Corporate address thesaurus building using text clustering

Nick Kemp

nmkemp@taz.dstl.gov.uk
DSTL (UK)

Robust and reliable scientometric analyses rely on high quality data. The aim of this ongoing research is to identify methods to support data cleansing for scientometric analysis. Corporate address data within bibliometric databases may contain many variations for individual institutions, which must be standardized to enable research contributions to be correctly assigned. Firstly, a brief review of methods applied to the problem of address standardisation is presented. The production of a thesaurus of Russian universities and research institutes from address data drawn from the Web of Science through a combination of text clustering and manual intervention is then described. Using the thesaurus as 'gold standard' the performances of different text clustering algorithms within the CLUTO clustering package, and character n-gram based methods are assessed.