UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
UNECE Machine Learning for Official Statistics Workshop 2023
5-7 June 2023, Geneva Switzerland
23 May 2023

Using Webdata to derive the Economic Activity of Enterprises

Manveer Mangat, Johannes Gussenbauer, Alexander Kowarik (Statistics Austria)
Manveer.Mangat@statistik.gv.at

Abstract

The economic activity of enterprises (NACE) is often a key characteristic for the production of business statistics. The NACE code is determined for all units in the population and stored in the statistical business register. It is imperative that misclassifications in NACE codes are avoided, since they can lead to seriously biased business statistics. Determining and maintaining the main economic activity of an enterprises is a challenging classification task which relies on manual editing. Accordingly, this leads to the swift depletion of time resources, especially for national statistical institutes with a large business population size. Due to an increasing web presence of enterprises, web data is becoming a viable data source to help classify the economic activity of enterprises. One project within the ESSNet Web Intelligence Network, which started in April 2021, aims to develop automatic procedures to support manual editing of NACE codes. Clearly whether a proposed classification model has the ability to support the manual editing processes will depend on its quality. Hence the purpose of this paper is two-fold: 1) to construct a hierarchical classification model to predict NACE 1-5 codes of enterprises on the basis of their scraped websites and 2) to apply evaluation measures, including a novel customized performance measure, which are more suitable to assess the quality of hierarchical models than the standard evaluation metrics. Our web data encompasses the web pages of enterprises that were part of the Information and Communication Technologies survey from 2019 to 2021.