Biodiversity Observations Miner: A web application to unlock primary biodiversity data from published literature

Gabriel Muñoz[‡], W. Daniel Kissling[§], E. Emiel van Loon[§]

‡ NASUA, Andrade Marin y Diego de Almagro, e7-76, Quito, Ecuador

§ Faculty of Science, Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, Netherlands

Corresponding author: Gabriel Muñoz (nasua.research@gmail.com)

Citation: () . https://doi.org/

Abstract

Background

The amount and diversity of digitally published literature pose many challenges for knowledge discovery and retrieval. A considerable portion of primary biodiversity data is still digitally locked, inside ecological literature which is often stored as pdf files. Large-scale approaches to biodiversity science could benefit from retrieving this knowledge in a standardized form and making it digitally accessible and machine-readable. Text mining has been used extensively for the discovery of data on large quantities of documents. However, the use of text mining in ecology and biodiversity science has been limited compared to other disciplines.

New information

Here, we present a novel open source text mining tool, the **Biodiversity Observations Miner (BOM).** This web application, written in R, allows the semi-automatized discovery of punctual observations (e.g mentioning biotic interactions, data on species functional traits, behavioral and/or natural history descriptions) associated to the scientific names present inside a corpus of scientific literature. This tool aims to increase the digital mobilization of primary biodiversity data and is freely accessible via GitHub or through a web server.

Keywords

biodiversity data, biodiversity knowledge, biotic interactions, data mobilization, natural history, scientific names, text mining

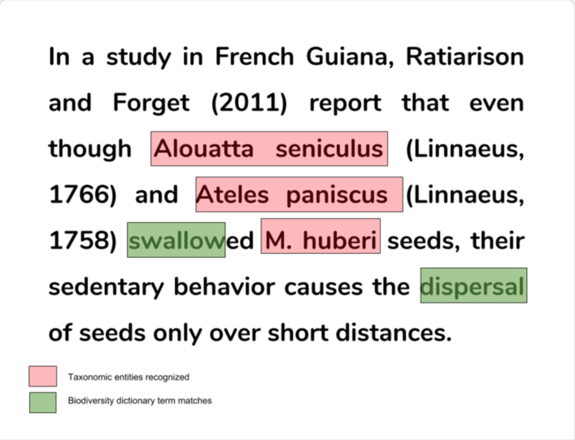
Introduction

Mobilization, digitalization, and interoperability of data on biodiversity are vital to share our global knowledge of nature (Wilkinson et al. 2016 , Kissling et al. 2015 , Edwards 2000). The need for digitally available biodiversity data has resulted in the development of global cyberinfrastructures such as the Global Biodiversity Information Facility

(GBIF: www.gbif.org) (Edwards 2001); the Plant Trait Database (TRY: www.try-db.org) (Kattge et al. 2011); the Data Observation Network for Earth (DataOne: www.dataone.org) (Michener et al. 2011), and Global Biotic Interactions (GloBi: www.globalbioticinteractions.org) (Poelen et al. 2014). Those efforts have made digital biodiversity data increasingly more available in recent years. However, a considerable amount of biodiversity data is still contained inside the current corpus of published literature (Nguyen et al. 2017). This pool of biodiversity data is often stored and shared as PDF files, limiting its interoperability, especially when the amount of literature to examine becomes large. With an increasing availability of literature over the internet, unlocking this biodiversity data to make it digitally interoperable becomes a challenge. Hence, there is a need for developing automatic and semi-automatic computational tools to discover and mobilize biodiversity data contained within this large corpus of literature (Senderov et al. 2017).

Text mining is a computational technique used for the automatic and semi-automatic discovery of useful information from large quantities of text (Hearst 2012). In bio-medicine research, text mining is applied for time demanding tasks such as document classification, the discovery of novel potential protein functions and protein-protein interactions (Petrič and Cestnik 2014, Saffer and Burnett 2014, Tari and Patel 2014). Biodiversity data stored within literature can be found stored in scientific articles (Thessen et al. 2012) or published in books and monographs (Kissling et al. 2014a). Recently, Algorithms and Application Programmatic Interfaces (API's) have been developed for the recognition of taxonomic entities and semantic tagging of ecological literature (Nunez-Mir et al. 2016, Pyle 2016, Sautter et al. 2006, Thessen et al. 2012). Furthermore, as ecology moves towards a data-driven science (Michener and Jones 2012), interest in the use of text mining frameworks for data discovery is growing (Miller et al. 2012 , Thessen et al. 2012, Thessen and Parr 2014, Nunez-Mir et al. 2016 , Senderov et al. 2017, Nguyen et al. 2017 , Senderov et al. 2018).

Biodiversity Observations Miner (BOM) is a web application that identifies snippets of text potentially containing biodiversity data within a given corpus of literature. BOM considers as a biodiversity observation any biological statement linked to one or more taxonomic entities (e.g. species names, genus, family). BOM finds this statement as text containing co-ocurrences of mentions of taxonomic entities with specific vocabulary terms describing a particular observation in biodiversity. These vocabularies are provided in BOM as biodiversity dictionaries. For example, the written description of the plant-animal interaction of frugivory might include terms such as fruit, eat, disperse, swallow, etc. (Fig. 1). Currently, biodiversity dictionaries available in BOM only includes frugivory and pollination vocabularies. However, in the near future, we are looking forward to including additional biodiversity dictionaries to include other types of interactions (e.g. parasitism); species traits (e.g. size, color, morphology); or natural history observations (e.g. traveling distances, habitat preferences). Biodiversity Observations Miner is an open source tool, freely accessible via GitHub (BiodiversityObservationsMiner) or via a web server (goo.gl/wt6V9R).



In a study in French Guiana, Ratiarison and Forget (2011) report that even though Alouatta seniculus (Linnaeus, 1766) and Ateles paniscus (Linnaeus, 1758) swallowed M. huberi seeds, their sedentary behavior causes the dispersal of seeds only over short distances.

🟥 Taxonomic entities recognized

🟩 Biodiversity dictionary term matches

Figure 1.

Example of one text snippet resulting from running Biodiversity Observations Miner with O'Farrill et al. (2013) as input. This text snippet (*i.e. biodiversity observation*) contains data about a **frugivory** interaction between plants and animals. Here, biodiversity data comes from the description of the monkeys *Alouatta seniculus* and *Ateles paniscus* being frugivores of *M.huberi* fruits. The terms "swallow" and "dispersal" were part of the frugivory biodiversity dictionary included in BOM. Red boxes highlight the taxonomical entities recognized using the Global Names Architecture API implemented with the taxize (Chamberlain and Szöcs 2013) R package. The green boxes show the matches of frugivory dictionary terms within the text snippet. Biodiversity Observations Miner indexes individual text snippets by finding co-occurrences such as this example. The final length (in characters) of the resulting text snippet is set by the user.

Project description

Design description: The literature corpus to be mined with Biodiversity Observations Miner is provided by a user and as a collection of individual PDF files. A user can select and retrieve information (as indexed snippets) about particular taxa of interest. To search for taxonomic entities inside the pdf corpus, Biodiversity Observations Miner makes use of the Global Names Recognition and Discovery (GNRD) API (Pyle 2016, Mozzherin et al. 2017). This API is part of the Global Names Architecture (GNA), a name-based cyberinfrastructure which offers a set of open and free web services to find, index and organize biological scientific names (Pyle 2016). Higher taxonomic ranks are retrieved by querying to the National Center of Biotechnology Information (NCBI) taxonomic database using the E-utilities RESTful API of NCBI. Functions to connect to both API's are implemented in the R package `taxize` (Chamberlain and Szöcs 2013).

The application indexes the co-occurrences of taxonomic entities and dictionary terms when finding biodiversity data using the provided biodiversity dictionaries. Terms composing the vocabulary of this dictionaries were hand-picked from a single term-frequency matrix from a collection of scientific articles known to contain relevant observations in frugivory and pollination. The frugivory dictionary was further pruned and improved after discussion among GM and WDK. A skip-n-gram (Huang et al. 1993, Mikolov et al. 2013) probabilistic model was applied to infer the context of each text snippet. This model calculates the likelihood of association of every word pair found in the text within a moving window. The skip-n-gram model is a practical, powerful model to infer context by calculating the likelihood of association among words of a particular piece of text. In Natural language processing (NLP), skip-n-gram models are usually applied in processes such as speech recognition (Huang et al. 1993, Mikolov et al. 2013). The value of "n" in the model defines the size (as a number of words) of the moving window applied to find word vectors in continuous text. This word vectors are constructed by selecting word pairs composed of a fixed word and all other possible combinations of words inside the moving window (Fig. 2). This is done for all the text, advancing the moving window one word at a time. Hence, a probabilistic measure of association can then be calculated for each distinct word vector based on its relative frequency. These probabilities can be normalized to evaluate the likelihood of a word appearing next to another one. Biodiversity Observations Miner applies this model to a sub-corpus of all indexed text snippets with a moving window size (n) of 6 words. Functions from the tidytext (Silge and Robinson 2016) and tidyverse (Lüdecke 2017) R packages were used to build the skip-n-gram model. Individual word vectors can then be matched back to the text, with the user setting the length of the returning snippets (Fig. 3).

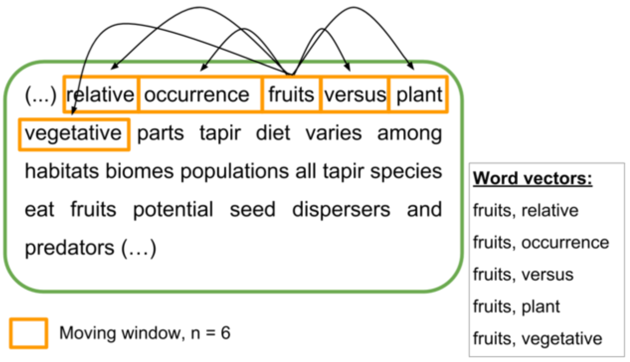
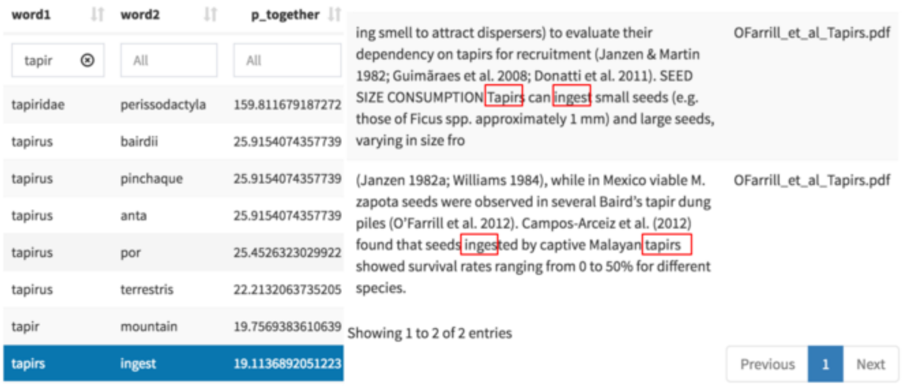

(...) relative occurrence fruits versus plant vegetative parts tapir diet varies among habitats biomes populations all tapir species eat fruits potential seed dispersers and predators (...)
Moving window, n = 6
Word vectors:
fruits, relative
fruits, occurrence
fruits, versus
fruits, plant
fruits, vegetative

Figure 2.

Example of a moving window of n = 6 of a skip-n-gram model over a piece of text from O'Farrill et al. (2013). The text has been cleaned of common stop words (e.g. "the", "all", "however"). Inside the moving window, a central word is fixated (randomly) and all possible word pairs are considered as word vectors. After this step is completed, the moving window advances one word and repeats the process again. Individual frequencies of word pairs within the pool of word vectors are further used to calculate the likelihood of association between specific word pairs.



word1	word2	p_together
tapir	All	All
tapiridae	perissodactyla	159.811679187272
tapirus	bairdii	25.9154074357739
tapirus	pinchaque	25.9154074357739
tapirus	anta	25.9154074357739
tapirus	por	25.4526323029922
tapirus	terrestris	22.2132063735205
tapir	mountain	19.7569383610639
tapirs	ingest	19.1136892051223

ing smell to attract dispersers) to evaluate their dependency on tapirs for recruitment (Janzen & Martin 1982; Guimãraes et al. 2008; Donatti et al. 2011). SEED SIZE CONSUMPTION Tapirs can ingest small seeds (e.g. those of Ficus spp. approximately 1 mm) and large seeds, varying in size fro	OFarrill_et_al_Tapirs.pdf
(Janzen 1982a; Williams 1984), while in Mexico viable M. zapota seeds were observed in several Baird's tapir dung piles (O'Farrill et al. 2012). Campos-Arceiz et al. (2012) found that seeds ingested by captive Malayan tapirs showed survival rates ranging from 0 to 50% for different species.	OFarrill_et_al_Tapirs.pdf

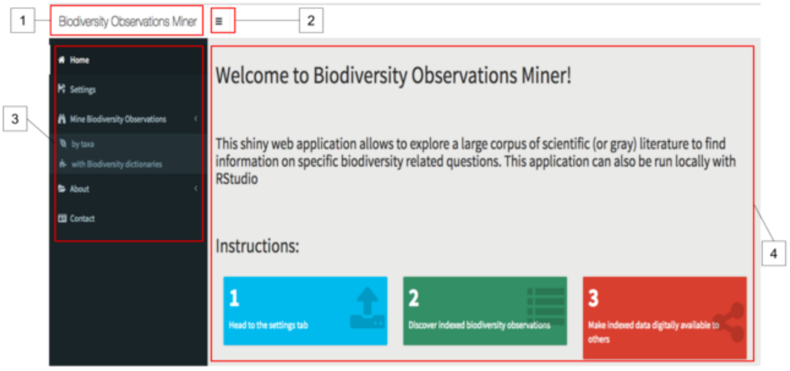
Showing 1 to 2 of 2 entries

Figure 3.

Example of the use of word pairs to classify and retrieve indexed text snippets. The table at the left shows the first 8 word pairs containing the term "tapir" in O'Farrill et al. (2013) text. In the example, the row containing the pair of words "tapirs" and "ingest" is selected. At the right, a data table shows the snippets in O'Farrill et al. (2013) where such word pairs occur. From those text snippets a user can manually derive frugivory related biodiversity observations like 1) Tapirs eats Ficus; 2) Baird's tapir eats M. zapota; 3) Malayan tapirs diet info can be found in Campos-Arceiz et al. (2012). The "p_together" column shows the normalized probability of finding those word pairs together in the text (i.e. how likely a word is to appear *nearby* another one). The word pair table can be filtered to show just the word pair associations which match only the given biodiversity dictionary terms. The resulting length of the text snippet can be modified by the user for both ends.

Application Structure:

Biodiversity Observations Miner was written in R (R Development Core Team 2015) using the shiny (Chang et al. 2017) R package. Application user interface (UI) was built using the shiny-dashboard R package (Chang and Borges Ribeiro 2018). This interface provides a collapsible sidebar menu at the left while the body page is at the right (Fig. 4). The sidebar menu of BOM contains five tabs. Each tab gives access to a determined functionality, rendered in the body page of the application. For a better user experience, we recommend keeping the sidebar menu collapsed while exploring the results of the text mining.



1
Biodiversity Observations Miner
≡
2
3
Home
Settings
Mine Biodiversity Observations
by taxa
with Biodiversity dictionaries
About
Contact
4

Welcome to Biodiversity Observations Miner!

This shiny web application allows to explore a large corpus of scientific (or gray) literature to find information on specific biodiversity related questions. This application can also be run locally with RStudio

Instructions:

1 Head to the settings tab

2 Discover indexed biodiversity observations

3 Make indexed data digitally available to others

Figure 4.

The Biodiversity Observations Miner homepage (goo.gl/wt6V9R). The shiny dashboard was used to create the user interface (UI). Sections of the UI are numbered in the figure as follows: 1) Application title section; 2) Button to collapse the sidebar menu; 3) Sidebar menu; 4) Body page of the application.

"Home" tab:

This provides welcome information and basic instructions for use for the application. It also mentions the license (CC BY-ND 4.0) which allows to share and adapt the material given attribution, non-commercial derivatives and the same license as original when distributed with an additional contribution.

"Settings" tab:

Clicking on this tab will render a series of boxes with widgets to upload PDF file(s) from a local drive as well as different options to connect to both GNA and NCBI APIs via taxize and selecting biodiversity dictionaries. Currently, to avoid issues of uploading heavy pdf files (which most probably contain text stored as images) a total file size limitation of 30Mb applies. However, if needed (e.g. to mine a very large corpus) this limitation can be changed by modifying the source code, either by changing line 27 of the server.R file (when running on a local computer) or by GitHub upon via issues or a push request (when running on the internet server).

"Mine biodiversity observations" tab:

This menu tab has two menu sub-items. First, the "by taxa" will display two boxes. The list of taxonomic entities recognized in the text is provided as a clickable data table, rendered inside the first box. Selecting items from this table will render another data table containing the snippets of text of the corpus corresponding to the selected taxonomic entity. Taxonomic entities could range from scientific names to mentions of higher taxa. We advise consulting the Global Names Architecture (GNA) documentation (http://globalnames.org/docs/home) to find details on how these entities are recognized in the text. The length of the text snippets, measured in characters to the left and to the right from the matching position, can be modified using the sliders. The default is a length of 100 characters to the left and right, respectively.

In the second sub-tab "with Biodiversity dictionaries", a user can explore the co-occurrences of taxonomic entities with dictionary terms in the text. The user has to select the article to be explored. The "Find Word Associations" action button will trigger a function that uses takes the previously indexed text and applies the skip-n-gram model to it. Word pairs and its normalized probabilities of association are displayed inside a clickable data table in the first box. The user can use a radio button to filter word pairs matching only the dictionary terms. Finally, clicking on a row of the table will render the corresponding snippet of text containing the selected word association in the second box. As before, the length of the snippet can be manually set with sliders (*t*he default is 100 characters to each side).

"About" tab:

This tab contains three additional sub-tabs: The first sub-tab "How does it work" provides general information about the web application workflow and R packages needed for running the application locally. Moreover, we provide of sources to find literature on the web as well as ways to download it. In addition links to global repositories of biodiversity data are mentioned and literature references on the text mining biological literature. The second sub-tab "Report a bug" shows direct links to report bugs in the application by filing a GitHub issue, GitHub issues can also be used to get feedback from users to improve functionalities of Biodiversity Observations Miner. The third sub-tab "Source Code" contains a link to the GitHub repository for all the code and resources used to build this web application. In this repository, we will be accepting commits of code and new biodiversity dictionaries from the public.

"Contact" tab:

Here there is available contact information for all authors of Biodiversity Observations Miner.

Web location (URIs)

Homepage: https://fgabriel1891.github.io/BiodiversityObservationsMiner/

Download page: https://fgabriel1891.github.io/BiodiversityObservationsMiner/

Bug database: https://github.com/fgabriel1891/BiodiversityObservationsMiner/issues/

Technical specification

Platform: shiny, R.

Programming language: R

Operational system: Windows, OSx, Linux

Interface language: shiny-dashboard, shiny

Repository

Type: Git

Browse URI: BiodiversityObservationsMiner

Usage rights

Use license: Other

IP rights notes: **Attribution-NoDerivatives 4.0 International (CC BY-ND 4.0)**

Implementation

Implements specification

Biodiversity Observations Miner web application can be run locally **(GitHub repo: "BiodiversityObservationsMiner")** with RStudio or by a web server (**goo.gl/wt6V9R**).

Audience

The target audience of this web application includes ecologists and biodiversity scientists at all career stages. The user interface of this application is straightforward to use and allows the user to make use of developed R packages to mine text and API's without the need of coding experience. Additionally, this application invites developers (ecologists or not) to suggest ideas for improvement. We are open to discussing additional ideas or new tools to expand the current functionalities of this web application.

Additional information

Application structure description

ui.R :

Inputs:

file1: fileInput widget, accepts one or more files with .pdf extension

SnamesOnly: checkboxInput widget, logical operator. If FALSE will render a conditional input to give more options to the user (conditionalInput, conditionalInput2).

GoButton: actionButton widget, render as the "Get Taxa" button.

up: sliderInput widget, ranges from 0-500. Sets the value for the number of characters up the matching position

down: sliderInput widget ranges from 0-500. Sets the value for the number of characters down the matching position

dictionary: selectInput widget, display a dropdown list of biodiversity dictionaries available

indexButton: actionButton widget, renders the "Index" button. Triggers Index events based on the biodiversity dictionaries

skGram: actionButton: actionButton widget, renders the "Find Word Associations" button. Triggers skipGram associations related functions.

up2: same as up.

down2: same as down.

checkbox: checkboxInput widget, logical argument. Should the output of the skipGram matches show only matches with the dictionary terms (TRUE/FALSE).

server.R:

The server side of this web application is composed of eight outputs and five reactive events.

Reactive Events:

read(): this event gets triggered after a user presses the "Get Taxa" button. This event will read the path to the files uploaded with the "Upload" tab and pass it to the *getSnames()* function.

Index(): this event is triggered by the "Index" button. This event will read the selected biodiversity dictionary (as .csv file) and pass the object to the *corpusIndexText()* function

skipGram(): this event occurs after pressing the "Find word associations" button. It looks for a match between the selected article from the drop-down menu with the results of the Index() event. This is passed to the *findSkipGram()* function.

renderContext(): This reactive event is dependent on the user row selection to the data.table rendered as a partial result in the "Mine by Scientific Names" tab. The scientific name of the selected row is passed to the *giveContext()* function.

renderContext2(): It does of the same as *renderContext(),* however, this event is triggered by the row selection of the data.table with the word associations rendered in the "Match with observation events" tab.

Outputs:

names: This output will read the filenames of the uploaded articles and display them as a dropping list with the selectInput() widget.

conditionalInput: Given that the result from input$SnamesOnly is FALSE, this conditional input will render additional user choices as checkboxes for the taxonomic classification of scientific names found.

conditionalInput2: Given that the result from input$SnamesOnly is FALSE, this conditional input will render additional user choices as checkboxes for databases to query

data_table: This will render a data.table with the scientific names found (plus additional taxonomic classifications) inside the "Select a species" box.

context: This will render a data.table with the indexed text snippets by scientific names that match the row selection of the "Text snippets" box from the "Mine by scientific names" tab.

names2: Similarly as names, this will display the filenames of articles uploaded as a dropping list.

skipGram: This output will render a data.table with the results of the word associations found resulting of the *skipGram()* reactive event. The data.table will show the pair of words found to be associated and the normalized probability of finding those pair of words together in the corpus of indexed snippets result of the *Index()* event.

context2: This, as context, will render a data.table with the indexed snippets of text that matches words from the row selection of the skipGram data.table.

Dependecies

Biodiversity Observations Miner makes use of R packages designed for text mining and base R functions. The taxize package is used to establish the API connection to the Global Names Recognition and Discovery (GNRF) tool. The same package is used for Optical Character Recognition (OCR) of the text in the PDFs is done by GNA using the Google Tesseract Tool. The stringr is used for string manipulation. Details about the specifics of the custom functions written for this application can be found in the Suppl. material 1. In addition, this application requires the following R packages to run locally:

- shiny
- shinydashboard
- stringi
- stringr
- taxize
- tidyverse
- tidytext
- tibble
- widyr
- fulltext
- tm
- DT

Acknowledgements

Biodiversity Observations Miner uses tools from GNA (GlobalNamesArchitecture) implemented in the taxize package. Thanks to Scott Chamberlain for modifications to taxize that improved the functionality of this application. Credits to the developers of the individual packages that Biodiversity Observations Miner is dependent on. Terms composing the pollination biodiversity dictionary were selected in collaboration with Joan Casanelles.

Author contributions

GM developed Biodiversity Observations Miner with guidance, comments and input from WDK and EvL. GM wrote the first draft of the manuscript and WDK and EvL provided input. Terms composing the frugivory interactions dictionary were discussed between GM and WDK.

References

- Chamberlain SA, Szöcs E (2013) taxize: taxonomic search and retrieval in R. F1000Research 2: 191. https://doi.org/10.12688/f1000research.2-191.v2
- Chang W, Cheng J, Allaire J, Xie Y, McPherson J (2017) shiny:Web Application Framework for R. 1.0.5. CRAN. URL: https://CRAN.R-project.org/package=shiny
- Chang W, Borges Ribeiro B (2018) shiny dashboard: Create Dashboards with Shiny. 0.7.0. CRAN. URL: https://CRAN.R-project.org/package=shinydashboard
- Edwards J (2001) The Global Biodiversity Information Facility: An international network of interoperabel biodiversity databases. Joho Chishiki Gakkaishi 10 (4): 58-61. https://doi.org/10.2964/jsik_kj00001039357
- Edwards JL (2000) Interoperability of biodiversity databases: Biodiversity information on every desktop. Science 289 (5488): 2312-2314. https://doi.org/10.1126/science.289.5488.2312
- Hearst M (2012) Text Data Mining. Oxford Handbooks Online https://doi.org/10.1093/oxfordhb/9780199276349.013.0034
- Huang X, Alleva F, Hon H, Hwang M, Lee K, Rosenfeld R (1993) The SPHINX-II speech recognition system: an overview. Computer Speech & Language 7 (2): 137-148. https://doi.org/10.1006/csla.1993.1007
- Kattge J, Diaz S, Lavorel S, Prentice IC, Leadley P, Bönisch G, Garnier E, Westoby M, Reich P, Wright I (2011) TRY-a global database of plant traits. Global change biology 17 (9): 2905-2935. https://doi.org/10.1111/j.1365-2486.2011.02451.x
- Kissling WD, Dalby L, Fløjgaard C, Lenoir J, Sandel B, Sandom C, Trøjelsgaard K, Svenning J (2014) Establishing macroecological trait datasets: digitalization, extrapolation, and validation of diet preferences in terrestrial mammals worldwide. Ecology and Evolution 4 (14): 2913-2930. https://doi.org/10.1002/ece3.1136
- Kissling WD, Hardisty A, García EA, Santamaria M, Leo FD, Pesole G, Freyhof J, Manset D, Wissel S, Konijn J, Los W (2015) Towards global interoperability for supporting biodiversity research on essential biodiversity variables (EBVs). Biodiversity 16: 99-107. https://doi.org/10.1080/14888386.2015.1068709
- Lüdecke D (2017) Data Transformation in R: The Tidyverse-Approach of Organizing Data. Unpublished https://doi.org/10.13140/RG.2.2.32973.33763
- Michener W, Vieglais D, Vision T, Kunze J, Cruse P, Janée G (2011) DataONE: Data Observation Network for Earth - preserving data and enabling innovation in the biological and environmental sciences. D-Lib Magazine 17 https://doi.org/10.1045/january2011-michener
- Michener W, Jones M (2012) Ecoinformatics: supporting ecology as a data-intensive science. Trends in Ecology & Evolution 27 (2): 85-93. https://doi.org/10.1016/j.tree.2011.11.016
- Mikolov T, Sustkever I, Chen K, Corrado GS, Dean J (2013) Distributed Representations of Words and Phrases and their Compositionality. Neural Information Processing Systems (NIPS)., 26. Advances in Neural Information Processing Systems.
- Miller J, Dikow T, Agosti D, Sautter G, Catapano T, Penev L, Zhang Z, Pentcheff D, Pyle R, Blum S, Parr C, Freeland C, Garnett T, Ford LS, Muller B, Smith L, Strader G, Georgiev T, Bénichou L (2012) From taxonomic literature to cybertaxonomic content. BMC Biology 10 (1): 87. https://doi.org/10.1186/1741-7007-10-87
- Mozzherin D, Myltsev A, Patterson D (2017) "gnparser": a powerful parser for scientific names based on Parsing Expression Grammar. BMC Bioinformatics 18: 279. https://doi.org/10.1186/s12859-017-1663-3
- Nguyen NH, Soto A, Kontonatsios G, Batista-Navarro R, Ananiadou S (2017) Constructing a biodiversity terminological inventory. PLOS ONE 12 (4): e0175277. https://doi.org/10.1371/journal.pone.0175277
- Nunez-Mir G, Iannone B, Pijanowski B, Kong N, Fei S (2016) Automated content analysis: addressing the big literature challenge in ecology and evolution. Methods in Ecology and Evolution 7 (11): 1262-1272. https://doi.org/10.1111/2041-210x.12602
- O'Farrill G, Galleti M, Campos-Arceiz A (2013) Frugivory and seed dispersal by tapirs: an insight on their ecological role. Integrative Zoology 8 (1): 4-17. https://doi.org/10.1111/j.1749-4877.2012.00316.x
- Petrič I, Cestnik B (2014) Predicting future discoveries from current scientific literature. Methods in Molecular Biology. https://doi.org/10.1007/978-1-4939-0709-0_10
- Poelen J, Simons J, Mungall C (2014) Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. Ecological Informatics 24: 148-159. https://doi.org/10.1016/j.ecoinf.2014.08.005

- Pyle RL (2016) Towards a Global Names Architecture: The future of indexing scientific names. ZooKeys 550: 261-281. https://doi.org/10.3897/zookeys.550.10009
- R Development Core Team (2015) R Software for statistical computing. 3.4.3 (2017-11-30) - "Kite-Eating Tree".
- Saffer J, Burnett V (2014) Introduction to biomedical literature text mining: context and objectives. Methods in Molecular Biology. https://doi.org/10.1007/978-1-4939-0709-0_1
- Sautter G, Böhm K, Agosti D (2006) A combining approach to find all taxon names (FAT). Biodiversity Informatics 3: 46-58. https://doi.org/10.17161/bi.v3i0.34
- Senderov V, Georgiev T, Agosti D, Catapano T, Sautter G, Tuama ÉÓ, Franz N, Simov K, Stoev P, Penev L (2017) OpenBiodiv: an implementaion of a semantic system running on top of the biodiversity knowledge graph. Proceedings of TDWG 1: e20084. https://doi.org/10.3897/tdwgproceedings.1.20084
- Senderov V, Simov K, Franz N, Stoev P, Catapano T, Agosti D, Sautter G, Morris RA, Penev L (2018) OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system. Journal of Biomedical Semantics 9: 5. https://doi.org/10.1186/s13326-017-0174-5
- Silge J, Robinson D (2016) tidytext: Text Mining and Analysis Using Tidy Data Principles in R. The Journal of Open Source Software 1 (3): 37. https://doi.org/10.21105/joss.00037
- Tari L, Patel J (2014) Systematic drug repurposing through text mining. Methods in Molecular Biology. https://doi.org/10.1007/978-1-4939-0709-0_14
- Thessen A, Parr CS (2014) Knowledge extraction and semantic annotation of text from the Encyclopedia of Life. PLoS ONE 9 (3): e89550. https://doi.org/10.1371/journal.pone.0089550
- Thessen AE, Cui H, Mozzherin D (2012) Applications of natural language processing in biodiversity science. Advances in Bioinformatics 2012: 391574. https://doi.org/10.1155/2012/391574
- Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: 160018. https://doi.org/10.1038/sdata.2016.18

Supplementary material

Suppl. material 1: Functions Reference

Authors: Gabriel Muñoz

Data type: reference for code, new functions, R.

Brief description: References for the custom functions written for the development of Biodiversity Observations Miner web application. Functions were written for the R environment.

Filename: BOMFunctionsReference.pdf - Download file (128.45 kb)