The Journal of Open Source Software

TextWiller: Collection of functions for text mining, specially devoted to the italian language

Dario Solari[1], **Andrea Sciandra**[2], **and Livio Finos**[3]

1 Bee Viva srl **2** STAR.Lab - Socio Territorial Analysis and Research, University of Padova **3** Department of Developmental Psychology and Socialisation, University of Padova

DOI: 10.21105/joss.01048

Software

- Review
- Repository
- Archive

Submitted: 24 October 2018
Published: 23 November 2018

Summary

`TextWiller` is the development version of a R package that collects some text mining utilities. It's available at https://github.com/livioivil/TextWiller. The aim of `TextWiller` is to help to deal with the pre-processing of a corpus and it also provides some functions about word classification and polarity. The main quality of this software is to be one of the few text mining R packages in Italian language. Moreover, `TextWiller` can help social media researchers with some specific functions for the data extracted from Twitter via APIs. In particular, `TextWiller` allows to: normalize (Miner, Elder IV, & Hill (2012), Bolasco & De Mauro (2013)) Italian text (stopwords, lower case, punctuation, plurarls, html, emoticons, slang, etc.); get the sentiment (Wilson, Wiebe, & Hoffmann (2005), Ceron, Curini, & Iacus (2014)) of a document, based on an internal lexicon or a custom one; classify users' gender by (Italian) names; classify Italian cities into 5 macro-areas (North East, North West, Centre, South, Islands); find re-tweet (Ferraccioli (2014)) by evaluation of texts similarity (and replace texts so that they become equals); extract short urls and get the long ones; extract users communication pattern. `TextWiller` was designed to be used by researchers (mainly statisticians and social scientists) and by students in courses on text mining (it has already been used in several Bachelor and Master's degree theses).

References

Bolasco, S., & De Mauro, T. (2013). *L'analisi automatica dei testi: Fare ricerca con il text mining*. Carocci Editore.

Ceron, A., Curini, L., & Iacus, S. M. (2014). *Social media e sentiment analysis: L'evoluzione dei fenomeni sociali attraverso la rete* (Vol. 9). Springer Science & Business Media.

Ferraccioli, F. (2014). Topic model workout: Un approccio per l'analisi di microblogging mass media e dintorni - m. Sc. Thesis.

Miner, G., Elder IV, J., & Hill, T. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 347–354). Association for Computational Linguistics.