Introduction to Text Mining

Ludovic Lebart

CNRS, Telecom-ParisTech
46 Rue Barrault 75013 Paris, France
ludovic.lebart@telecom-paristech.fr

Abstract Principal axes techniques and classification methods play a major role in the computerized exploration of textual corpora. They produce visualizations and/or groupings of elements (free responses in marketing and socioeconomic surveys, discourses, scientific abstracts, patents, broadcast news, financial and economic reports, literary texts, etc.); they highlight associations and patterns; they devise decision aids for attributing a text to an author or a period, for choosing a document within a database, for coding information expressed in natural language. They help also to achieve more technical objectives such as lexical disambiguation, parsing, selection of statistical units, description of semantic graphs, speech and optical character recognition. However, the basic concepts of statistical data analysis must be modified in text analysis. Variables, instead of being declared a priori, are derived from the text. Statistical units (or: observations, subjects, individuals, examples) can be documents (described by their titles or abstracts) in documentary databases, respondents (described by their responses to open questions) in surveys, or segments of texts (sentences, context units, paragraphs) in literary applications. Four additional characteristics increase the complexity of the basic data tables: These tables are large (thousands of documents, thousands of words), often sparse (a document may contain a relatively small number of words) and are provided with a huge amount of available meta-data (rules of grammar, semantic networks). Finally, textual data deal with sequences of occurrences (or: strings) of items, whose order could be of importance, another non standard feature in the multidimensional data analysis. We will focus our presentation on the assessments of visualizations, and the use of meta-data. The examples of application concern open-ended questions in an international survey.

References

[1] Lebart, L., Salem, A., Berry, E.: Exploring Textual Data. Kluwer Academic Publisher, Dordrecht (1998).