Digital Preservation for Machine-Scale Access and Analysis

Lisa Green

Digital Preservation 2013
24 July 2013

#digpres13

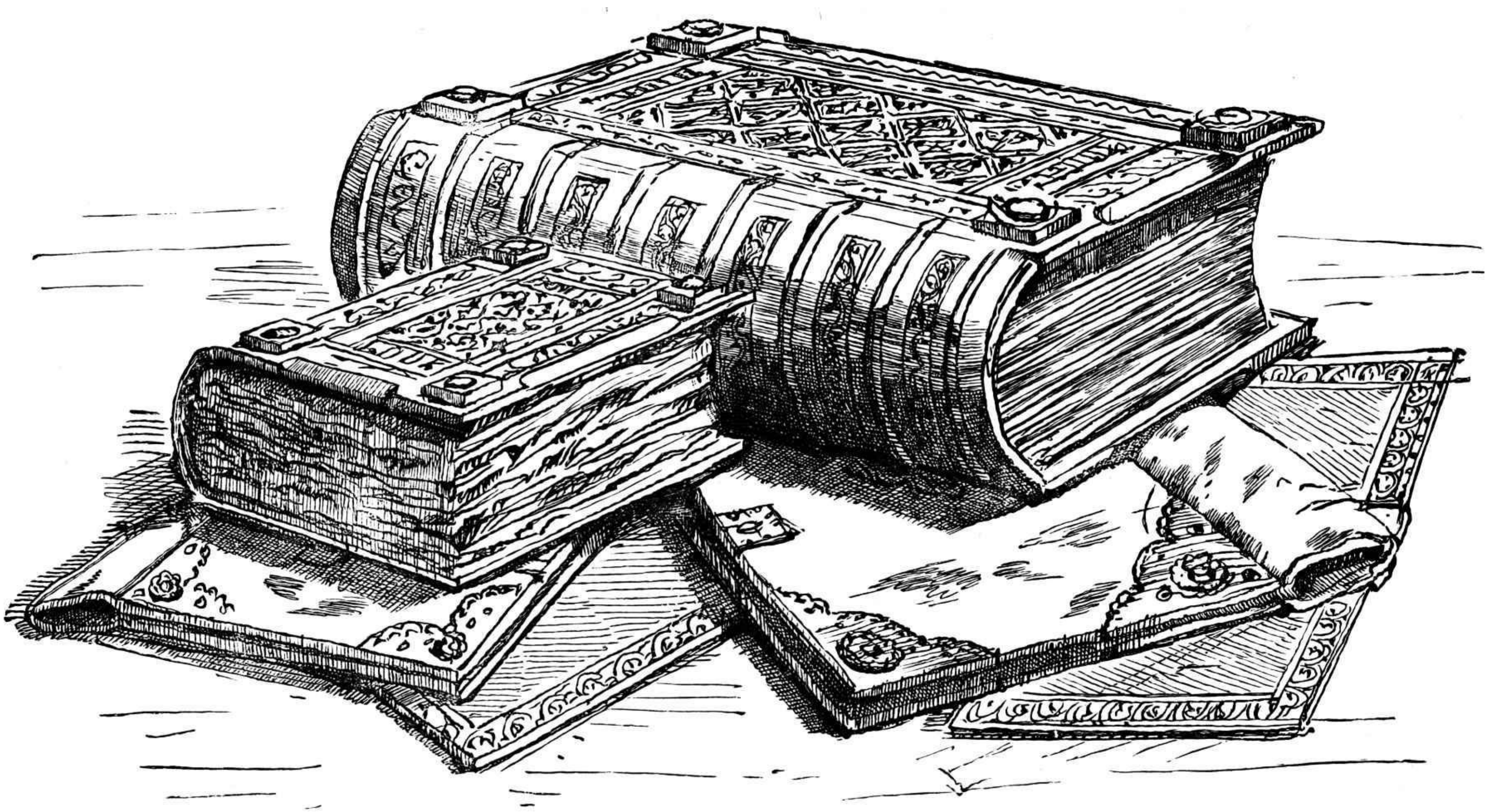


Image license: Public Domain Origin: http://www.fromoldbooks.org/Andrews-HistoricBywaysAndHighways/pages/33-domesday-book/

Bundesarchiv, B 145 Bild-F031434-0006
Foto: Gathmann, Jens | 23. März 1970

Origin: http://en.wikipedia.org/wiki/File:Bundesarchiv_B_145_Bild-F031434-0006,_Aachen,_Technische_Hochschule,_Rechenzentrum.jpg#file

Photo license: Public Domain Origin: http://en.wikipedia.org/wiki/File:Floppy_disk_2009_G1.jpg

Library of Congress Mission

To make its resources available and useful to Congress and the American people and to sustain and preserve a universal collection of knowledge and creativity for future generations.

Photo license: CC-BY Origin: http://www.flickr.com/photos/anonymouscollective/2291896028/

Library of Congress Mission

To make its resources available and useful to Congress and the American people and to sustain and preserve a universal collection of knowledge and creativity for future generations.

Photo license: CC-BY-SA Origin: http://www.flickr.com/photos/kellymccarthy/2345153323

Photo license: CC-BY-SA Origin: http://en.wikipedia.org/wiki/File:Wikimedia_Foundation_Servers-8055_08.jpg

Natural Language Processing

Sentiment Analysis

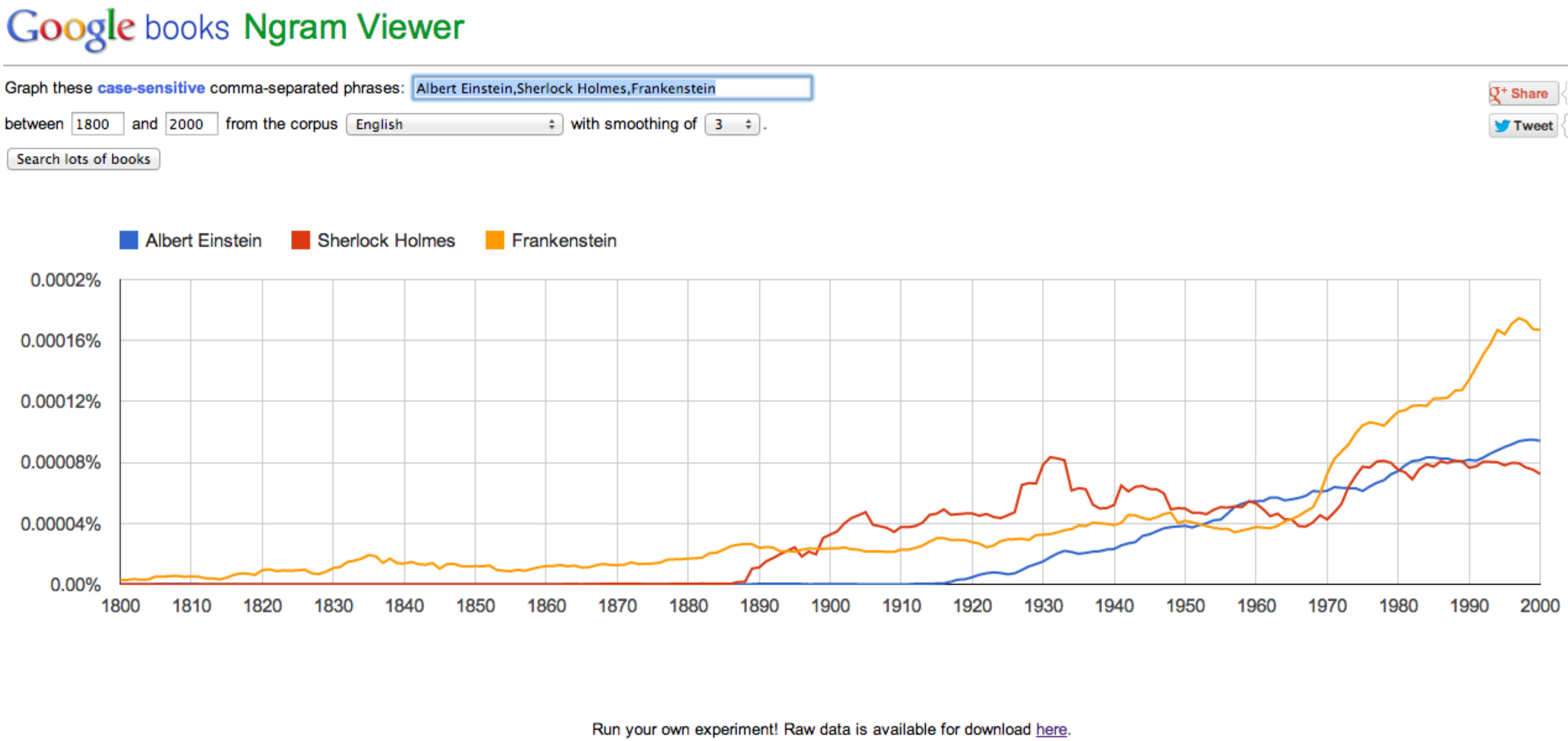
Cluster Analysis

Algorithms

Machine Learning

Artificial Neural Networks

Statistics

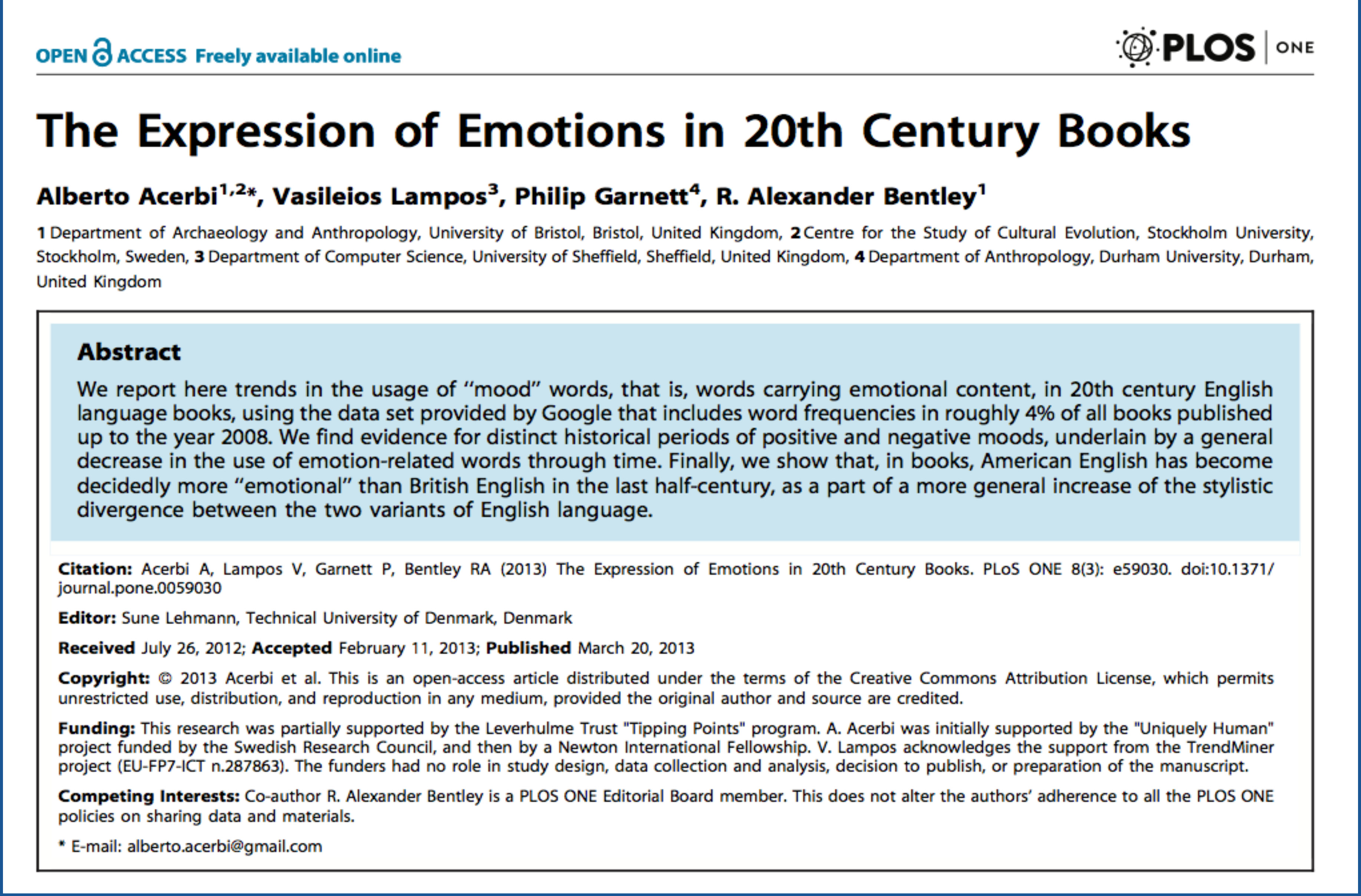


Google books Ngram Viewer

Graph these case-sensitive comma-separated phrases: Albert Einstein,Sherlock Holmes,Frankenstein

between 1800 and 2000 from the corpus English with smoothing of 3 .

Run your own experiment! Raw data is available for download here.



OPEN ACCESS Freely available online

PLOS | ONE

The Expression of Emotions in 20th Century Books

Alberto Acerbi[1,2]*, Vasileios Lampos[3], Philip Garnett[4], R. Alexander Bentley[1]

1 Department of Archaeology and Anthropology, University of Bristol, Bristol, United Kingdom, **2** Centre for the Study of Cultural Evolution, Stockholm University, Stockholm, Sweden, **3** Department of Computer Science, University of Sheffield, Sheffield, United Kingdom, **4** Department of Anthropology, Durham University, Durham, United Kingdom

Abstract

We report here trends in the usage of "mood" words, that is, words carrying emotional content, in 20th century English language books, using the data set provided by Google that includes word frequencies in roughly 4% of all books published up to the year 2008. We find evidence for distinct historical periods of positive and negative moods, underlain by a general decrease in the use of emotion-related words through time. Finally, we show that, in books, American English has become decidedly more "emotional" than British English in the last half-century, as a part of a more general increase of the stylistic divergence between the two variants of English language.

Citation: Acerbi A, Lampos V, Garnett P, Bentley RA (2013) The Expression of Emotions in 20th Century Books. PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030

Editor: Sune Lehmann, Technical University of Denmark, Denmark

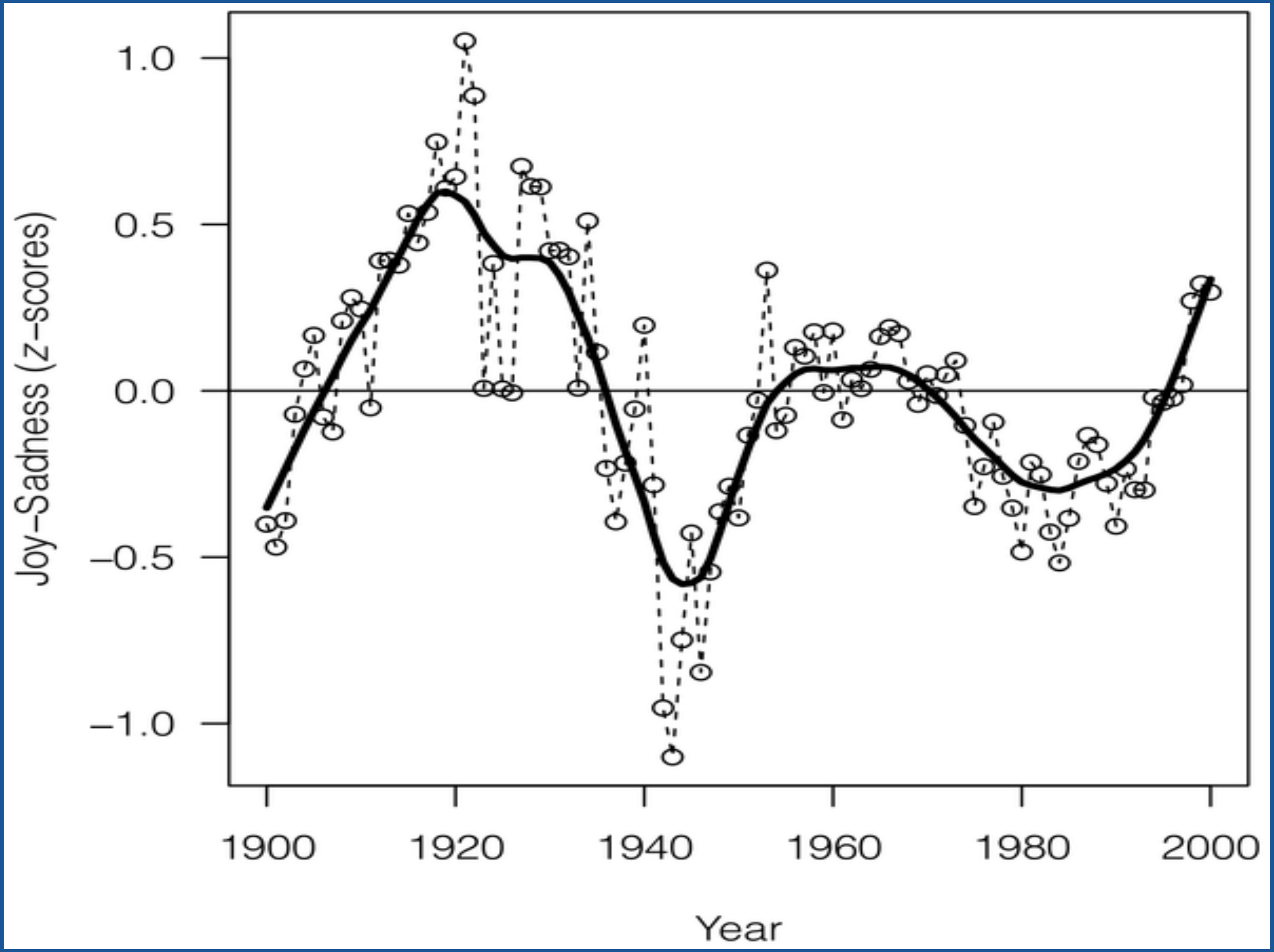
Received July 26, 2012; **Accepted** February 11, 2013; **Published** March 20, 2013

Funding: This research was partially supported by the Leverhulme Trust "Tipping Points" program. A. Acerbi was initially supported by the "Uniquely Human" project funded by the Swedish Research Council, and then by a Newton International Fellowship. V. Lampos acknowledges the support from the TrendMiner project (EU-FP7-ICT n.287863). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Co-author R. Alexander Bentley is a PLOS ONE Editorial Board member. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: alberto.acerbi@gmail.com

Figure 1. Historical periods of positive and negative moods.

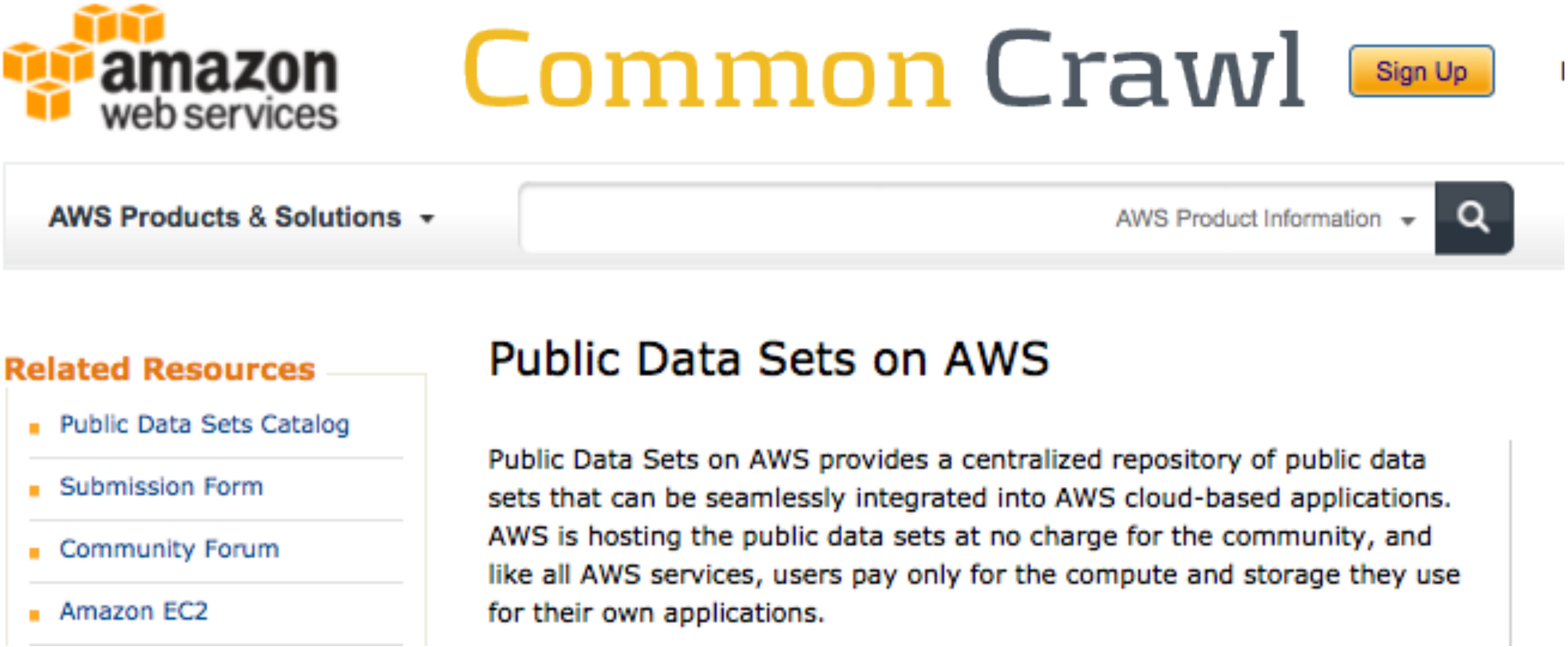


Acerbi A, Lampos V, Garnett P, Bentley RA (2013) The Expression of Emotions in 20th Century Books. PLoS ONE 8(3): e59030.
doi:10.1371/journal.pone.0059030
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0059030

#digpres13

Photo license: CC-BY-SA Origin: http://www.flickr.com/photos/kellymccarthy/2345153323

Common Crawl



Public Data Sets on AWS

Public Data Sets on AWS provides a centralized repository of public data sets that can be seamlessly integrated into AWS cloud-based applications. AWS is hosting the public data sets at no charge for the community, and like all AWS services, users pay only for the compute and storage they use for their own applications.

Common Crawl

Photo license: CC-BY Origin: http://www.flickr.com/photos/53370644@N06/4976497160/

Library of Congress Mission

To make its resources available and useful to Congress and the American people and to sustain and preserve a universal collection of knowledge and creativity for future generations.

We must enable machine-scale access and analysis.

NDIIPP

#digpres13

Thank you!

www.commoncrawl.org

lisa@commoncrawl.org

@commoncrawl

@boudicca