Evaluation of term-weighting measures for grouped text documents with a target variable: a simulation study

Riccardo Ricciardi[1], Marica Manisera[1]
[1]Department of Economics and Management, University of Brescia, Italy.

Abstract

In Text Mining applications, count-based models are often used to represent text documents. When two document variables are available, i.e. an outcome and a grouping variable, the weight of a word for the documents may depend on the group memberships. The contribution of this work is to frame this context with a statistical approach, by modelling the corpus of documents with a Multivariate Binomial distribution (Hudson et al., 1986). The advantage of this solution is two-fold: it allows (1) to review, in a statistical framework, some term-weighting measures used in the literature (Samant et al., 2019), and (2) to simulate corpora with predefined characteristics by means of the Gaussian Copula method (Genest and McKay, 1986). This simulation is useful to investigate the ability of the existing measures, computed on the group-word interaction, to capture both the group-word relationship itself and the target-word association. Results from the simulation study show interesting relationships that can be exploited by nice visualization tools.

Keywords: *Term-weighting measures; Gaussian Copula; Simulation.*

References

Hudson, William N., Howard G. Tucker, e Jerry A. Veeh. 1986. Limit Theorems for the Multivariate Binomial Distribution. *Journal of Multivariate Analysis* 18(1):32–45.

Genest, Christian, e Jock MacKay. 1986. The Joy of Copulas: Bivariate Distributions with Uniform Marginals. *The American Statistician* 40(4):280–83.

Samant, Surender Singh, N. L. Bhanu Murthy, e Aruna Malapati. 2019. Improving Term Weighting Schemes for Short Text Classification in Vector Space Model. *IEEE Access* 7:166578–92.