# NLP pour comptable

Rambier Estelle, estelle.rambier@hotmail.fr, Exact cloud développement, Delft (Pays Bas)

Résumé : *Exact est un logiciel de comptabilité néerlandais. Dans le but d'automatisé le travail de comptable, l'équipe Data Science a développé un modèle capable the catégoriser parmi 900 classes, seulement à partir de la description de relevé bancaire, les transactions dans le bon livret de compte.*

Mots clés : *Neural Network, NLP, word embedding, LSA, Accountancy*

## 1. Introduction

Exact est un logiciel de comptabilité néerlandais. Dans l'optique de réduire les tâches redondantes et ennuyantes, depuis 3 ans, Exact a investi dans la data science. Une de ces tâches avec un potentiel pour l'automatisation est la classification manuelle des transactions bancaire en livre de comptes. Le comptable tout les mois va parcourir ses relevés bancaires et 'ranger' chacunes des transactions dans les bons livres de compte (ex : un plein d'essence irait dans le livre de compte correspondant aux coûts automobile et mon salaire dans le livre de compte correspondant à la charge salarial). Afin de pouvoir automatiser ces processus, nous avons développé une normalisation de ces livres de compte puis un model répondant à la problématique. Le model est en production et offre à nos clients quelques millions de suggestions par mois.

## 2. Méthodologie

Avant de pouvoir espérer classifier ces relevés bancaire, une normalisation est nécessaire. En effet, le comptable a la créativité d'appeler son livre de compte comme il l'entend. Le nombre de transaction par entreprise ne serait pas assez nombreuse pour pouvoir espérer construire un model par companie. La première étape est donc de normaliser les livres de compte correspondant aux mêmes schémas de taxe (https://www.referentiegrootboekschema.nl/). Ceci revient dans un premier temps a mapper tous les livres des compte à travers les 350 000 entreprises a un identifiant . 10% de nos clients ont déjà renseigné ce code unificateur, ceci représente 10 millions de livret de compte. Ils seront utilisés pour développer un model performant capable de catégoriser le 90 million restant. Comme dans de nombreux problème de NLP, le pre-processing est crucial ainsi que le featuring. On a choisi de combiner deux méthodes de featuring permettant d'extraire aussi bien la sémantique que l'importance des mots pour chaque description. Le populaire TF IDF pour comprendre l'importance des mots et FastText word embedding pour tenter d'extraire la signifiance. Obtenant des dimensions extravagantes, on a ensuite utilisé une technique de réduction de dimensions avant de d'entrer les donner dans un neural network.

## 3.Originalité / perspective

A notre connaissance, ce cas d'études appliqué à la compatibilité est une première. Le model en production a fait plus d'un client heureux.  Maintenant quant aux techniques utilisées, elles sont le 'state of the art' en NLP, on n'a pas réinventé la roue, mais on l'a appliqué à un nouveau sujet d'étude.

Aussi, d'un point de vue de l'entreprise, le succès de ce projet a fini par convaincre le reste de la boite de la légitimité de l'équipe data science et a permis de nous apporter beaucoup plus de projet et budget.