The MIT Press

Journals

Sign In / Register

Books　　Journals　　Digital　　Resources　　About　　Contact

Article navigation

Automatic Text Categorization in Terms of Genre and Author

Efstathios Stamatatos, Nikos Fakotakis and George Kokkinakis

◎ Download Options ›

Journal Resources

Editorial Info
Abstracting and Indexing
Release Schedule
Advertising Info

Abstract　Authors

The two main factors that characterize a text are its content and its style, and both can be used as a means of categorization. In this paper we present an approach to text categorization in terms of genre and author for Modern Greek. In contrast to previous stylometric approaches, we attempt to take full advantage of existing natural language processing (NLP) tools. To this end, we propose a set of style markers including analysis-

Author Resources

Submission Guidelines
Publication Agreement
Author Reprints

Reader Resources

Rights and Permissions
Most Read
Most Cited

More About Computational Linguistics ⌄

Metrics ⌄

level measures that represent the way in which the input text has been analyzed and capture useful stylistic information without additional cost. We present a set of small-scale but reasonable experiments in text genre detection, author identification, and author verification tasks and show that the proposed method performs better than the most popular distributional lexical measures, i.e., functions of vocabulary richness and frequencies of occurrence of the most frequent words. All the presented experiments are based on unrestricted text downloaded from the World Wide Web without any manual text preprocessing or text sampling. Various performance issues regarding the training set size and the significance of the proposed style markers are discussed. Our system can be used in any application that requires fast and easily adaptable text categorization in terms of stylistically homogeneous categories. Moreover, the procedure of defining analysis-level markers can be followed in order to extract useful stylistic information using existing text processing tools.

Forthcoming

154

154 Total citations
26 Recent citations

32 Field Citation Ratio
n/a Relative Citation Ratio

Open Access ⌄

Computational Linguistics Computational Linguistics is Open Access. All content is freely available in

Most Read   See More

🔗 **Lexicon-Based Methods for Sentiment Analysis** (14129 times)
Maite Taboada et al.
Computational Linguistics
Volume: 37, Issue: 2, pp. 267-307

🔗 **Computational Linguistics and Deep Learning** (10558 times)
Christopher D. Manning
Computational Linguistics
Volume: 41, Issue: 4, pp. 701-707

🔗 **Near-Synonymy and Lexical Choice** (3688 times)
Philip Edmonds et al.
Computational Linguistics
Volume: 28, Issue: 2, pp. 105-144

(Note that the Most Read numbers are based on the number of full text downloads over the last 12 months.)

Most Cited   See More

electronic format (Full text HTML, PDF, and PDF Plus) to readers across the globe. All articles are published under a CC BY-NC-ND 4.0 license. For more information on allowed uses, please view the CC license.
Support OA at MITP

Lexicon-Based Methods for Sentiment Analysis (436 times)
Maite Taboada et al.
Computational Linguistics
Volume: 37, Issue: 2, pp. 267-307

A Systematic Comparison of Various Statistical Alignment Models (174 times)
Franz Josef Och et al.
Computational Linguistics
Volume: 29, Issue: 1, pp. 19-51

Opinion Word Expansion and Target Extraction through Double Propagation (147 times)
Guang Qiu et al.
Computational Linguistics
Volume: 37, Issue: 1, pp. 9-27

(Note that the Most Cited numbers are based on Crossref's Cited-by service and reflect citation information for the past 24 months.)
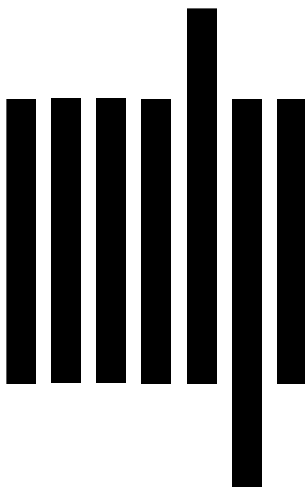
Download Options

Favorite

Sign up for Alerts

Download Citation

RSS TOC

RSS Citation

Submit your article

Support OA at MITP

Journals

Books

Terms & Conditions
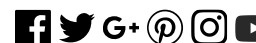
Privacy Statement

Contact Us

US

One Rogers Street Cambridge MA 02142-1209

UK

Suite 2, 1 Duchess Street London, W1W 6AN, UK

Connect

© 2018 The MIT Press
Technology Partner: Atypon Systems, Inc.
CrossRef Member
COUNTER Member
The MIT Press colophon is registered in the

U.S. Patent and Trademark Office.