DSC-18 Categorising contents of lorries in cross-border goods

2018-04-12

The Data Science Campus has been exploring how to process unlabelled list data that are collected manually in an uncontrolled fashion with no supplementary information to allow aggregation of data.

## Team members

* Steven Hopkins
* Gareth Clews
* Arturas Eidukas

## The need

The enabling of analysis on datasets acquired from several ferry operators.

## Impact

The main output is the processing of the datasets into well-structured hierarchical datasets that enable aggregation across categories for analytical understanding of trade flows. The project, on a wider scope, is aiming to open source a generalised tool for these sorts of problems that can be used by analysts to understand similar free-text variables in their own work.

## Data science

The unsupervised processing of free-text using current methods such as word embeddings and clustering algorithms.

## Stakeholders

For processed datasets - Department for Environment Food and Rural Affairs (Defra) and indirectly the cross-Whitehall group on UK trade. For the generalised tool - the analytical community who use Python for natural language analysis.

## Code and outputs

[Optimus](https://github.com/datasciencecampus/optimus) - Github repository [o p t i m u s – turning free-text lists into hierarchical datasets](https://datasciencecampus.ons.gov.uk/o-p-t-i-m-u-s-turning-free-text-lists-into-hierarchical-datasets/)

## Further information

Please contact datasciencecampus@ons.gov.uk for more information.

## Updates

**2019-11-27T09:48:43Z**

Gareth Clews wrote a [report](https://datasciencecampus.ons.gov.uk/o-p-t-i-m-u-s-turning-free-text-lists-into-hierarchical-datasets/) on this work highlighting the findings in September 2018. [Steven Hopkins](https://datasciencecampus.ons.gov.uk/author/steven-hopkins/) wrote a [report](https://datasciencecampus.ons.gov.uk/projects/optimus-a-natural-language-processing-pipeline-for-turning-free-text-lists-into-hierarchical-datasets/) on this work in September 2018 also.