Sujet – Généralisation d'un système de notation automatique de productions écrites en français langue étrangère

Nom du tuteur : Dominique CASANOVA

Entreprise ou organisme : Le français des affaires de la CCI Paris Ile-de-France

1. Contexte

Représentant près de 800 000 entreprises, soit 29% du PIB national, la Chambre de commerce et d'industrie de région Paris Ile-de-France (CCIR) est activement engagée aux côtés des acteurs qui font l'économie régionale – qu'ils soient chefs d'entreprise, décideurs, élèves, apprentis, étudiants ou encore salariés en formation continue.

Sur l'ensemble d'un territoire qui comprend Paris, la Seine-et-Marne, Versailles-Yvelines, l'Essonne, les Hauts-de-Seine, la Seine-Saint-Denis, le Val-de-Marne et le Val d'Oise, la Chambre de région a pour missions de représenter les entreprises pour favoriser leur croissance, de former les hommes et les femmes aux défis de demain, de faire grandir les projets d'entreprise et de promouvoir la région capitale pour accroître son rayonnement.

Créé en 1958, Le français des affaires est un acteur historique et pionnier de la certification et de la formation, en français à visée professionnelle. Son ambition est de promouvoir un français utile et professionnel. Sa mission et de promouvoir le français comme outil des échanges économiques. Le département du développement scientifique du Français des affaires veille au développement de la qualité des tests et certifications au moyen d'analyses statistiques et d'une activité recherche appliquée. Il souhaite exploiter dans les pratiques d'évaluation les possibilités offertes par le traitement automatique des langues et l'apprentissage automatique.

2. Profil recherché :

Élève de 3e année d'école d'ingénieur et étudiant de Master 2 en Statistiques / Mathématiques appliquées, vous avez un intérêt prononcé pour l'apprentissage automatique (machine learning) et le traitement automatique des langues, et des compétences éprouvées de programmation en R. Autonome et organisé, vous savez conduire un projet et le finaliser dans les temps. Bon communicant, vous savez présenter la nature de vos travaux à un public non initié et vous intégrer dans une équipe pluridisciplinaire.

3. Mission

Le français des affaires a élaboré un premier système de notation automatique de copies d'expression écrite pour l'une des versions de son test international de français langue étrangère. Votre mission sera de généraliser cet outil aux autres versions du test, en adaptant le recueil et la transformation des variables aux exigences particulières de ces versions et en optimisant les modèles de prédiction. La nouvelle version devra être mise en production à l'issue du stage.

4. Méthodologie envisagée

Analyse de l'existant, recherche de variables caractéristiques complémentaires (traitement automatique des langues, exploitation de corpus), modélisation (apprentissage automatique), test et optimisation, intégration à la chaîne de correction (mise en production).

Programmation en R. Modèles de prédiction privilégiés : SVM, forêts aléatoires, régression logistique ordinale. Base de données : productions écrites d'un test international de français langue étrangère.

5. Résultats attendus

Le programme réalisé génèrera, à partir de fichiers textes comportant les productions écrites de candidats, une prédiction du score et du niveau de chacun des candidats qui sera intégrée à la base de données psychométrique du test d'évaluation de français. Le rapport décrira le(s) modèle(s) retenu(s) pour les prédictions, les variables influençant le plus le résultat et rendra compte de la performance du programme en comparaison avec l'accord entre évaluateurs.

Informations pratiques :

- Durée du stage : 4 à 6 mois entre le 1er janvier 2021 et le 31 juillet 2021
- Gratification : 3.75 € par heure
- Date limite de candidature : 30 novembre 2020
- Contact : Dominique Casanova (dcasanova@cci-paris-idf.fr)