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

Noms, prénoms et spécialité :

MASTER 1 Données Apprentissage Connaissances (DAC)
Sofia Borchani
Souleymane Mbaye
Nolwenn PIGEON

Sujet :

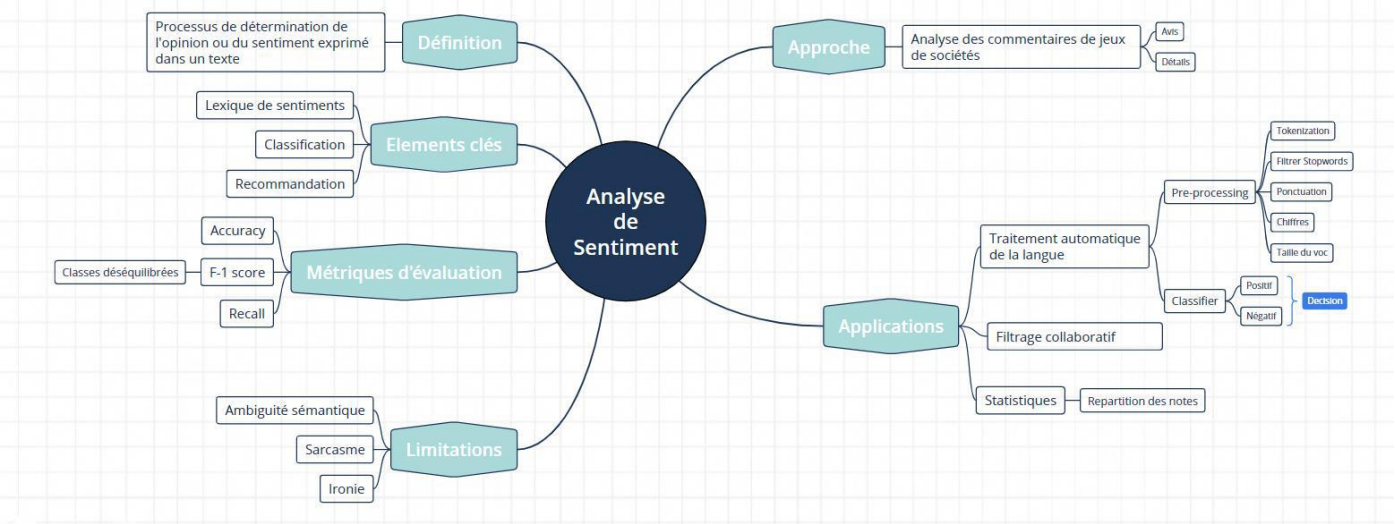
Analyse de reviews et recommandation

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 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 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 classifier les sentiments. Nous traiterons aussi de la problématique de transfert du classifieur vers des données twitter. Enfin, en utilisant le filtrage collaboratif, nous allons étudier un système de recommandation 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.

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



3. **Descriptif de la recherche documentaire (10-15 lignes) :** Décrivez votre utilisation des différents outils de recherche (moteurs de recherche, base de donnée, 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 ?

Nous avons commencé notre recherche littéraire sur Wikipedia, puis nous avons utilisé SUDOC, un catalogue qui référence les collections des bibliothèques universitaires françaises. Il est possible d'effectuer une recherche simple ou avancée et de filtrer cette recherche. La bibliothèque numérique Jstor offre une variété de contenu académique, tandis que Web Of Science est une base de données bibliographiques utile pour trouver des articles pertinents dans le domaine scientifique. 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. Theses.fr nous a permis de mieux comprendre l'avancée des recherches sur notre sujet d'étude en consultant le travail de nos prédécesseurs. Bien que Wikipedia soit généraliste et populaire, il est parfois considéré comme peu fiable. En comparaison, Jstor et WOS sont plus fiables et de haute qualité pour les documents scientifiques, mais l'accès à Jstor peut être limité sans abonnement. Le catalogue SUDOC est spécialisé dans les ouvrages universitaires, il ne couvre donc que les collections des bibliothèques universitaires françaises. 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.

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

[1] Antonio Feraco, Sivaji Bandyopadhyay, Dipankar Das, and Erik Cambria. 2016. Affective Computing and Sentiment Analysis. IEEE Intelligent System 31, 2 (2016), 102–107. DOI:https://doi.org/10.1109/MIS.2016.31

[2] Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. Training Classifiers with Natural Language Explanations. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 1884–1895. DOI:https://doi.org/10.18653/v1/P18-1175

[3] Clara Gainon de Forsan de Gabriac. 2021. Deep Natural Language Processing for User Representation. Données textuelles. Sorbonne Université Lip6, Paris.

[4] François Buet. 2022. Modèles neuronaux pour la simplification de parole, application au sous-titrage. Données textuelles. Université Paris Saclay, Paris.

[5] 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

[6] Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep Learning for Sentiment Analysis : A Survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, e1253.

[7] 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).

[8] 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).

[9] 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.

[10] William Yang Wang, Jiwei Li, Xiaodong He, and Jacob Eisenstein. 2018. Deep reinforcement learning for NLP. Association for Computational Linguistics, Melbourne Convention and Exhibition Centre, 19–21. DOI:https://doi.org/10.18653/v1/P18-5007

[11] 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

[12] 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. In IEEE Transactions on Services Computing. 685–695.

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

5. **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.

[3]
Clara Gainon de Forsan de Gabriac a soutenu en décembre 2021 une thèse de doctorat à Sorbonne Université, dirigée par Vincent Guigue et Patrick Gallinari. Sa thèse de 137 pages, structurée en six parties, porte sur le NLP et la recommandation. Après avoir présenté le contexte et les précédents travaux, elle expose ses méthodes et modèles d'expérimentation. Diplômée de l'école d'ingénieure INSA Rouen en 2015, l'auteure a réalisé sa thèse au laboratoire Lip6 de Paris 6. Sa thèse est disponible sur l'archive ouverte pluridisciplinaire HAL theses et sa bibliographie de 10 pages respecte les normes en vigueur pour une thèse, en incluant des sources françaises et étrangères remontant jusqu'à 2006. La thèse présentée par Clara Gainon de Forsan de Gabriac est pertinente pour notre projet, car elle aborde la problématique de la recommandation en utilisant des méthodes de traitement du langage naturel pour apprendre des représentations d'utilisateurs riches et versatiles, et en combinant une méthode de représentation par factorisation matricielle traditionnelle avec un modèle d'analyse de sentiments. De plus, la thèse traite également de la problématique de l'apprentissage de profils professionnels, qui est similaire à l'objectif de notre projet.

[7]
Cette source est un article scientifique publié en juin 2022 dans la revue PeerJ computer science et rédigé par les chercheurs chinois Fanqi Meng, Yujie Zheng, Songbin Bao, Jingdong Wang et Shuaisong Yang, dont certains sont des professeurs agrégés et des auteurs qualifiés dans les domaines de la vision par ordinateur, de la conception d'architectures neuronales, du traitement du langage naturel et de l'intelligence artificielle. L'article qui suit une structure logique IMRAD (introduction, méthodologie, résultat et discussion) a été cité 18 fois depuis sa publication, témoignant de son impact dans le domaine de la recherche en traitement automatique des langues. La bibliographie se concentre sur des thèmes tels que l'identification et la détection

d'expressions, les modèles d'apprentissage automatique et les réseaux de neurones. La conclusion est en accord avec les données présentées et propose des pistes d'amélioration pour les recherches futures. La source proposée est pertinente pour notre projet car elle traite de l'identification de la langue formulée, qui est une notion essentielle pour l'analyse de sentiments et la classification des commentaires clients. En utilisant des graphes de construction de phrases et des réseaux neuronaux convolutifs pour extraire des informations associées entre les mots, la méthode proposée peut améliorer l'exactitude de l'identification de la langue formulée. En outre, l'utilisation de Graph Neural Networks (GNN) pour améliorer la recommandation de jeux est également mentionnée comme ouverture pour notre projet, et la source suggère l'utilisation de GCN pour extraire des informations associées dans la construction de graphes de phrases. Par conséquent, cette source peut fournir des idées utiles pour la mise en œuvre de GNN pour améliorer la recommandation de jeux en fonction des avis et notes des utilisateurs.

[8]
Cet article est une revue systématique de la littérature qui suite une méthodologie scientifique portant sur l'analyse des sentiments des consommateurs à l'aide de techniques de machine learning. Écrit par Praphula Kumar Jain, de l'Institut indien de la technologie, et publié en 2021 dans le journal Computer Science Review d'Elsevier, il présente une méthodologie rigoureuse basée sur l'étude de 182 références scientifiques. Les co-auteurs Rajendra Pamula et Gautam Srivastava ont participé à la relecture et à la correction de l'article. L'objectif de cette revue était d'examiner toutes les recherches sur l'analyse des sentiments de manière impartiale. La source est pertinente pour notre projet car elle traite également de l'analyse des sentiments des consommateurs à partir de leurs avis en ligne, mais dans le domaine de l'hospitalité et du tourisme. Elle utilise également des techniques de machine learning pour traiter la grande quantité de données en ligne. Les résultats de cette étude pourraient être utiles pour différencier les critiques positives et négatives et à évaluer l'impact de mots spécifiques pour recommander des jeux similaires. Cependant, la source ne traite pas spécifiquement de l'analyse des avis de joueurs de jeux de société, donc des ajustements et des adaptations pourraient être nécessaires pour appliquer les résultats de l'étude à notre projet en particulier.