haute école arc ingénierie
neuchâtel berne jura www.he-arc.ch

Hes·so
Haute Ecole Spécialisée
de Suisse occidentale

PICC

Guillaume CHACUN

Travail de Bachelor 2019

Filière Informatique - Orientation Développement Logiciel et Multimédia

Professeur-e-s: Hatem GHORBEL

Expert-e-s: Maria SOKHN

Description

Ce travail de Bachelor est proposé par l'entreprise Exelop SA, dont l'application PICC consiste à regrouper le savoir-faire d'une entreprise sous la forme d'un graphe. Les nœuds de ce dernier représentent des problèmes ou solutions que les utilisateurs doivent, pour l'instant, ajouter à la main.

Le but de ce travail est ainsi d'automatiser la collecte de ces entrées en allant les puiser directement dans les archives textuelles de l'entreprise. Ces documents peuvent être des rapports, présentations PowerPoint, PVs de séances, manuels d'utilisation, etc. Les entrées textuelles sont donc extraites, puis traitées et classifiées par des algorithmes d'apprentissage approfondi afin de déterminer si elles représentent des problèmes ou solutions.

Ce travail, à titre exploratoire, pave la voie d'un projet plus important qui sera mené par une équipe différente. Ainsi, de nombreuses approches ont été comparées, évaluées et de multiples pistes d'amélioration ont été proposées.

Déroulement

Le travail s'est déroulé de la manière suivante :

- Étude de travaux similaires
- Rédaction d'un état de l'art sur les techniques d'apprentissage profond pour la classification de textes
- Prétraitement et préparation des données textuelles pour leur utilisation
- Calcul de caractéristiques permettant la séparation des données selon leur classe
- Entraînement, évaluation et comparaison de différents algorithmes non supervisés de classification
- Extraction du texte des archives textuelles d'une entreprise
- Amélioration des modèles et algorithmes

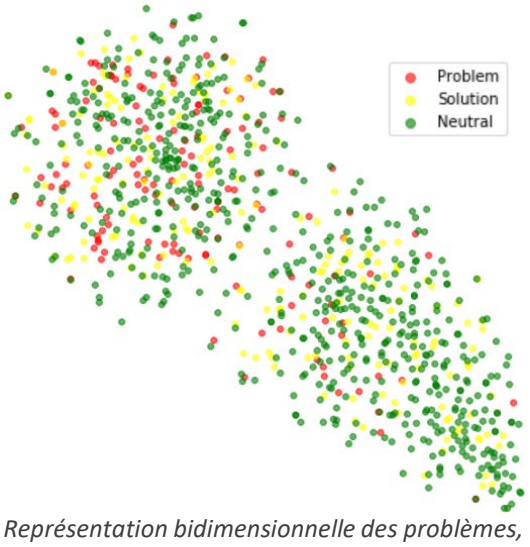
Résultats

Un pipeline permettant l'automatisation de l'entraînement de classificateurs et l'extraction des données textuelles a été développé. Il est capable de traiter l'anglais comme le français.

Représentation bidimensionnelle des problèmes et solutions de la base de données de PICC avant leur classification.

Il permet de classifier très précisément les données existantes de la base de données de PICC.

La classification est, comme prévu, moins efficace sur les entrées extraites des archives textuelles.
Les résultats ont été détaillés, expliqués et des pistes d'amélioration ont été proposées.



Représentation bidimensionnelle des problèmes, solutions et entrées «neutres» extraits des archives textuelles d'une entreprise avant leur classification.

Perspectives

Le pipeline implémenté est fonctionnel et son utilisation a été facilitée. Les classificateurs ont prouvé être performants. La phase d'extraction des données pourrait cependant être améliorée. Le projet étant de nature exploratoire, de nombreuses approches ont été étudiées et comparées. Elles permettent, à l'équipe reprenant ce travail, de poursuivre sur une base solide et avec de multiples pistes d'amélioration.