

International Journal of Population Data Science

Journal Website: www.ijpds.org



Exploring Text Classification Systems for Automatically Coding Historical Occupations and Causes of Death

Luiza Antonie¹, Peter Christen², Chris Dibben², Jeremy Foxcroft¹, and Lee Williamson²

¹University of Guelph

²University of Edinburgh

Objectives

Text classification models can be used to automatically categorize occupations and causes of death within historical documents. It is important to classify/code these categories as different words or textual descriptions could refer to the same occupation or cause of death. Given the many historical documents that are becoming available for research, accurate classification systems can be valuable resources.

Approach

We explore different text classification techniques, from traditional machine learning to deep learning, and investigate methodologies that transform occupations and causes of death into a vectorial space and use these representations as features to train text classification systems. Our data come from IPUMS USA/International, and SCADR.

Results

Historians have coded occupations and causes of death for some census collections (e.g., US, Canada), but not yet for others (e.g., Scotland). We train and evaluate our classification systems using data from the US and Canada and then deploy it on data from Scotland. We quantitatively measure the performance of the classification systems for historical documents that have codes available. Additionally, once we deploy the model to data that does not yet have codes, we qualitatively evaluate our results by engaging with historians working on those data. We report and discuss these results to understand where the models are performing well and where the models are underperforming.

Conclusions

Results suggest that there is value in building and deploying these classification models. We recommend the use of such models in conjunction with engaging with domain experts.

