openMINTED
Open Mining Infrastructure for Text & Data

Facebook.com/openminted

Twitter.com/openminted_eu
#TDMstory

A TEXT AND DATA MINER TELLS HER STORY

THE STRUCTURE OF PAPERS

Iana Atanassova, Centre Tesnière - CRIT, University of Bourgogne Franche-Comté

TELL US ABOUT YOUR PROJECT

We process the full-text content of papers - citation contexts and the Introduction, Methods, Results and Discussion structure (IMRaD) - and use this information to study scientific writing. For example, we identify in-text references that appear more than once per article. This suggests a strong link between the citing paper and the cited paper. From the aggregated results, we produce visualizations that give insights into the structure of papers.

HOW DO YOU ANALYZE AND STORE INFORMATION?

We analyze the text structure and segment the text into sentences. After the initial analysis, we store only those sentences that are of interest for our study (eg. sentences containing in-text references), while keeping all related metadata of the paper. The data is stored locally and only the aggregated results are published in the form of visualizations or statistical analyses.

WHAT HAVE YOU DISCOVERED?

In the Methods section, the average age of references is somewhat lower than in the other sections. This means that the Methods section tends to cite sources that are more recent than those of the other sections. Another important result is the invariant distribution of the in-text references along the IMRaD structure.

CAN YOU EASILY ACCESS MATERIAL FOR TDM?

We work on the full text of scientific articles but many corpora are not available in a format that is ready for processing. Most papers today are published as PDFs and are quite difficult to process. The extraction of full text often generates errors and slows down the research process. The Open Access movement has favored scientific publishing in standardized and machine-readable formats, especially in the bio-medical domain. However, multidisciplinary corpora remain difficult to access.

WHAT IS THE VALUE OF TDM FOR YOU?

Text mining applied to scientific writing is an important tool to understand the process of creating and representing knowledge. Scientific production has become so prolific that researchers can no longer keep track of all publications. For this reason, TDM is indispensable for creating new tools for the efficient exploitation of large-scale corpora and knowledge extraction.

Atanassova and Bertin, Temporal properties of recurring in-text references, D-Lib Magazine Vol 22, September/October 2016

Bertin, Atanassova, Larivière and Gingras. The Invariant Distribution of References in Scientific Papers. JASIST Vol 67(1), January 2016

European Commission

www.openminted.eu

OpenMinTeD is funded by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 654021.

ATHENA Research & Innovation Information Technologies
AGRO-KNOW
EMBL-EBI
gesis Leibniz Institute for the Social Sciences
grnet
MANCHESTER 1824 The University of Manchester
TECHNISCHE UNIVERSITÄT DARMSTADT
UNIVERSITY of STIRLING
BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación
EPFL ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
frontiers
INRA SCIENCE & IMPACT
LIBER
The Open University
UNIVERSITY OF AMSTERDAM
University of Glasgow
CREATe