International Journal of Population Data Science

Journal Website: www.ijpds.org

Automatic coding of occupation and cause-of-death records

Tobin, R[1*], Farrow, E[1], Grover, C[1], and Alex, B[1]

[1]The University of Edinburgh

The Digitising Scotland project aims to digitise 24 million Scottish vital event records of births, marriages and deaths from 1856 to 1973. To use these records effectively for large-scale research they must not only be made machine-readable, but also coded in a form suitable for statistical analysis.

The digitised birth, marriage, and death certificates include textual descriptions of occupations and causes of death. Our aim is to map these descriptions to standard HISCO and ICD-10 codes.

It is impractical to have experts code all the records manually, so we treat the problem as a text classification task and apply machine learning techniques. A proportion of the records will be manually coded and used to train the system. More recent records are already coded and these can also be used for training. Following earlier work by [Kirby et al] and [Carson et al] we are experimenting with Bayesian classifiers for this task.

By combining exact matching for texts that have been seen in the training data and Bayes for the rest, we get an accuracy in cross-validation of 92% for causes of death and 94-97% for occupations.

We are investigating methods to improve this, including automatic spelling correction and synonym detection, use of age and sex information, and (for causes of death) the presence of co-occurring causes.

We are also investigating the value of coarser-grained but more reliable coding, and reporting second- and third-choice codes.

This is work in progress, and the final paper will consider whether the improvements we are making are sufficient to produce useful data for further research. We will also make recommendations about further manual annotation to provide training data covering the whole timespan of the records.

*Corresponding Author:
Email Address: richard@inf.ed.ac.uk (R Tobin)

https://doi.org/10.23889/ijpds.v4i3.1202