

M1 Informatique – UE Projet

Carnet de bord : les coulisses de la recherche documentaire

Les éléments que vous indiquez dans ce carnet donneront lieu à une notation

NOTE:7,75 /10

Noms, prénoms et spécialité :

Master 1 DAC

Sofia Borchani

Souleymane Mbaye

Nolwenn Pigeon

Sujet :

Analyse de reviews et recommandation

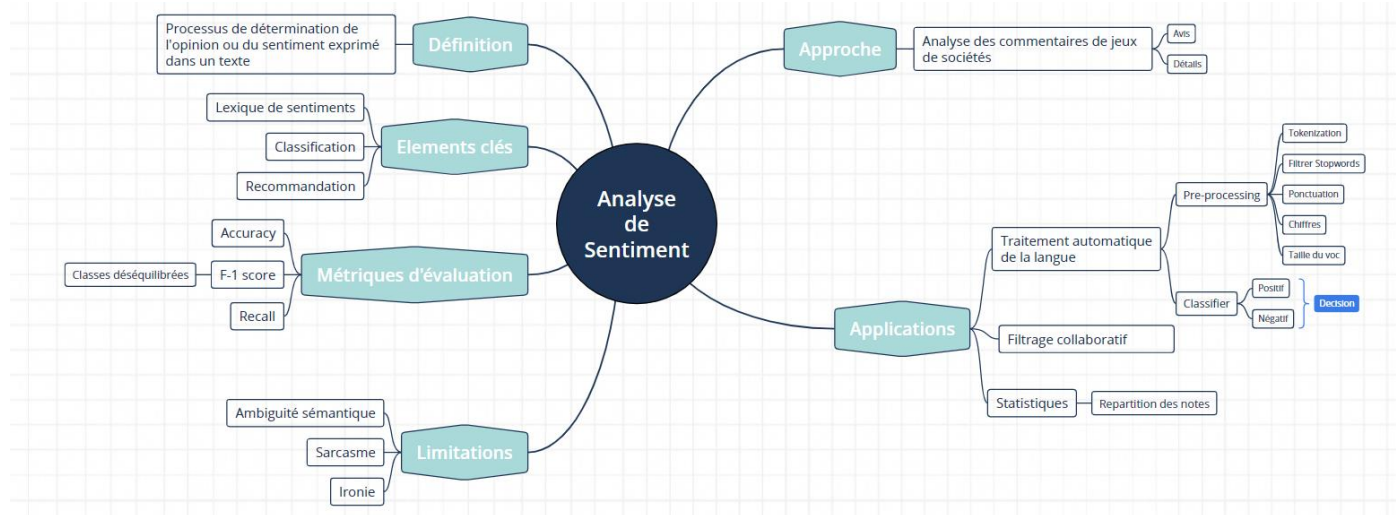
Consigne :

1. **Introduction (5-10 lignes max) :** Décrivez rapidement votre sujet de recherche, ses différents aspects et enjeux, ainsi que l'angle sous lequel vous avez décidé de le traiter.

Les consommateurs[5] partagent de plus en plus leurs avis sur les produits et services, ce qui incite les entreprises à accorder plus d'attention aux commentaires des clients pour améliorer leur expérience. Dans ce projet de développement, l'objectif est d'analyser les sentiments[4,6] des joueurs de jeux de société, exprimés sous forme d'avis et de critiques sur le site TricTrac. La base de données est composée de 150 000 avis et 200 000 notes sur plus de 15 000 jeux. L'enjeu principal est de différencier les critiques positives des critiques négatives et d'évaluer l'impact de mots spécifiques pour comprendre comment l'analyse d'avis peut permettre aux machines de recommander des jeux similaires. Après la mise en forme et le nettoyage des données, nous allons utiliser des méthodes de traitement automatique du langage pour classer les sentiments. Nous traiterons aussi de la problématique de transfert du classifieur vers des données twitter[9]. Enfin, en utilisant le filtrage collaboratif[3,7], nous allons étudier un système de recommandation[8] en fonction des notes et avis donnés par les utilisateurs. Le projet s'ouvrira éventuellement sur l'utilisation de Graph Neural Networks pour améliorer la recommandation.

Intro : 2/2 Très bonne introduction. Le contexte et les enjeux sont bien définis.

2. **Les mots clés retenus :** Listez les mots-clés que vous avez utilisés pour votre recherche bibliographique. Organisez-les sous forme de carte heuristique.



Mots -clés : 1/1 Carte heuristique claire, mots-clés organisés et hiérarchisés.

3. **Descriptif de la recherche documentaire (10-15 lignes) :** Décrivez votre utilisation des différents outils de recherche (moteurs de recherche, base de données, catalogues, recherche par rebond etc.). Comparez ensuite les outils entre eux. A quelles sources vous ont-ils permis d'accéder ? Quelles sont leurs spécificités ? Quel est leur niveau de spécialisation ?

Après une rapide recherche littéraire des mots clés de notre sujet sur Wikipédia, nous avons commencé la recherche documentaire en consultant le catalogue universitaire SUDOC. Ce catalogue recense toutes les collections des bibliothèques universitaires françaises. Il est possible d'effectuer une recherche simple ou avancée et de filtrer cette recherche. Parmi les ressources bibliographiques pluridisciplinaires, nous avons consulté **Jstor, une plateforme** qui offre une grande variété de contenu académique. Pour trouver des articles plus pertinents dans le domaine scientifique, nous avons principalement consulté la base de données bibliographiques, Web Of Science, introduite dans les cours de Recherche Documentaire. Cette plateforme est particulièrement intéressante car elle génère des indicateurs bibliométriques qui analysent les statistiques des publications. Notamment le filtre 'highly cited papers' nous garantit que ces sources sont pertinentes et **écrites** par un large panel d'auteurs scientifiques. Afin de mieux comprendre l'avancée des recherches sur notre domaine d'étude nous tenions à lire le travail de nos prédécesseurs sur le moteur de recherche theses.fr. Nous avons également consulté les travaux supervisés par nos responsables de projet.[1,2]

Nous allons à présent comparer ces outils. Wikipedia est un outil de recherche généraliste qui n'a pas de niveau de spécialisation spécifique. Il est accessible à tous et très populaire, si bien qu'il est parfois considéré comme peu fiable. Le catalogue SUDOC est spécialisé dans les ouvrages universitaires, il ne couvre donc que les collections des bibliothèques universitaires françaises. Jstor est un outil de recherche spécialisé pour des sources académiques et scientifiques généralement fiables et de haute qualité. Cependant n'ayant pas d'abonnement, il était parfois difficile d'accéder à l'entièreté de ces sources. Jstor et WOS sont utiles pour consulter différents types de documents scientifiques tandis que SUDOC et theses.fr sont plus adaptés pour trouver des documents de bibliothèques et des travaux de recherche.

Descriptif de la recherche : 1,5/2 La démarche est bien expliquée et claire. Attention vous devez faire votre description en 15 lignes maximum. Attention aux termes utilisés pour désigner les outils utilisés. Par ex : Jstor : bibliothèque numérique + système d'archivage en ligne de publications universitaires et scientifiques

4. **Bibliographie produite dans le cadre du projet :** Utilisez la norme ACM.

Références bibliographiques :

- [1] Clara Gainon de Forsan de Gabriac. 2021. Deep Natural Language Processing for User Representation. Données textuelles. Sorbonne Université Lip6, Paris.
- [2] François Buet. 2022. Modèles neuronaux pour la simplification de parole, application au sous-titrage. Données textuelles. Université Paris Saclay, Paris.
- [3] Khadija Naji and Abdelali Ibriz. 2022. Approach for Eliciting Learners' Preferences in Moocs Through Collaborative Filtering. *International Journal of Emerging Technologies in Learning (iJET)* 17, 235–245 (2022). DOI:https://doi.org/10.3991/ijet.v17i14.29887
- [4] Meng Fanqi, Zheng Yujie, Bao Songbin, Wang Jingdong, and Yang, Shuaisong. 2022. Formulaic language identification model based on GCN fusing associated information. *Peejr Computer Science* 8, (June 2022).
- [5] Praphula Kumar Jain, Rajendra Pamula, and Gautam Srivastava. 2021. A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. *Computer Science Review* 41, (Aout 2021).
- [6] Schouten, Kim and Frasincar, Flavius. 2016. Survey on Aspect-Level Sentiment Analysis. *IEEE Transactions on knowledge and data engineering* 28, 3 (March 2016), 813–830.
- [7] Xiaoxian Yang, Sijing Zhou, and Min Cao. 2020. An Approach to Alleviate the Sparsity Problem of Hybrid Collaborative Filtering Based Recommendations: The Product-Attribute Perspective from User Reviews. *Mobile Networks and Applications* 25, 2 (April 2020), 376–390. DOI:https://doi.org/10.1007/s11036-019-01246-2

[8] Zhihua Cui, Xianghua Xu, Fei XUE, Xingjuan Cai, Yang Cao, Wensheng Zhang, and Jinjun Chen. 2020. Personalized Recommendation System Based on Collaborative Filtering for IoT Scenarios. *IEEE Transactions on Services Computing* 13, 4 (August 2020), 685–695. DOI:<https://doi.org/10.1109/TSC.2020.2964552>

[9] Zulfadzli Drus and Haliyana Khalid. 2019. Sentiment Analysis in Social Media and Its Application: Systematic Literature Review. *Procedia Computer Science* 161, (2019), 707–714.

Bibliographie : 1,25/2 : Biblio cohérente mais il vous faut au moins 10 références. Attention, plusieurs références ne sont pas dans la bonne norme. Diversité des sources.

Evaluation des sources (5 lignes minimum par source) : Choisissez 3 sources parmi votre bibliographie, décrivez la manière dont vous les avez trouvées et faites-en une évaluation critique en utilisant les critères vus sur les supports de TDs.

Source [1]

Cette source est une thèse de doctorat présentée par Clara Gainon de Forsan de Gabriac pour obtenir le grade de Docteur de Sorbonne Université. Dirigée par Vincent Guigue et Patrick Gallinari, la thèse a été présentée et soutenue publiquement en décembre 2021. Composée de 137 pages, la thèse est structurée en six parties sans compter les premières pages contenant le résumé, les remerciements, le sommaire, la liste des figures, le glossaire, ni la bibliographie à la fin. L'introduction présente le contexte ainsi que les motivations de la thésarde et ses contributions dans le domaine du NLP (natural language processing) et de la recommandation. L'auteure y décrit également les précédents travaux avant d'exposer ses méthodes et modèles d'expérimentation. Diplômée de l'école d'ingénieur INSA Rouen en 2015, l'auteure a fait sa thèse deux ans plus tard au laboratoire d'informatique Lip6 de Paris 6, après quelques expériences professionnelles dans la science des données. Sa thèse est accessible sur l'archive ouverte pluridisciplinaire HAL theses. La bibliographie est longue de 10 pages ce qui correspond aux normes pour une thèse et regroupe aussi bien des sources françaises que étrangères pouvant dater jusqu'à 2006.

Source [4]

Cette source est un article de recherche scientifique qui suit une structure logique IMRAD (introduction, méthodologie, résultat et discussion). Les auteurs Fanqi Meng, Yujie Zheng, Songbin Bao, Jingdong Wang et Shuaisong Yang de nationalité chinoise sont pour la plupart des professeurs agrégés et des auteurs qualifiés. Leurs recherches portent sur divers sujets, notamment la vision par ordinateur, la conception d'architectures neuronales, le traitement du langage naturel et l'intelligence artificielle. Plus particulièrement, le chercheur Fanqi Meng a reçu plusieurs prix pour ses contributions à la recherche en informatique. Jingdong Wang est actuellement le scientifique en chef de la vision par ordinateur chez Baidu. Avant de rejoindre Baidu, il était chercheur principal senior chez Microsoft Research Asia. Wang a reçu de nombreuses distinctions pour ses contributions à la compréhension et à la récupération du contenu visuel. L'article a été récemment publié en juin 2022 dans la revue scientifique PeerJ computer science et cité 18 fois depuis, ce qui témoigne de son impact dans le domaine de la recherche en traitement automatique des langues (TAL). En ce qui concerne la bibliographie, on remarque qu'elle s'articule autour de thèmes majoritaires : méthodes d'identification et de détection d'expressions, modèles d'apprentissage automatiques, réseaux de neurones. De manière générale, la conclusion est en accord avec les données présentées. Elle résume également les limites de l'approche proposée et les pistes d'amélioration pour les recherches futures.

Source [5]

Cet article est une revue systématique de la littérature qui suit une méthodologie scientifique et présente une étude visant à déterminer comment les alliances de techniques de machine learning permettent d'analyser les sentiments des consommateurs. Cette étude a été faite au sein du département d'informatique et d'ingénierie à l'Institut indien de la technologie. Praphula Kumar Jain est l'auteur principal de cette source. Il a étudié les données et formulé le concept et le modèle. Rajendra Pamula, son directeur de thèse et Gautam Srivastava ont participé à la correction et la relecture de l'article. Publiée en 2021, dans le journal Computer Science Review par la maison d'édition scientifique renommée Elsevier, nous avons eu accès à cette source via la plateforme en accès libre ScienceDirect. La bibliographie est composée de 182 références à des articles et travaux de recherches scientifiques autour de l'analyse de sentiments, ses méthodes et applications. Elle est aussi longue car le but de cet article était d'inclure tous les articles de recherche sur l'analyse de sentiment et de les examiner avec précision tout en gardant une opinion impartiale.

Evaluation des sources : 2/3 : On attend de vous une évaluation de la pertinence des documents, leurs provenances, la rigueur du contenu, le niveau de l'information.

Vous avez très bien discuté rigueur des documents, leurs provenances et le niveau de l'information. Par contre, à aucun moment vous n'évoquez la pertinence de ces sources quant à votre projet.