**Measuring sustainability reporting using web scraping and Natural Language Processing**

Alessandra Sozzi; Kimberley Brett; Office for National Statistics

The Big Data team in the Office for National Statistics (ONS) have been exploring the benefits of using Natural Language Processing (NLP) techniques. NLP is concerned with using computer algorithms to understand, and sometimes classify, large volumes of unstructured human text. The speakers will start by providing a short overview of some NLP techniques before covering how these have been used in two projects:

* Using descriptions of properties in housing website data to identify caravan properties and whether they are more likely to be residential or holiday homes. It is vital to understand where people may be living for the Census in 2021 to ensure that everyone in the country is counted. Caravan properties, and whether people are living permanently in them, are not recorded well in many data sources so the team used descriptions from the Zoopla website to add insight about these properties. This work will improve the efficiency of the Census, the largest peacetime operation carried out in the country.
* Applying sentiment analysis techniques to analyse the content of messages for the presence of defined expressions describing whether they were positive/negative, or displayed different emotional states e.g. joy, sadness, fear, anger. Two examples of applications include:
  + Automating feedback from internal staff seminars to understand sentiment.
  + Monitoring the level of daily social sentiment from Facebook comments towards events/topics. As part of the Eurostat task force in Big Data, the project aimed to identify new sources and techniques that can help towards understanding the level of daily satisfaction of citizens.

**Key Words:** Big Data, natural language processing, free text analysis, housing data, survey data

Contact Details for Author 1:

Telephone: +44 (0)1329 444047

Email: alessandra.sozzi@ons.gov.uk