Web scraping, topics and classification: comparing methods

Arjuna Tuzzi and Stefano Sbalchiero
University of Padova

A diachronic corpus is a collection of texts including information on their time period, e.g. the publication date of a post, the publication year of a document, etc. In many situations these texts are arranged into groups (subcorpora) that refer to the same time interval, thus generating a sequence of text sets.

The growing availability of large diachronic corpora and the development of methods to (distant) read them open a new strain of research in time-based analysis of textual data. In the light of the occurrence and co-occurrence over time of words and words' clusters in large amounts of texts, finding methods to assess and trace the words' history would be useful in order to understand the evolution of opinions, ideas, concepts and retrieve which were in the past and which are today the most relevant topics. By means of different approaches and methods (Correspondence Analysis, Reinert's Method, Latent Dirichelet Allocation, Curve Clustering, etc.) scholars aim at finding procedures to retrieve relevant information, summarize contents, find patterns, etc.

This study explores the opportunities of reading the temporal evolution of topics in a large diachronic corpus and aims at comparing and contrasting different methods in order to highlight advantages, limitations and constrains. Results highlight the opportunity of integrating different methods to achieve a thorough understanding of the relevant content embedded in texts.