ISBN: 9789073592421

OTTAWA 2023
64TH WORLD STATISTICS CONGRESS

CPS Paper

Using Natural Language Processing to Classify Administrative Data of Purchased Products

Author: Mr Gergely Attila Kiss

Submission ID: 549

Reference Number: 549

Presentation File

abstracts/ottawa-2023_6ea7b27b805897fa20ee52feff39d50e.pdf

Brief Description

The Hungarian Central Statistical Office is currently developing new statistical processes to increase the qualities of the household consumption expenditure estimation.

In this innovation multiple data sources are used, most importantly data from Online Cash Registers and Online Invoice System.

We experiment with ML models to identify the appropriate categories.

The new methodology is expected to be useful for other related statistical domains, such as for the improvement of CPI estimation.

Abstract

In cooperation with the National Tax and Customs Administration and the Institute of Agricultural Economics Non-profit Ltd., the Hungarian Central Statistical Office is currently developing new statistical processes to increase the quality of the household consumption expenditure estimates. In this innovative project multiple data sources are used, most importantly data from the Online Cash Registers(OCR) and the Online Invoice System(OIS). Based on these data sources, we aim at building a Machine Learning based methodology to identify the appropriate COICOP and CN categories.

Compared to scanner data our sources do not include bar code identifiers and therefore our current experiments heavily rely on using Natural Language Processing(NLP) techniques on the item names recorded in the purchases. Furthermore, these item names are given by the retailer itself and many times, especially in the OCR, are just abbreviations of the product's name which makes even manual coding difficult. In the literature of NLP the usual approach for short texts are related to twitter data and their processing where using word vectors and TF-IDF representations of the text are the most common. Our data, however, make an exception again as they are tend to be even shorter (our texts are about 5-50 character long) and usually do not have the coherence of a sentence. Thereof, we choose the simplest approaches to preprocess the texts, e.g. One-Hot-Encoding, and frequency based Bag-of-Words.

During our research we experimented with several ML techniques so far to see what could bring an optimal solution in creating the appropriate depth and precision for the classifications. Experiments with Unsupervised models, especially Latent Dirichlet Allocation, show that it cannot provide the necessary depth nor precision for a general solution although it preforms reasonably well for some categories when the text data is detailed enough. However, supervised learning provides a set of tools that performed surprising well on our test samples even though the train sets were small and unbalanced across categories. The three best performing model types are logistic regressions, random forests and multi layer perceptrons. The document will elaborate in details on the mentioned methods and results of the research process.