Analyzing the Biological Literature with evoText

UQAM, 1/11/2019

Charles H. Pence @pencechp · @pencelab





Outline

- **1**. Tool development and future work
- 2. Current corpus status
- 3. evoText in action: analyzing biodiversity

The take-home: Could evoText be useful for your work? Get in touch!

Tool Development

Why a New Tool?

Journal articles are weird.

- They're small.
- They carry unusual metadata.
- People are used to searching/indexing them in particular ways.
- They're hard to get and come in an infinite variety of formats.

Tricks

- Have to analyze using only basic, plain text. (No guarantee we can always get anything better.)
- Ways to search/filter by journals, years (categories coming soon)
- Have to be robust to low-quality data (often entirely dependent on publishers for data quality; too many documents for manual cleaning)
- Also coming soon: citation analysis support

Current Infrastructure

Frontend: Ruby on Rails web application **Analysis backend:** Ruby **Article database:** Apache Solr

https://www.evotext.org/

Corporate DNA Corporate DNA Editor's Note (2.1) Editor's Note (2.1) Energence: Complexity and Organization, Vol. 2, No. 1 (volume 2-tissue 1) Energence Why Collisition (Dr krit Richardson Energence: Complexity and Organization, Vol. 2, No. 1 (volume 2-tissue 1) Editor's Note (2.1) Energence Why Collisition (Correct Editor's Note (2.1No. 1 (volume 2-tissue 1)) Editorial (16.4) Peter Alters Energence: Complexity and Organization, Vol. 2, No. 1 (volume 2-tissue 1). Editorial (16.4) Peter Alters Energence: Complexity and Organization, Vol. 2, No. 1 (volume 2-tissue 1). Editorial (16.4) Peter Alters Energence: Complexity and Organization, Vol. 2, No. 1 (volume 2-tissue 1). Editorial (16.4) Peter Alters Energence: Complexity and Organization, Vol. 1, No. 4 (volume 2-tissue 1). Editorial (16.4) Peter Alters Energence: Complexity and Organization, Vol. 1, No. 4 (volume 2-tissue 1). Editorial (16.4) Peter Alters Energence: Complexity and Organization, Vol. 1, No. 4 (volume 2-tissue 1). Editorial (16.4) Peter Alters Energence: Complexity and Organization, Vol. 1, No. 4 (volume 2-tissue 1). Editorial (16.4) Peter Alters Energence: Complexity and Organization, Vol. 1, No. 4 (volume 2-tissue 1). Editorial (16.4) Peter Alters Energence: Complexity and Organization, Vol. 1, No. 4 (volume 2-tissue 1). Editorial (16.4) Peter Alters Energence: Complexity and Organization, Vol. 1, No. 4 (volume 2-tissue 1). Editorial (16.4) Peter Alters Energence: Complexity and Organization, Vol. 1, No. 4 (volume 2-tissue 2). Peter Alters Energence: Complexity and Organization, Vol. 1, No. 4 (volume 2-tissue 2). Peter Alters Energence: Complexity and Organization, Vol. 1, No. 4 (volume 2-tissue 2). Peter Alters Energence: Complexity and Organization, Vol. 1, No. 4 (volume 2-tissue 2). Peter Alters Energence: Complexity and Organization, Vol. 1, No. 4 (volume 2-tissue 2). Peter Alters Energence: Complexity and Organization, Vol. 1, No. 4 (volume 2-tissue 2). Peter Alters Energence: Complexity and Organization,	Corporate DNA Co	We are getting a few job failures on our new infrastructure that we're in the process of diagnosing. If you've to gives us valuable diagnostic information!	ieu jobs that ha	ve raneo, please do retry	y unem as .
Corporate DNA Co	Corporate DNA key and Cognitation, Vol. 2, No. 1 (volume -2-issue -1) Filter search Authors Filter search Authors C. P. Swarnon 428 What and hears from a theory of complexity and Cognitation, Vol. 2, No. 1 (volume-2-issue -1) What and hears from a theory of complexity? What and key and Cognitization, Vol. 2, No. 1 (volume-2-issue -1) Editorial (16.4) Why Businesses Fail Ton Rend, Dr Kurt Richardson, Emergence: Complexity vol. 18, No. 4 (volume-2-issue -1). Market Science 4.1 What and Cognitization, Vol. 18, No. 4 (volume-2-issue -1). What and cognitization, Vol. 18, No. 4 (volume-2-issue -1). Why Businesses Fail Ton Rend, Dr Kurt Richardson (Emergence: Complexity Viet 19, No. 4 (volume-2-issue -1). The Querty Control of Complexity? Proceedings Complexity and Cognitization, Vol. 18, No. 4 (volume-2-issue -1). Proceedings Complexity and Cognitization, Vol. 18, No. 4 (volume-2-issue -1). Proceedings Complexity and Cognitization, Vol. 18, No. 4 (volume-2-issue -1). Proceedings Complexity and Cognitization, Vol. 18, No. 4 (volume-2-issue -1). Proceedings Complexity and Cognitization, Vol. 19, No. 4 (volume-2-issue -1). Proceedings Complexity and Cognitization, Vol. 19, No. 4 (volume-2-issue -1). Proceedings Complexity and Cognitization, Vol. 19, No. 4 (volume-2-issue -1). Proceedings Complexity and Cognitization, Vol. 19, No. 4 (volume-2-issue -1). Proceedings Complexity and Cognitization, Vol. 19, No. 4 (volume-2-issue -1). Proceedings Complexity and Cognitization, Vol. 19, No. 4 (volume-2-issue -1). Proceedings Complexity and Cognitization, Vol. 19, No. 4 (volume-2-issue -1). Proceedings Complexity and Cognitization, Vol. 19, No. 4 (volume-2-issue -1). Proceedings Complexity and Cognitization, Vol. 19, No. 4 (volume-2-issue -1). Proceedings Complexity and Cognitization, Vol. 19, No. 4 (volume-2-issue -1). Proceedings Complexity and Cognitization, Vol. 19, No. 4 (volume-2-issue -1). Proceedings Complexity and Cognitization, Vol. 19, No. 4 (volume-2-issue -1). Procee	Search for articles		¢ Sort ▼	🖬 Savi
		Corporate DNA	More +	1625790 articles	s in databa
Michael Lassak, Dr. Kurt Richardson, Dr. Michael Lissak (Emergence: Complexity and Oganisation, Vol. 2, No. 1 Emergence Emergence Emergence: Minute Volume 2-Issue: 1). What can we learn from a theory of complexity? Paul Cilliers (Emergence: Complexity and Oganisation, Vol. 2, No. 1 Ivolume: Paul Cilliers (Emergence: Complexity and Oganisation, Vol. 2, No. 1 Ivolume: Editorial (16.4) Peter Alten (Emergence: Complexity and Oganisation, Vol. 2, No. 1 Ivolume: Minute Volume: Editorial (16.4) Minute: Minute Volume: Minute Volum	Michael Lusack, Dr. Karl Richardson, Dr. Michael Lusack (Emergence: Complexity and Organization, Vol. 2, No. 1 Emergence Emergence Herry Colditions (Dr. Karl Richardson, Dr. Michael Lusack (Emergence: Complexity) Part Cilliens (Emergence: Complexity)? Part Cilliens (Emergence: Complexity)? Part Cilliens (Emergence: Complexity)? Editorial (L6.4) What can we learn from a theory of complexity? Editorial (L6.4) Why Businesses Fail Tom Rend, Dr Kurt Richardson, Vol. 18, No. 4 (volume-2-lusue-1). Marry of Sources C. P. Swamon 428 Source C. Swamon 428	Ken Baskin, Dr Kurt Richardson Emergence: Complexity and Organization, Vol. 2, No. 1 (volume-2-issue-1).		Filter search	
Vordame 2-Issue-13. Emergence Mathew Collision DK Art Richardson Emergence: Complexity and Organization, Vol. 2, No. 1 (volume - Marked Schward - 77) Article Complexity and Organization, Vol. 2, No. 1 (volume - C. P. Sources) - C. P. Sources - 42 What can we learn from a theory of complexity? Mater in the complexity and Organization, Vol. 2, No. 1 (volume - 2-Issue - 1). Editorial (16.4) Mater in the United Schward - 2-Issue - 1). Mater in the United Schwar	Verdume 2-4 sour -11.	Editor's Note (2.1)	More *	Authors	
Emergence Emerg	Emergence Emerge			Jr.	16143
Intergence	Term Partice Term Partice Term Partice Term Partice Term Partice Term Partice Term Part Contents De Kurt Richardson (Energence: Complexity)? Part Colliters (En	AALMENG-7-122000-77			951
2/stude-11. C. R. Sommon et al. What can we learn from a theory of complexity? Multiliers / Emergence: Complexity and Ogenization, Vol. 2, No. 1 (volume-2-issue-1). Multiliers / Emergence: Complexity and Ogenization, Vol. 2, No. 4 (volume-16-issue-4). Multiliers / Emergence: Complexity and Ogenization, Vol. 16, No. 4 (volume-16-issue-4). Multiliers / Emergence: Complexity and Ogenization, Vol. 16, No. 4 (volume-16-issue-4). Multiliers / Emergence: Complexity and Ogenization, Vol. 16, No. 4 (volume-16-issue-4). Multiliers / Emergence: Complexity and Ogenization, Vol. 16, No. 4 (volume-16-issue-4). Multiliers / Emergence: Complexity and Ogenization, Vol. 1, No. 4 (volume-16-issue-4). Multiliers / Emergence: Complexity and Ogenization, Vol. 1, No. 4 (volume-16-issue-4).	2-Issue 1). C. P. Swamon 22 2-Issue 1). C. P. Swamon 22 2-Issue 1). Sucr 2 2-Issue 1). Sucr 3 2-Issue 1). Sucr 3 2-Issue 1). Sucr 4 2-Issue 1). Sucr 4 2-Issue 1). Sucr 4 2-Issue 1). Sucr 4 Correct States of 4 Correct States Correct States Correct States Correct	Emergence	More *	Bentley Glass	811
C. R. Swanson et al. Source et al.	What can we learn from a theory of complexity? What can we learn from a theory of complexity? What can we learn from a theory of complexity? User a source of the complexity of the complexity? Description of the complexity of t	Jeffrey Goldstein, Dr Kurt Richardson Emergence: Complexity and Organization, Vol. 2, No. 1 (volume-		Rudolf Schmid	d 778
Pad Cillers Emrgance: Complexity and Organization, Vol. 2, No. 1, Ivolume 2-Issue 11. Induce 14.11 Editorial (16.4) Editorial (16.4) Marce - Engrance: Complexity and Organization, Vol. 16, No. 4 (volume 16-Issue -4), Marce	Pad Cillers Energence: Complexity and Organization, Vol. 2, No. 1 (volume 2-lissue 1), Nature 3 e3711 Editorial (16.4) Editorial (16.4) Peter Alen Energence: Complexity and Organization, Vol. 16, No. 4 (volume 16-lissue 4), Academy of Selected States of Academy of Academy of Selected States of Academy of Academy of Academy of Selected States of Academy of Ac			C. P. Swansor	428
Editorial (16.4) Proceedings 11377 cf the National Proceedings 11377 cf the National Academy of Sciences cf the United Science Academy of Sciences cf the United Science America Minus Current Science 11 Minus Current Scien	Editorial (16.4) Feter Alen J Emergence: Complexity and Oganization Vol. 16, No. 4 (volume-16-issue-4), Why Businesses Fail Tom Rand, Dr Kurt Richardson J Emergence: Complexity and Oganization, Vol. 1, No. 4 (volume-16-issue-4), Tom Rand, Dr Kurt Richardson J Emergence: Complexity and Oganization, Vol. 1, No. 4 (vol1)sr44,	What can we learn from a theory of complexity?	More -	Journal	
Editorial (16.4) Mon = Peter Alten J Energance: Complexity and Organization, Vol. 16, No. 4 (voltame-16-issue-4), Mon = VMVB Balanceses Fail Mon = Mine multi Charlandono (Energance: Complexity and Organizations, Vol. 1, No. 4 (voltame), 4	Editorial (16.4) Burn - Anderson - States - St	Paul Cilliers Emergence: Complexity and Organization, Vol. 2, No. 1 (volume-2-issue-1),		Nature	363791
Academy of Sciences Academy Academy of Sciences Academy A	Here Allen [Energence: Complexity and Organization, Vol. 16, No. 4 (voltme-16-issue-4), Academy of Silences or the Eucled States of America Why Businesses Fail Tom Rend, Dr Kurt Richardson [Emergence: Complexity and Organization, Vol. 1, No. 4 (vol1is/4), The Quarter Science: 1117 The Quarter Science: 1117	Editorial (14-4)			
the United States of the United States of America America America America America Current Science: sust:	Why Businesses Fail Why Businesses Fail Why Gusinesses Fail Current Science: 54127 Tom Rind, Dr Kurt Richardson J. Rengence: Complexity and Organization, Vol. 1, No. 4 (vol.11s.94), The Quarterly 20227			Academy of S	iciences
Tom Rand, Dr Kurt Richardson Emergence: Complexity and Organization, Vol. 1, No. 4 (vol1iss4),	Tom Rand, Dr Kurt Richardson Emergence: Complexity and Organization, Vol. 1, No. 4 (vol1iss4), The Quarterly 22837				States of
	The Quarterly 32837	Why Businesses Fail	More +	Current Scien	ce 56187
		Tom Rand, Dr Kurt Richardson Emergence: Complexity and Organization, Vol. 1, No. 4 (vol1iss4),		The Oursterlu	

Coming soon: New version!

Frontend: Angular web app (new interface!) Analysis backend: Go (faster analysis!) Article database: Solr (or maybe MongoDB?)

RLetters

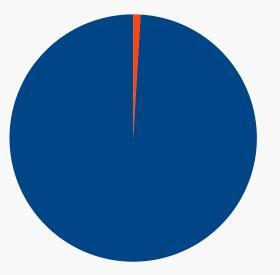
Home	Search.		Save as dataset 47 Sort *
Q	Search Results		Filter Results
Search	Whitehorn, James et al. 2012. 'Prophylactic Platelets in Dengue: Survey Responses Highlight Lack of an Evidence Base.' PLoS Neglected Tropical Diseases 6(6):e1716.		Author
Browse	Kirko, Brian H. et al. 2012. "Burkholderia pseudomallei Known Siderophores and Hemin Uptake Are Dispensable for Lethal Murine Melioidosis." PLoS Neglected Tropical Diseases 6(6):e1715.	More	Peter J. Hotez Jürg Utzinger T Marleen Boelaert Donald P. McManus 18
	Gower, Emily W. et al. 2012. "Definitions and Standardization of a New Grading Scheme for Eyelid Contour Abnormalities after Trichiasis Surgery." PLoS Neglected Tropical Diseases 6(6):e1713.	More	Philippe Büscher 15 Show more
	Gopal, Hemavathi et al. 2012. "Oligonucleotide Based Magnetic Bead Capture of Onchocerca volvulus DNA for PCR Pool Screening of Vector Black Files." PLoS Neglected Tropical Diseases 6(6):e1712.	More	Journal PLoS Neglected Tropical Diseases 1500 Actually a Novel
	Duppre, Nadia C, et al. 2012. "Impact of PGL-1 Seropositivity on the Protective Effect of BCG Vaccination among Leprosy Contacts: A Cohort Study." PLoS Neglected Tropical Diseases 6(6):e1711.	More	Show more Publication Date
	de Souza Leoratti, Fabiana Maria et al. 2012. "Neutrophil Paralysis in Plasmodium vivax Malaria." PLoS Neglected Tropical Diseases 6(6):e1710.	More	2011 443 2010 350
	Savage, Amy F. et al. 2012. "Transcript Expression Analysis of Putative Trypanosoma brucei GPI-Anchored Surface Proteins during Development in the Tsetse and Mammalian Hosts." PLoS Neglected Tropical Diseases 6(6):e1708.	More	2012 202 2009 224 2008 179
	Carrasco, Hernán J. et al. 2012. "Geographical Distribution of Trypanosoma cruzi Genotypes in Venezuela." PLoS Neglected Tropical Diseases 6(6):e1707.	More	Show more
	Escadabl, Camille et al. 2012. "First International External Quality Assessment of Molecular Detection of Crimean-Congo Hemorrhagic Fever Virus." PLoS Neglected Tropical Diseases 6(6):e1706.	More	
	Zihlmann, Karina Franco et al. 2012. "Living Invisible: HTLV-1-Infected Persons and the Lack of Care in Public Health." PLoS Neglected Tropical Diseases 6(6):e1705.	More	
	Katara, Gajendra Kumar et al. 2012. "Evidence for Involvement of Th17 Type Responses in Post Kala Azar Dermal Leishmaniasis (PKDL)." PLoS Neglected Tropical Diseases 6(s):e1703.	More	

Current Corpus

The Corpus

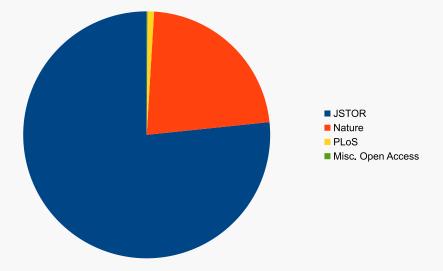
1,625,790 documents 334 journals



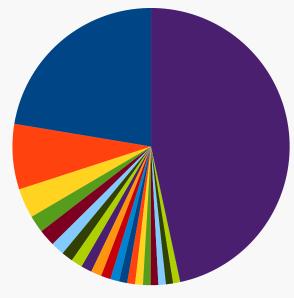




Data Sources



Journals



Nature Nature Nature Nature Current Science The Quarterly Review of Biology Plant Physiology The Auk The Science Teacher American Scientist Ecology

- The American Naturalist
- The Journal of Parasitology
- The American Biology Teacher
- The New Phytologist
- American Journal of Botany
- BioScience
- Copeia
- Science and Children
- Annals of Botany
- Marine Ecology Progress Series
- Botanical Gazette
- Other

Example: Biodiversity

Conceptual Analysis

Classic conceptual analysis question: What do scientists mean by **biodiversity**?



Articles from *Conservation Biology*, from 1987–2012.

5,459 articles; 27M tokens; 427k types.

Craig-Zeta Algorithm

Take a corpus that you believe to consist of two "sub-corpora," A and B.

Goal: Find 'marker words' that distinguish A papers as opposed to B papers, and vice versa.

Craig-Zeta Algorithm

Idea: Analyze as though papers that mention biodiversity are a different sub-corpus from those that do not.

CZ Marker Words

Markers for biodiversity:

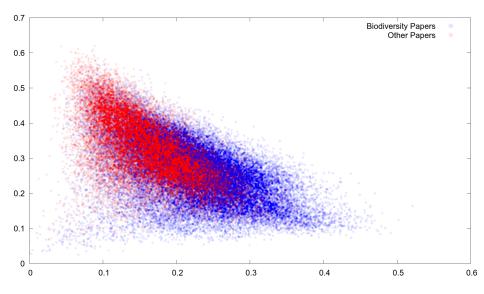
- diversity
- richness
- ecological
- protected
- planning
- conservation
- development
- policy
- economic
- assessment
- international
- management
- (Year numbers, 1997–2007)

Markers for not-biodiversity:

- population
- genetic
- breeding
- individuals
- survival
- rate
- variation
- reproductive
- mortality
- adult
- mean
- demographic
- (Year numbers, 1980–1988)

Craig-Zeta Algorithm

How do we know that the algorithm was successful?



Cooccurrence

What words occur (significantly often) within a given distance of 'biodiversity' (our focal word of interest)?

Cooccurrences (500-word window)

- biointegrity
- bioresources
- macroclimatic
- distributive
- *bureaucratically*
- countdown
- neoliberalization
- postmodernism
- manifesto
- cataloguing
- underprotected
- biopiracy
- hotspots
- coextinctions

A first suggestion...

Biodiversity is unusually related with words indicating **social and political context**.

A first suggestion...

Biodiversity is unusually related with words indicating **social and political context**.

Good! This lines up both with what historians of science and the practitioners themselves have told us about the self-image of the biodiversity paradigm.

Comparing Definitions

Could we try to look for places where **different** definitions of biodiversity have been used in different places in the literature?

Comparing Definitions

Could we try to look for places where **different** definitions of biodiversity have been used in different places in the literature?

One idea: Biodiversity is an inherently **spatial** concept. What parts of the globe has it been applied to in different contexts?

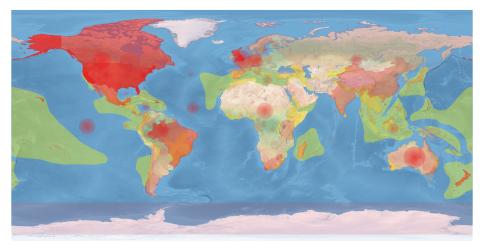
Named Entity Recognition

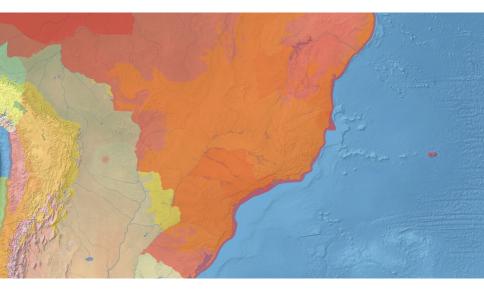
Look for any instances of proper place names throughout the *Conservation Biology* corpus, map them.

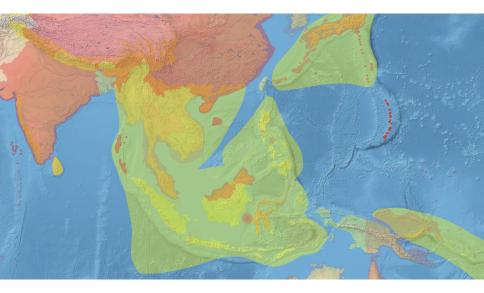
Named Entity Recognition

Look for any instances of proper place names throughout the *Conservation Biology* corpus, map them.

Unsolved problem for evoText: Geocoding! How to get from "Australia" to a set of geographic coordinates on a map? Currently using 3rd-party API (with a large free service tier).







A second suggestion...

Various definitions of biodiversity seem to **come apart** when we compare scientific and political uses of the concept.

Back to evoText

Other tools available:

- General word frequency analysis (1-gram or N-gram, with tf-idf scores, configurable block creation behavior, etc.)
- Collocation analysis (like cooccurrence but for direct bigrams)
- Some simple graphing-by-date
- Export to various reference manager citation formats

Questions?

charles@charlespence.net https://pencelab.be @pencechp · @pencelab





Wallonie - Bruxelles International.be



Craig-Zeta Algorithm

- 1. Divide corpus A and corpus B into blocks of 500 words.
- **2**. For each word, determine how many blocks in which that word appears. (Discard any words that appear in *every* block.)
- **3**. For each word, compute the Zeta score:

$$Z_w = \frac{\text{count}_A}{n_A} + \frac{n_B - \text{count}_B}{n_B} \tag{1}$$

That is: the fraction of blocks of *A* that contain *w* plus the fraction of blocks of *B* that *do not* contain *w*. Z_w values range from 2 (appears in every block of *A* and no blocks of *B*) to 0 (appears in no blocks of *A* and every block of *B*).

Collocation Algorithm

- **1**. Inputs: word of interest *t*, window *s*.
- **2**. Divide the corpus into *n* blocks of size *s*.
- **3**. For every word *w* in the corpus, compute both f_w (how many blocks contain *w*) and f_{wt} (how many blocks contain both *w* and *t*).
- **4**. Compute a score for the significance of the *wt* pair. In this case, we used mutual information:

$$I_{wt} = \log\left(\frac{f_{wt} \cdot n}{f_t \cdot f_w}\right) \tag{2}$$