Methodologies for Big Data at Istat:

State of the Art and Open Challenges

Mauro Bruno[1], Monica Scannapieco[2]

Abstract

In line with the path taken by the European Statistical System, Istat is investing on innovative methods to harness Big Data sources and to use them to produce new and enriched Official Statistics products. Big Data sources are not, in general, directly processable with traditional statistical techniques, indeed the nature and structure of such data sources requires the adoption of new data processing methods. This motivates and justifies the growing interest of National Statistical Institutes in Machine Learning and, more generally, Data Science techniques, which represent a (somewhat) new methodological approach to data analysis.
Istat is currently using Data Science techniques in research and innovation projects, e.g., Big Data experimental statistics. This paper will provide an overview of Istat's more relevant projects in the Big Data ecosystem. It will focus on two specific Big Data-based production pipelines, related to the processing of respectively text sources and imagery sources, thus addressing the variety dimension of Big Data. Later the velocity Big Data dimension, which can be exploited to address the need of publishing timely statistics, will be illustrated through a specific recent project that Istat has been investing on. The paper will also highlight the main open challenges, and, in some cases, the preliminary solutions implemented to solve them.

Keywords: Machine Learning, Natural Language Processing, Text Processing, Image Processing, Big Data.

[1] Tecnologo (Istat), e-mail: mbruno@istat.it.
[2] Dirigente di Ricerca (Istat) – currently on leave at the Italian National Agency for Cybersecurity, e-mail: scannapi@istat.it.