Statistics Canada Statistique Canada

CANSSI INCASS

Machine Learning and Operational Research project for Statistics Canada's Business Register

Description

The Business Register team in Business Survey Methods Division at Statistics Canada is looking for a student to work on research and development to improve the industrial classification of Canadian businesses. Statistics Canada's Business Register (BR) is currently exploring numerous opportunities related to big data, more specifically administrative data. Administrative data may include descriptions related to the industrial activity that are vague, incomplete and vary in time. In order to use the administrative data for statistical purposes, the descriptions have to be converted to standard industrial codes, which is a challenge given the large volume of data, and the availability and quality of auxiliary information.

The student will have to innovate to develop a classification process of the industrial activity of businesses into the current North American Industrial Classification System (NAICS). In particular, the student will have to test different machine learning techniques in order to classify the industrial description data, evaluate the results of these techniques, identify one (or many) technique (if any) leading to results of a sufficient quality in order to be used in regular production. This position offers a unique opportunity to work on a project of high visibility. It is expected that the student will contribute significantly to the development, application and evaluation of machine learning techniques in a key BR component, so it is essential to master the appropriate knowledge in this area. The student will be asked to document the techniques used and the results obtained. The work will be done within a multidisciplinary team with statisticians, economists and programmers.

If time permits, the student could be asked to explore the integration of machine learning to the BR's record linkage processes.

Required knowledge

- Advanced skills with programming languages to do machine learning, such as R, Python, or JAVA as well as other statistical programming languages such as SAS and SAS Enterprise Miner;
- Knowledge of machine learning techniques.

Duration

4 months, with possibility of renewal. The term will begin in early 2018, according to student's availability.

To apply:

Interested candidates should send a current CV including contact information for one reference person to Javier Oyarzun (javier.oyarzun@canada.ca), Senior Methodologist at Statistics Canada.