|  |  |  |  |
| --- | --- | --- | --- |
| Institut Supérieur des Technologies de l’Information et de la Communication  de Borj Cedria | Master DeRecherche ………………………  (Régime LMD) | Année | 2019/2020 |
| Semestre | 2 |

Cahier des Charges du Mémoire de Master

|  |  |
| --- | --- |
| **Identification du Mémoire de Master** | |
| Intitulé du Mémoire | **Titre anglais:** Learning transaction embeddings via high utility itemsets  **Titre français:**Apprentissage du plongement (embedding) de transactions via des ensembles d’éléments à haute valeur utile |
| Proposant (Encadreur) | Khaled Belghith |
| Structure d’accueil (1) |  |
| Partenaire |  |
| Cadre (2) |  |
| Privilège (3) |  |

1. **Unité/Laboratoire de Recherche,…**
2. **Projet PRF, CMCU, PVR…**
3. **Bourse, contrat,…**

|  |  |
| --- | --- |
| Coordonnées encadreur | |
| Adresse personnelle : 16 av la Qualité de la vie, Chotrana 1, La soukra, Ariana, 2036, Tunis. | Adresse professionnelle : Istic, Borj Cedria |
| Téléphone fixe : | Téléphone portable : +216 29 610 582 |
| E-mail : khaled.belghith@istic.rnu.tn | Fax : |

|  |  |
| --- | --- |
| Coordonnées étudiant (e) | |
| Nom et prénom : | Adresse : |
| |  |  | | --- | --- | | **e-mail :** |  | | Téléphone portable : |

|  |  |  |
| --- | --- | --- |
| Echéances | | |
| Démarrage : | Dépôt rapport final : | Soutenance : |
| Février 2020 |  |  |

|  |
| --- |
| Problématique |
| L’extraction d’ensemble d’éléments (itemsets)à haute valeur utile est un domaine émergent de la data science, qui consiste à découvrir des ensembles d’éléments (itemset) qui ont une grande importance dans les bases de données. L’utilité d’un ‘itemset’ peut être mesurée suivant plusieurs critères comme le profit, la fréquence ou le poids. Une application populaire des ‘itemsets’ à grande valeur utile est de découvrir dans les transactions tous les ensembles d’articles achetés ensemble par les clients et qui ont généré des bénéfices élevés.  Cependant, l’application de ces approchessouffrede problèmes de clarté des données (data sparsity) et de problèmes de grande dimensionnalité.  Pour pallier ces problèmes, une idée serait d’adapterun modèle d’apprentissage automatique (machine learning) pour apprendre ces représentations d’’itemsets’ à haute valeur utile puis à catégoriser efficacement les données sur la base de ces dernières. |

|  |
| --- |
| Objectifs |
| L’objectif de ce mémoire est de développer un modèle pour l’apprentissage automatique du plongement (embedding) des transactions sur la base des ‘itemsets’ à haute valeur utile. |

|  |
| --- |
| Approche méthodologique |
| Notre approche est une extension de l’approche Trans2Vec [1] pour le développement d’un modèle d’apprentissage des plongements des transactions sur la base des ‘itemsets’ fréquents.  En se basant sur les ‘itemsets’ à haute valeur utile et donc avecdes données plus riches nous pensons que notre approche va nous permettre d’obtenir de meilleurs résultats.  Ces deux approches sont une adaptation des travaux réalisés en [3] du word embedding pour le traitement du langage naturel (Natural language processing ou NLP) au domaine du pattern/itemset mining. |

|  |
| --- |
| Plan de travail |
| 1. Étudier les différentes approches pour l’extraction d’ensemble d’éléments (itemsets)à haute valeur utile et choisir un des algorithmes en utilisant comme critère l’efficacité. On préconise le choix de l’algorithmeEFIM qui est un des plus performants. [2]. 2. Étudier l’approche Trans2Vec [1] pour le développement d’un modèle d’apprentissage des plongements des transactions sur la base des ‘itemsets’ fréquents. 3. Développer le modèle pour l’apprentissage automatique du plongement (embedding) des transactions sur la base des ‘itemsets’ à haute valeur utileen combinant les idées implémentées dans l’algorithme d’extraction d’itemsets choisi dans l’étape 1 avec la logique du modèle d’apprentissage des plongements des transactions vues dans l’étape 2.   Après une première phase au cours de laquelle le candidat devra se familiariser avec les outils  de l’approche neuro-flous, il se concentrera à la mise en œuvre d'un algorithme efficace afin  de considérer la plus grande classe possible de systèmes non linéaires en tenant compte des  données de mesures bruitées.  Unevalidationdecetrav  Après une première phase au cours de laquelle le candidat devra se familiariser avec les outils  de l’approche neuro-flous, il se concentrera à la mise en œuvre d'un algorithme efficace afin  de considérer la plus grande classe possible de systèmes non linéaires en tenant compte des  données de mesures bruitées.  Unevalidationdecetr |

|  |
| --- |
| Principales Références bibliographiques récentes |
| 1. Nguyen D., Nguyen T.D., Luo W., Venkatesh S. (2018) [Trans2Vec: Learning Transaction Embedding via Items and Frequent Itemsets](https://link.springer.com/chapter/10.1007%2F978-3-319-93040-4_29). In: Phung D., Tseng V., Webb G., Ho B., Ganji M., Rashidi L. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2018. Lecture Notes in Computer Science, vol 10939. Springer, pp 361-372, 2018. 2. Fournier-Viger., P., Lin, J. C.-W., Truong, T., Nkambou, R. (2019). [A survey of high utility itemset mining](http://www.philippe-fournier-viger.com/Survey_high_utility_itemset2019_draft.pdf).In: Fournier-Viger et al. (eds). High-Utility Pattern Mining: Theory, Algorithms and Applications, Springer (to appear), p. 1-46. 3. Mikolov, Tomas; et al. (2013). [Efficient Estimation of Word Representations in Vector Space](https://research.google/pubs/pub41224/).In 1stInternational Conference on Learning Representations (2013). 4. Code pour word2Vec : <https://code.google.com/archive/p/word2vec/> |

**Date :**

**Noms et signatures**

**L’encadrant Le Coordinateur du Mastère**

.