# Classification multinomiale et supervisée de comptes rendus d'incidents

Alina Ana-Maria Petrescu

Master thesis in Computer Science

Dans cette thèse, on s'intéresse à la classification discrète multinomiale supervisée des documents. Trois modèles sont proposés, implémentés à l'aide du langage Python et testés grâce aux données mises à notre disposition par l'entreprise TechWan qui utilise actuellement une suite logicielle pour la gestion opérationnelle et la gestion de crises.
Dans ce contexte, les documents sont des comptes rendus ou des exposés des faits décrivant des incidents et leur classification doit être faite en fonction d'un catalogue de catégories d'intervention bien précises. Le but recherché est de prédire la ou les catégories appropriées pour tout nouveau compte rendu à classifier.
Le premier modèle proposé est basé sur la fonction d'ordonnancement BM25. Des définitions ad-hoc par catégorie sont élaborées et la catégorie proposée pour un nouvel exposé des faits est celle qui obtient le plus élevé score de similitude avec cet exposé.
La deuxième approche présentée est basé sur le modèle Doc2Vec et ses deux versions DBOW et DM. Pour chaque version, on crée des modèles qui prennent en compte soit des catégories principales (chacune avec plus de 1000 exposés des faits associés), soit des super-catégories regroupant des catégories principales apparentées. De plus, chaque tel modèle a deux variantes de labellisation des documents : par catégorie ou avec identifiants uniques.
La troisième approche est basée sur le modèle LogisticRegression. En fait, chaque modèle Doc2Vec déjà entraîné infère les vecteurs numériques correspondant aux documents à classifier qui constituent ensuite l'ensemble d'entraînement d'un modèle LogisticRegression associé.
Finalement, on compare les modèles proposés en fonction de leur exactitude. L'approche BM25 s'avère particulièrement intéressante pour des catégories avec un nombre réduit d'exposés des faits associés. Par contre, si les données sont suffisantes pour implémenter un apprentissage automatique, les deux autres modèles sont préférés. Pour l'approche Doc2Vec, on recommande la labellisation des documents sans doublons par identifiants uniques. Les meilleures exactitudes correspondent à l'approche LogisticRegression basée sur un modèle Doc2Vec en version DBOW et avec labellisation par catégorie. En outre, une classification en cascade super-catégories→catégories principales→catégories ordinaires est envisageable.
Finalement, on présente des recommandations pour la rédaction des exposés des faits et la réorganisation de leur taxonomie, ainsi que des suites possibles du travail.

Professeur Docteur Philippe Cudré-Mauroux