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** Department of Communication and Economics, University of Modena and Reggio Emilia **3** Department of Developmental Psychology and Socialisation, University of Padova

DOI: 10.21105/joss.01256

Software

- Review
- Repository
- Archive

Submitted: 24 October 2018
Published: 08 September 2019

Summary

`TextWiller` is the development version of a R package that collects some text mining utilities intended for the Italian language. 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.

This software is one of the few text mining R packages in the Italian language. The main differences compared to other popular packages like `tm` are that TextWiller allows sentiment analysis; it includes classification tools for Italian cities and names; and it can help social media researchers with some specific functions for the data extracted from a social networking site via APIs. In particular, `TextWiller` allows to normalize (Miner et al. (2012), Bolasco & De Mauro (2013)) Italian text, i.e. transforming a corpus in a canonical form, useful for text mining. Normalization includes several functions for removing punctuation, stopwords and for managing:

- upper-lower cases;
- plurals;
- URLs and emoticons recognition;
- some slang expressions (which are brought back to the correct form in Italian).

Specifically, other relevant `TextWiller` functions allow to:

- get the sentiment (Wilson, Wiebe, & Hoffmann (2005), Ceron, Curini, & Iacus (2014)) of each document in a corpus, 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-tweets (`RTHound` function; Ferraccioli (2014)) by evaluation of texts similarity (and replace texts so that they become equals) through hierarchical clustering on Levenshtein distance (dissimilarity) matrix;
- extract short URLs and get the long ones;
- extract users' communication pattern, with the `patternExtract` function, based on the '@' sign followed by a username, as in several social media platforms and forums it's used to denote a reply.

`TextWiller` is 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*. Springer Milan. doi:10.1007/978-88-470-5532-2

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., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Elsevier. doi:10.1016/c2010-0-66188-8

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 - HLT 05*. Association for Computational Linguistics. doi:10.3115/1220575.1220619