M1 Informatique –UE Projet
Carnet de bord

Noms, prénoms et spécialité :

HADDADI Hacene - DAC
AKNOUCHE Anis - DAC

Sujet :

Analyse d'un corpus du CV, apprentissage de représentation sur des documents structurés et démêlage des facteurs explicatifs

Table des matières

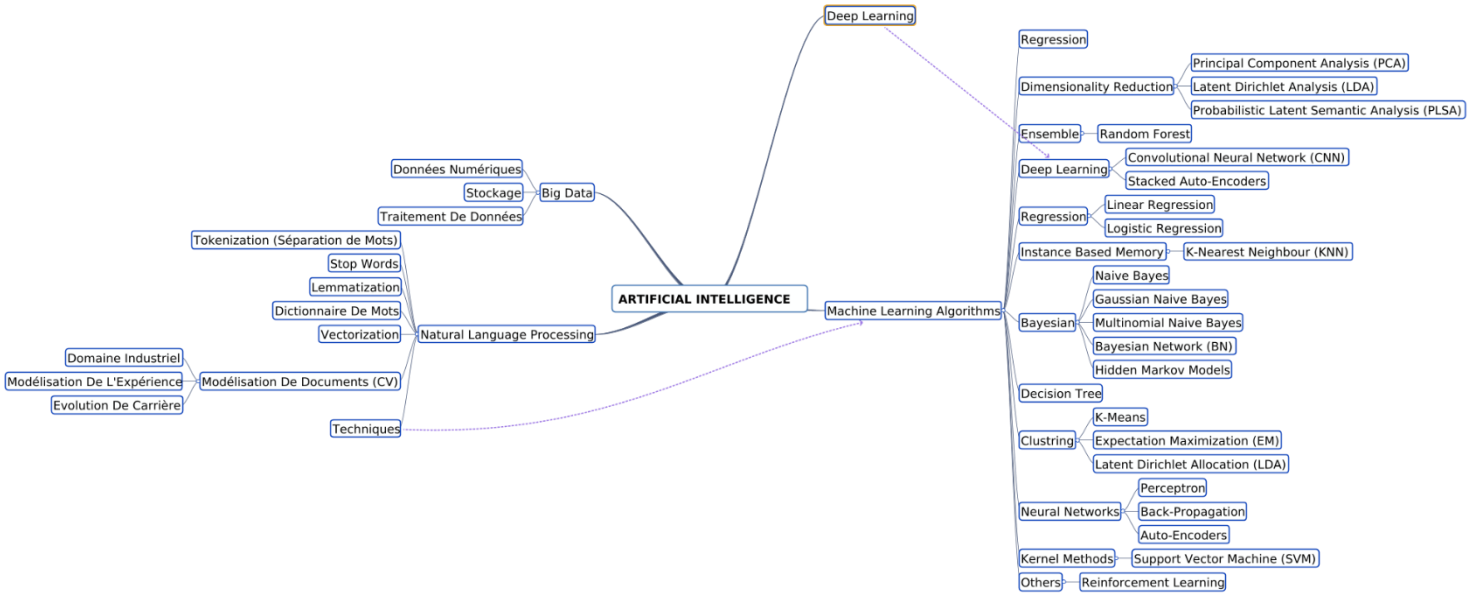
1. Introduction :

De nos jours, les CV font partie intégrante du monde professionnel, d'où l'intérêt que porte les entreprises à l'automatisation du traitement de ces derniers.

Dans ce projet, nous souhaitons développer des modèles capables de prédire le domaine industriel d'un individu quelconque on se basant uniquement sur le contenu de son CV (expériences, compétences …), pour ce faire nous utiliserons des algorithmes de Machine Learning adaptés aux problèmes de NLP. Nous voulons aussi pouvoir jauger le niveau d'expertise d'un CV et pour cela nous allons recourir à différentes techniques se basant sur le Deep Learning.

Le but est donc de pouvoir démêler les différents profils et ce de manière automatique, ce qui nous donnera accès par la suite à de nombreuses utilisations pratiques, telles que la recommandation des différents profils aux entreprises, ou la prédication de l'évolution de carrière.

2. Les mots clés retenus :



3. Descriptif de la recherche documentaire :

Notre projet porte sur le traitement automatique des CV, nous avons donc débuté notre recherche par le traitement automatique de texte. Pour ce faire, nous avons utilisé le moteur de recherche Google, qui nous a fait prendre conscience qu'on était face à toute une branche de la science des données. Par rebond, nous somme vite arrivés sur le traitement automatique du langage naturel (TAL).

Nous avons utilisé Google Scholar pour entamer notre recherche sur le TAL ce qui a permis d'accéder à une multitude d'articles et publications scientifique traitant ce sujet. Nous avons opté pour cet outil car les résultats renvoyés par Google n'étaient pas très pertinents, un mélange de forums et de guides pour débutants. Nous avons par la suite sélectionné les publications des chercheurs réputés dans le domaine du TAL mais aussi de ceux recommandé par nos encadrants. Pour approfondir notre recherche, nous avons utilisé le portail documentaire de sorbonne université, qui nous a donné accès à toute une panoplie d'ouvrages traitant sur le TAL.

La principale difficulté rencontrée est que les articles traitant précisément notre sujet sont assez rares et ce car l'intérêt porté par la communauté scientifique à ce dernier est récent.

4. Bibliographie produite dans le cadre du projet :

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv :1810.04805, 2018.

[2] Charles-Emmanuel Dias, Vincent Guigue, and Patrick Gallinari. Passé, présent, futurs : induction de carrières professionnelles à partir de cv. In CORIA, pages 281–296, 2017.

[3] Clara Gainon de Forsan de Gabriac, Vincent Guigue, and Patrick Gallinari. Resume: A robust framework for professional profile learning & evaluation. In European Symposium on Artificial Neural Networks, 2020.

[4] Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. Multiple-attribute text rewriting. In International Conference on Learning Representations, 2019.

[5] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), volume 1, pages 2227–2237, 2018.

[6] THOMAS HOFMANN. *Unsupervised Learning by Probabilistic Latent Semantic Analysis*. Kluwer Academic Publishers, 2001, https://www.cs.helsinki.fi/u/vmakinen/stringology-k04/hofmann-unsupervised_learning_by_probabilistic_latent_semantic_analysis.pdf. Machine Learning, 42, 177–196.

[7] Gui-Rong Xue, Wenyuan Dai, Qiang Yang et Yong Yu. *Topic-bridged PLSA for Cross-Domain Text Classification*. Association for Computing Machinery New York, United States, juillet 2008, http://info7.cse.ust.hk/~qyang/Docs/2008/fp352-xue.pdf.

[8] McAuley Julian et Jure Leskovec. *From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews*. 18 mars 2013, http://infolab.stanford.edu/~julian/pdfs/www13.pdf.

[9] Nicolas Béchet, Mathieu Roche et Jacques Chauché. *ExpLSA et classification de textes*. 2008, https://hal-lirmm.ccsd.cnrs.fr/file/index/docid/335878/filename/jadt08_NB_MR_JC.pdf. HAL.

[10] Di Wang, Marcus Thint et Ahmad Al-Rubaie. Semi-Supervised Latent Dirichlet Allocation and Its Application for Document Classification. IEEE. 2012, https://ieeexplore.ieee.org/abstract/document/6511698

[11] Yaoyong Li, Kalina Bontcheva, Hamish Cunningham. Adapting SVM for Natural Language Learning: A Case Study Involving Information Extraction. 2006. Department of Computer Science, The University of Sheffield, UK. Corpus ID : 174790242, https://gate.ac.uk/sale/nle-svm/svm-ie.pdf

[12] Houda Khrouf, Raphaël Troncy. De la modélisation sémantique des événements vers l'enrichissement et la recommandation. Multimedia Communications Department, EURECOM Campus SophiaTech, 06904 Biot Sophia Antipolis, France. Art, Computer Science Revue d'Intelligence Artificielle, 2014. https://www.eurecom.fr/en/publication/4407/download/mm-publi-4407.pdf

5. **Evaluation des sources** :

a. Evaluation de la source 6 :

L'article intitulé *Unsupervised Learning by Probabilistic Latent Semantic Analysis* présente une forme standard avec introduction, développement et conclusion et est donc conforme à un article scientifique. L'auteur THOMAS HOFMANN est professeur en analyse de données dans le département informatique de l'ETH Zurich, et parmi ses domaines de recherche on retrouve le Machine Learning ainsi que la compréhension du langage naturel. L'article a été publié en 2001 mais l'algorithme traité dans cette publication n'a pas connu de changement majeur et reste viable, la maison d'édition Kluwer Academic Publishers (maintenant rattaché au groupe Springer qui est le troisième groupe au niveau mondial d'édition spécialisé dans le secteur des Sciences, Technologie et médecine) est un média spécialisé dans la publication scientifique. La bibliographie de l'article respecte la norme imposée sur la citation des sources, se référons à des sources fiables après vérification de notre part. L'article traite sur le PLSA, qui correspond à l'analyse sémantique latente et probabiliste. Après la lecture de l'article, son contenu ainsi que sa conclusion semblent cohérents, et les résultats expérimentaux confirme la conclusion.

b. Evaluation de la source 8 :

L'article intitulé *From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews* présente une forme standard avec introduction, développement et conclusion et est donc conforme à un article scientifique. L'auteur McAuley Julian est maitre de conférences dans le département d'informatique de l'université de Californie à San Diego. Ce dernier à de nombreuses publication dans le domaine de la science des données et jouit de plusieurs sponsors de renom (Amazon, Département de la défense des USA …), le co-auteur Jure Leskovec est quant à lui maitre de conférences dans le département d'informatique à l'université de Stanford, spécialisé en data science appliquée sur les réseaux.
L'article a été publié en 2013 et est donc plutôt récent. La bibliographie de l'article respecte la norme imposée sur la citation des sources, se référons à des sources fiables après vérification de notre part. L'article traite sur la modélisation de l'expérience utilisateur dans le but d'améliorer les systèmes de recommandation. Après la lecture de l'article, son contenu ainsi que sa conclusion semblent cohérents.

c. Evaluation de la source 2 :

L'article intitulé *Passé, présent, futurs : induction de carrières professionnelles à partir de cv* présente une forme standard avec introduction, développement et conclusion et est donc conforme à un article scientifique. L'auteur Vincent GUIGUE est maitre de conférences dans le département d'informatique de Sorbonne Université Sciences et Ingénieries. Ce dernier à de nombreuses publication dans le domaine de la science des données et du TAL, Charles-Emmanuel Dias est un docteur travaillant au sein de l'équipe MLIA (Machine Learning and Information Access) au lip6 et le professeur Patrick GULLINARI est aussi membre de cette équipe et est l'un des pionniers dans le domaine du Machine Learning en Europe.
L'article a été publié en 2017 et est donc récent. La bibliographie de l'article respecte la norme imposée sur la citation des sources, se référons ainsi à des sources fiables après vérification de notre part. L'article traite sur l'apprentissage d'un espace latent permettant l'induction de carrières professionnelles. Après la lecture de l'article, son contenu ainsi que sa conclusion semblent cohérents.