WebSty – an Open Stylometric System based on Multilevel Text Analysis

Maciej Eder
Institute of Polish Language PAS & Pedagogical University of Kraków maciejeder@gmail.com
Maciej Piasecki
G4.19 Research Group, Wrocław University of Technology maciej.piasecki@pwr.edu.pl
Tomasz Walkowiak
Department of Computer Engineering, Wrocław University of Technology
tomasz.walkowiak@pwr.edu.pl

DARIAH-PL, Digital Philology Working Group

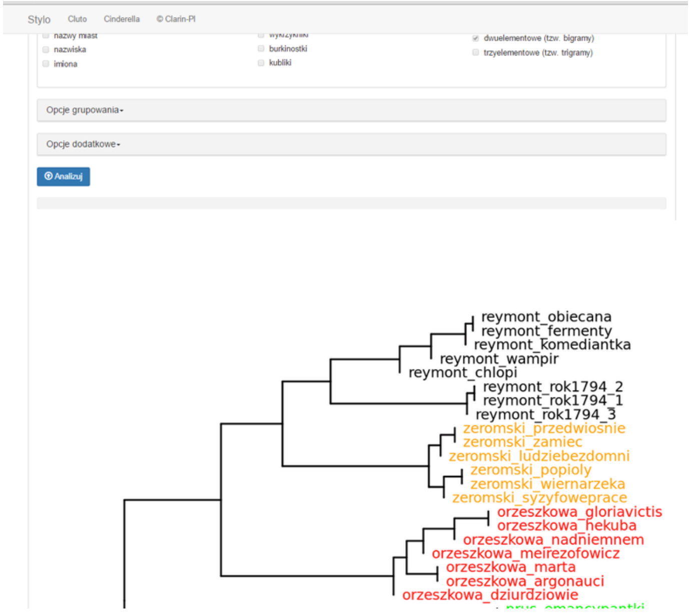
Stylometric techniques are known for their high accuracy of text classification, but at the same time they are usually quite difficult to be used by, say, an average literary scholar. Presumably, stylometry would have been routinely applied in many research tasks in the Humanities, if it had been more accessible to researchers with no programming skills. In this paper we present a general idea, followed by a fully functional prototype of an open stylometric system that facilitates its wide use with respect to two aspects: technical and research flexibility. The system relies on a server installation combined with a web-based user interface. This frees the user from the necessity of installing any additional software. Moreover, we planned to enlarge the set of standard stylometric features with style-markers referring to various levels of the natural language description and based on NLP methods.

Computing word frequencies is simple in English, but in the case of highly inflected languages, characterised by a large number of possible word forms, e.g. Polish, one faces the problem of data sparseness. Thus, it might be better first to map the inflected word forms to lexemes and grammatical attributes, and next to calculate the frequencies of the lexemes. The mapping can be performed by using a morpho-syntactic tagger. Such attributes can be also used as elements of the text document description. Moreover, the documents can be further processed and enriched with Proper Names identification, or even with disambiguated word senses (e.g. as recorded in a semantic lexicon). We will analyse the applicability of the aforementioned language tools to the document description for the needs of stylometry.

The workflow is as follows. Input documents are processed in parallel. The uploaded documents are first converted to uniform text format. Next, each text is analysed by a part-of-speech tagger (we use WCRFT2 for Polish [5]) and then it is piped to a name entity recognizer (in our case it is Liner2 [3]). When the annotation phase is completed for all the texts, the feature extraction module comes into stage (we use the tool Fextor [1]). The extracted raw features can be filtered and transformed by a range of methods. Finally, the R package Stylo [2] or a well known clustering tool called Cluto [6] are run to perform explanatory analysis, e.g. multidimensional scaling (inclusion of other systems is planned). The results obtained in graphical format are displayed by the web browser (see Fig. 1). The web interface allows uploading input documents from a local machine or from a public repository, provides some options of selecting a feature set and options for selecting a grouping algorithm.

For the constructed groups features that are characteristic for them can be identified by a range of algorithms, next presented as the explanation or stored to CSV files.

The system is currently focused on processing Polish. However, as the feature representation is quite language independent, we plan to add converters for the results of application of language tools for other languages.



□ nazwy miast
□ nazwiska
□ imiona
□ wykrzykniki
□ burkinostki
□ kubliki
☑ dwuelementowe (tzw. bigramy)
□ trzyelementowe (tzw. trigramy)

Fig.1 Stylometric system GUI

The web based interface and the lack of the technical requirements facilitates the application of text clustering methods beyond the typical tasks of the stylometry, e.g. analysis of types of blogs [4], recognition of the corpus internal structure, analysis of the subgroups and subcultures, etc.

Bibliography

[1] Broda, B., Kędzia, P., Marcińczuk, M., Radziszewski, A., Ramocki, R., Wardyński, A.: Fextor: A feature extraction framework for natural language processing: A case study in word sense disambiguation, relation recognition and anaphora resolution. Studies in Computational Intelligence, vol. 458, Springer, pp. 41-62 (2013)

[2] Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools. In: Digital Humanities 2013: Conference Abstracts. University of Nebraska--Lincoln, NE, pp. 487-89.

[3] Marcinczuk, M.; Kocon, J., Janicki, M.: Liner2 - A Customizable Framework for Proper Names Recognition for Polish. Studies in Computational Intelligence, vol. 467, pp. 231-253 (2013)

[4] Maryl Maciej. „Kim jest pisarz (w internecie?)” w: Teksty Drugie 2012 nr 6.

[5] Radziszewski, A.: A tiered CRF tagger for Polish, Intelligent Tools for Building a Scien-tific Information Platform. Studies in Computational Intelligence, vol. 467, pp. 215-230 (2013)

[6] Zhao, Ying and Karypis, George. Hierarchical Clustering Algorithms for Document Datasets. Data Mining and Knowledge Discovery, Vol. 10, No. 2, pp. 141 - 168, 2005.