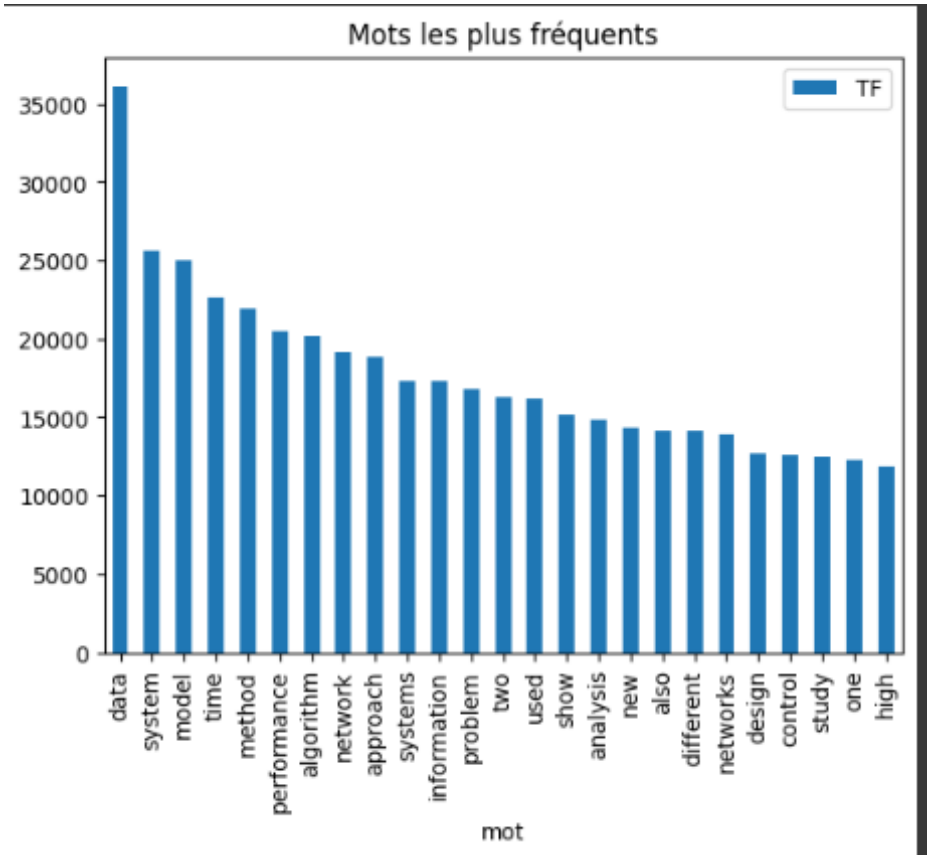

Fig. 1. Liste des 25 mots les plus fréquents du corpus

2 LDA

La LDA, ou Allocation Latente de Dirichlet, est une technique de modélisation de thèmes qui explore les sujets cachés présents dans un ensemble de documents textuels. Elle repose sur l'idée que chaque document est un mélange de plusieurs sujets, et chaque sujet est caractérisé par un ensemble spécifique de mots qui ont tendance à co-occurrence dans le corpus. Après avoir appliqué le modèle LDA à notre corpus,il semble visuellement que le modèle a identifié des regroupements thématiques bien distincts :

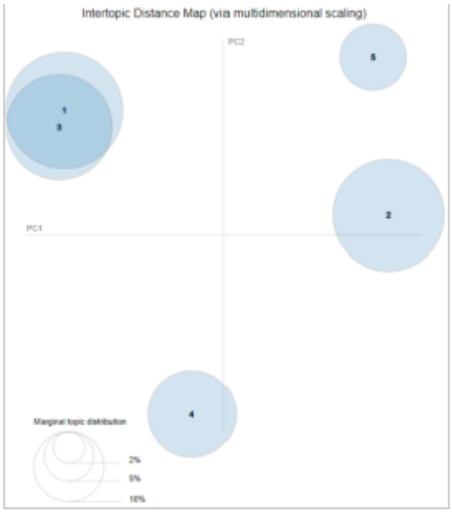

Fig. 2. Représentation des clusters avec LDA

Nous avons examiné les diverses répartitions thématiques et les mots clés liés à chaque thème. Pour une meilleure visualisation de ces thèmes, nous avons utilisé la méthode de réduction dimensionnelle t-SNE.

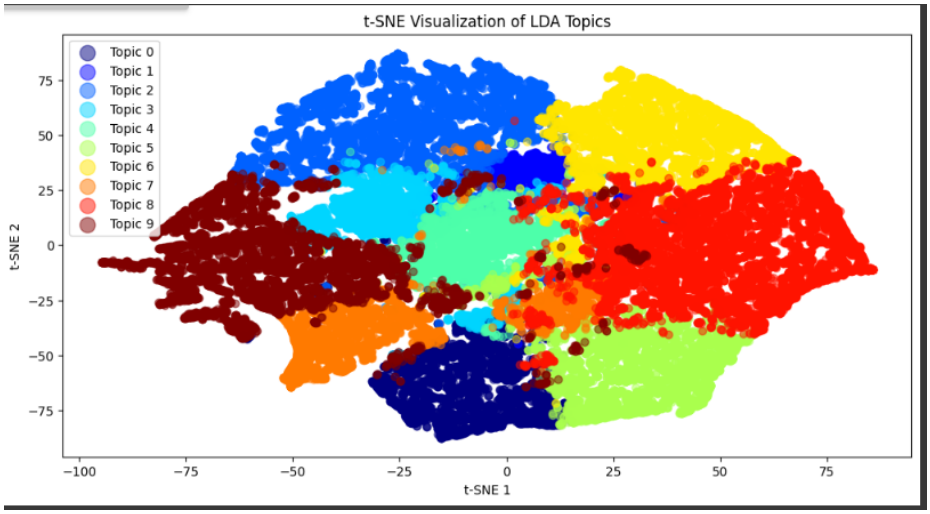

Fig. 3. Représentation avec t-SNE des thématiques extraites via une LDA

3 Moteur de Recherche

Dans cette partie, nous utilisons les donnees disponible pour developper un moteur de recherche permettant aux utilisateurs d'inserer une query en forme de texte en language naturel. Dans ce processus nous utilisons une methode de recherche d'information basee sur la similarite entre la query et les documents, pouvant ainsi proposer au user les TOP K documents plus semblables a sa query.

Deux types de representations de documents ont ete implemente, TF-IDF et l'autre basee sur l'embedding BERT.
Le challenge etant d'executer sur tout le corpus contant plus de 188k documents.

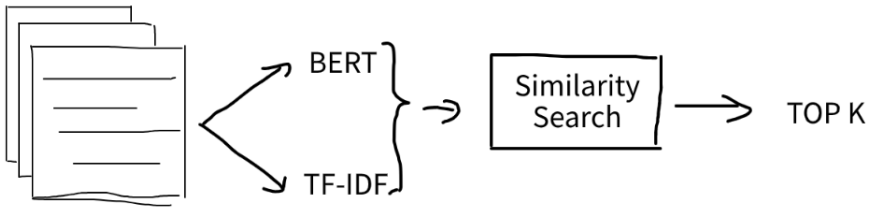

Fig. 4. Conception d'un moteur de recherche

3.1 Exemple d'execution

Query: Artificial intelligence and deep learning

Table 1. Résultats de la recherche TF-IDF; TOP 10

Index du document	ID du document
122269	77c575cd-094b-49d9-8e71-d915b66d2f13
97762	b4e86470-1668-425a-a017-97388d070f10
32256	dcbdee8a-7218-40ff-8f08-ec1fa98f7654
67501	280a717e-518b-4d75-996e-456994c1d449
13324	0c611553-8e1e-491a-961a-c4fedb264081
49998	2054d65d-016d-434d-b500-a9b66b6c3fe0
184773	5f3e7767-5271-47e9-9107-4147e1d5cf3f
164058	a1cc4e80-1bbb-4bfa-8791-68fd4c98aee7
179781	32e4132b-2ae7-49ca-b883-56eeb435789a
4415	11a2c8dd-ce69-450d-86e1-ec4c23950952

Result-BERT:

Table 2. Résultats de la recherche BERT: TOP 10

Index du document	ID du document
99574	bd234a55-acd9-49ad-9e1f-83df54a9169b
14790	1cc573b3-84fb-4e92-b23e-1ef73d3331a7
115693	180412ba-65ab-45fb-9502-10d9bd2a50ea
46081	3a7e6114-0290-4b1b-9466-6c8a48c0e6ad
141745	27018ae4-f29b-408d-a971-c42e9660f838
146393	ad27dace-d76d-401a-b91a-933e2d274664
57442	dc646350-2c4c-4dde-a44f-e217d961abf4
168129	820caaa9-01e2-4965-a34c-0012c6cc2c35
58581	f80c5ded-0146-45ea-b768-2058d942f06b
136752	740daeaf-8165-4e42-b07a-7817ed2359bf

4 Structure de Graphe

Nous avons opté pour une modélisation de notre corpus sous forme de réseau, en considérant chaque document comme un nœud du graphe. Les liens entre ces nœuds sont établis sur la base des références bibliographiques : si un document cite un autre, un lien est créé entre eux. Cette approche repose sur l'hypothèse selon laquelle deux documents reliés par une citation sont susceptibles d'aborder des sujets similaires.

Nous avons utilisé le nom de la conférence, tel qu'indiqué dans la colonne "venue" de nos données, comme étiquettes pour classifier les documents. Nous avons postulé que les documents présentés lors d'une même conférence couvrent généralement des sujets semblables, et cette hypothèse a été confirmée par la visualisation des clusters au sein du graphe.

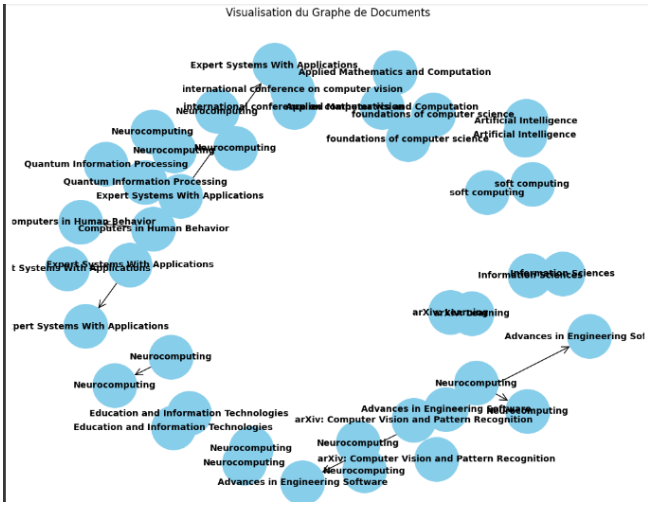

Fig. 5. Visualisation du graphe

Il apparaît que les nœuds portant le même nom de conférence sont fréquemment connectés par des liens de citation.

Nous avons envisagé l'analyse des plus grandes composantes connexes du graphe. En pratique, cela peut révéler des regroupements de documents qui se citent mutuellement, indiquant des domaines de recherche étroitement liés ou des sujets spécifiques qui captivent l'attention de communautés scientifiques particulières.

Nous avons calculé et analysé d'autres statistiques qui se basent sur la structure du réseau, telles que la distribution des degrés (qui indique combien de connexions chaque article a avec d'autres), la largeur du réseau (qui peut donner une idée de l'étendue du réseau à travers différents niveaux de connexions),

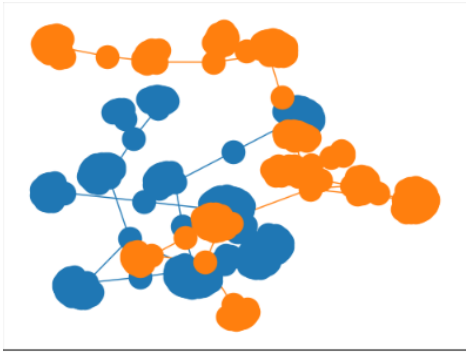

Fig. 6. Représentation des plus grande composantes connexes

et la densité (qui mesure à quel point les articles sont interconnectés). Ces données ont été soigneusement traitées et sont clairement exposées dans notre carnet Jupyter.

5 Clustering de noeuds

5.1 Kmeans

En raison de limitations liées à la mémoire RAM, nous avons opté pour l'application d'algorithmes de clustering sur une sélection restreinte de données. Cette sélection a été faite en choisissant les documents liés aux cinq noms de conférences les plus récurrents. Dans un premier temps, nous avons employé l'algorithme K-means pour créer cinq clusters, et c'est ainsi que nous avons obtenu les résultats suivant:

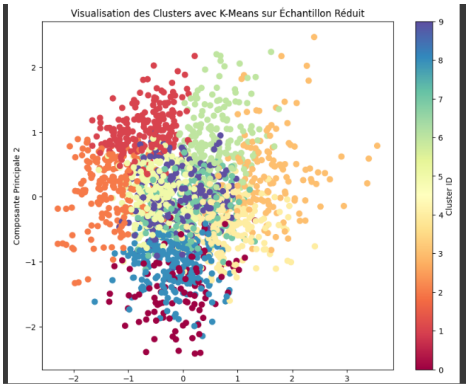

Fig. 7. Clusters (Kmeans)

Nous avons évalué les différences entre les classifications pré-définies des documents et les groupes formés par notre processus de clustering. Pour cela, nous avons mené une analyse à la fois de manière qualitative, en observant directement les résultats, et de manière quantitative, en faisant appel à des métriques telles que l'Indice de Rand Ajusté (ARI) et l'Information Mutuelle Ajustée (AMI). Les scores obtenus, avoisinant 0,5 pour les deux indicateurs, sont encourageants. Ils montrent une correspondance solide entre les regroupements obtenus par clustering et les vraies catégories identifiées par les noms des conférences.

5.2 Hdbscan

cette méthode a renvoyé des résultats moins satisfaisants que K-Means pour deux principales raisons : d'abord, sa sensibilité aux paramètres, tels que la taille minimale des clusters et la distance minimale entre eux, qui peuvent être difficiles à ajuster de manière optimale ; ensuite, son attribution de certains points à du bruit ou à des outliers, ce qui peut sous-représenter ou ignorer des groupes réels dans les données, contrairement à K-Means qui assigne chaque point à un cluster, même s'il est éloigné des autres, ce qui peut entraîner une plus grande variabilité dans la densité et la taille des clusters.

5.3 Clustring Spectral

Le clustering spectral nous fait ressortir les clusters suivants :

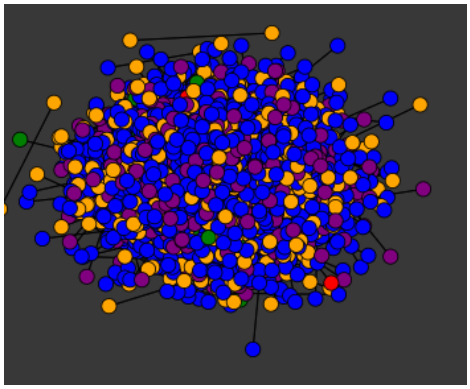

Fig. 8. Clustring spectral

Enfin, nous avons projeté nos nœuds dans un nouvel espace en utilisant la technique de réduction de dimension t-SNE. Cependant, la représentation en deux dimensions n'apparaît pas comme étant significative. Lorsque nous avons coloré les points en fonction des clusters identifiés auparavant, les couleurs ne semblent pas correspondre aux différents clusters attendus.

6 Classification de noeuds

Dans cette partie, l'etude de comment peut-on classfier les documents de base de donnee, en plusieurs classes. Ceci aidera dans la comprehension du corpus, mais aussi pour developper des futures systemes de recommandations.
Deux structure de documents a ete etudie, une representation semantiques des documents, grace au vecteurs BERT deja en calcule au paravant dans l'etape de creation du moteur de recherche. Un autre type de representation de documents est utilise permettant d'exploiter les liens de citations entre les papiers de recherches en utilisant le matrice d'adjacence.

Le challenge avec la classification supervise est l'obtention des lablels de sortie pour chaque document du corpus.

Une methode basee sur l'exploitation des LLMs afin de generer les labels automatiquement a ete exploite

PROMPT-GPT: i have a list of article titles in a csv, i want to get generate a class label for each title, first give me a list of possible lables, given that the the papers are in computer science, give me a list of categories a can use as paper labels like artificial intelligence, cloud computing...
PROMPT-GPT: you will assign each title the corresepodant category for example a title in which talk about an ai technique should be assigned to artificial intelligence and so on. the output should be in a csv i can download,

Table 3. Resultat du Labelling

ID	Titre	Abstract	Venue	Texte	Category
04b38d22	Monetization as a Motivator for the Freemium E...	The paper describes user behavior as a result ...	arXiv: Computers and Society	the paper describes user behavior result intro...	E-Commerce and E-Business
086c091e	Towards intelligent distributed computing : ce...	Distributed computing systems are of huge impo...	NaN	distributed computing systems huge importance ...	Other
3089c330	Changes in Urban Area Discovered by Analysis o...	A period of almost 150 years (since the middle...	NaN	a period almost years since middle th century ...	Urban Computing and Smart Cities
328ff460	A Survey of Social Web Mining Applications for...	The final publication is available at Springer...	NaN	the final publication available springer via h...	Social Web and Web Technologies
38f94267	Quantized Control and Data-Rate Constraints	This article briefly describes the topic of qu...	NaN	this article briefly describes topic quan tize...	Control Systems and Robotics

Le tableau ci dessous represente le nombre d'instance pour chaque classe generee

Table 4: Distribution Des Categories

Category	Count
Other	174752
Algorithm Design and Analysis	8409
Database Systems	965
Cloud Computing	604
Machine Learning (ML)	576
Networking and Communications	533
Internet of Things (IoT)	466
Deep Learning	284
Software Engineering	235
Robotics	215
Virtual Reality (VR) and Augmented Reality (AR)	177
Graphics and Visualization	125

Continued on next page

Table 4 – *Continued from previous page*

Category	**Count**
E-Commerce and E-Business	109
Distributed Systems	89
Computer Vision	87
Artificial Intelligence (AI)	78
Ethical and Social Aspects of Computing	73
Cybersecurity	61
Natural Language Processing (NLP)	55
Urban Computing and Smart Cities	45
Data Science and Analytics	43
Programming Languages and Compilers	29
Human-Computer Interaction (HCI)	26
Game Development and Design	17
Educational Technology	16
Social Web and Web Technologies	15
Quantum Computing	14
Computational Biology	14
Blockchain and Cryptocurrencies	10
Energy-Efficient Computing	8
Edge Computing	6
Control Systems and Robotics	3
Systems and Architecture	3
Theoretical Computer Science	1

Nous avons pris la decision d'eliminer les categories comportant peu ou trops d'instances, pour eliminer le bias et avoir des donnee avec une balance. Les categories suivantes on ete choisis

(['Cloud Computing',
'Machine Learning (ML)',
'Networking and Communications',
'Internet of Things (IoT)',
'Deep Learning ',
'Software Engineering',
'Robotics'])

6.1 En utilisant l'Embeddings des documents

Table 5: Model Evaluation Metrics

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
RidgeClassifierCV	0.91	0.91	None	0.91	0.87
LinearDiscriminantAnalysis	0.89	0.89	None	0.89	1.32
CalibratedClassifierCV	0.89	0.88	None	0.89	28.28
RidgeClassifier	0.89	0.88	None	0.89	0.21
LogisticRegression	0.87	0.88	None	0.87	0.85
PassiveAggressiveClassifier	0.88	0.88	None	0.88	1.51
LinearSVC	0.86	0.86	None	0.86	5.76
SGDClassifier	0.86	0.85	None	0.86	1.88
Perceptron	0.86	0.85	None	0.86	0.62
SVC	0.86	0.84	None	0.86	1.82
LGBMClassifier	0.82	0.80	None	0.82	65.70
NuSVC	0.82	0.78	None	0.81	2.75
ExtraTreesClassifier	0.77	0.74	None	0.77	1.08
RandomForestClassifier	0.77	0.74	None	0.76	6.83
GaussianNB	0.71	0.74	None	0.71	0.11
NearestCentroid	0.70	0.73	None	0.70	0.12
BernoulliNB	0.68	0.71	None	0.68	0.27
KNeighborsClassifier	0.72	0.71	None	0.72	0.19
BaggingClassifier	0.62	0.59	None	0.61	27.30
AdaBoostClassifier	0.54	0.56	None	0.53	22.67
DecisionTreeClassifier	0.48	0.46	None	0.48	1.95
ExtraTreeClassifier	0.40	0.38	None	0.40	0.10
QuadraticDiscriminantAnalysis	0.20	0.19	None	0.20	0.76
DummyClassifier	0.25	0.17	None	0.10	0.07
LabelSpreading	0.25	0.17	None	0.10	0.45
LabelPropagation	0.25	0.17	None	0.10	0.38

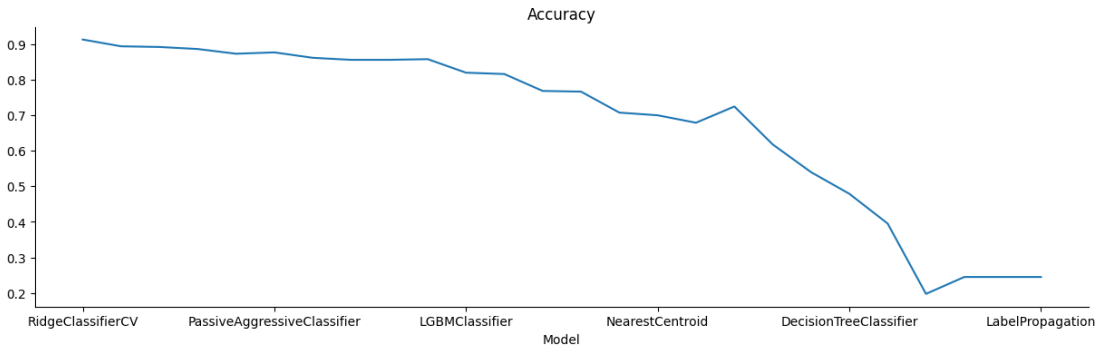

Fig. 9. Valeur d'Accuracy pour differents modeles

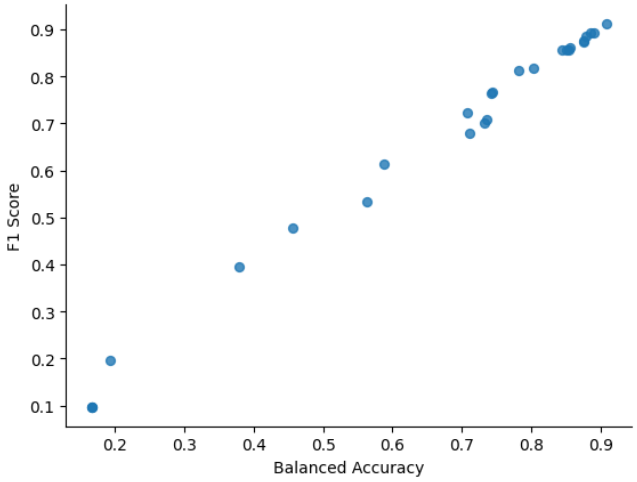

Fig. 10. Relation Accuracy, F1-Score

La figure ci-dessus represente la correlation entre l'accuracy et la mesure de F1-score, une relation importante afin de juger la qualite d'une classfication

ci dessous la matrice de confusion du modele le plus performant: RidgeClassifier

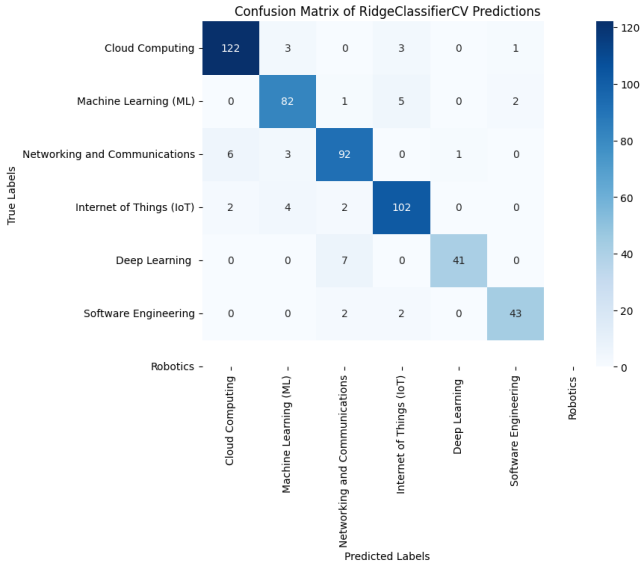

Fig. 11. Matrice de confusion: RidgeClassifier

6.2 En utilisant la structure de graphe

Pour effectuer la classification, la matrice d'adjacence du graphe a ete utilise en tant que feature map pour le modele. Le resultat de la classfication par RidgeClassifier donne une Accuracy de **0.90**. En gardant un oeuil critique sur cette approche, puisque ce resultat est donnee par un subset de la donnee vue la difficulte combinatoire de travailler avec une matrice d'adjacence pour tout le corpus.

7 Conclusion

Dans ce projet, une etude complete a ete effectue sur un corpus de donnee de papiers de recherche. nous avons créé un système pour nettoyer et organiser les données, puis utilisé des graphes pour voir comment les documents se rapportent entre eux. Grâce à un moteur de recherche spécial, il est devenu facile de retrouver des articles avec des requetes en langage naturel. Ainsi de bon resultat de classfication et clustering ont ete presente, ouvrant d'autres opportunitees d'outils tel qu'un systeme de recommandation (personalise our par groupe).