Cambridge Elements

Elements in Corpus Linguistics
edited by
Susan Hunston
University of Birmingham

NATURAL LANGUAGE PROCESSING FOR CORPUS LINGUISTICS

Jonathan Dunn
University of Canterbury

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781009074438
DOI: 10.1017/9781009070447

First published 2022

A catalogue record for this publication is available from the British Library.

ISBN 978-1-009-07443-8 Paperback
ISSN 2632-8097 (online)
ISSN 2632-8089 (print)

Additional resources for this publication at www.cambridge.org/dunnresources

Natural Language Processing for Corpus Linguistics

Elements in Corpus Linguistics

DOI: 10.1017/9781009070447
First published online: March 2022

Jonathan Dunn
University of Canterbury
Author for correspondence: Jonathan Dunn,
jonathan.dunn@canterbury.ac.nz

Abstract: Corpus analysis can be expanded and scaled up by incorporating computational methods from natural language processing. This Element shows how text classification and text similarity models can extend our ability to undertake corpus linguistics across very large corpora. These computational methods are becoming increasingly important as corpora grow too large for more traditional types of linguistic analysis. We draw on five case studies to show how and why to use computational methods, ranging from usage-based grammar to authorship analysis to using social media for corpus-based sociolinguistics. Each section is accompanied by an interactive code notebook that shows how to implement the analysis in Python. A stand-alone Python package is also available to help readers use these methods with their own data. Because large-scale analysis introduces new ethical problems, this Element pairs each new methodology with a discussion of potential ethical implications.

This Element also has a video abstract: www.cambridge.org/dunnabstract

Keywords: computational linguistics, natural language processing, corpus linguistics, text classification, text similarity, usage-based grammar, corpus-based sociolinguistics, computational stylistics, computational syntax

JEL classifications: A12, B34, C56, D78, E90

ISBNs: 9781009074438 (PB), 9781009070447 (OC)
ISSNs: 2632-8097 (online), 2632-8089 (print)

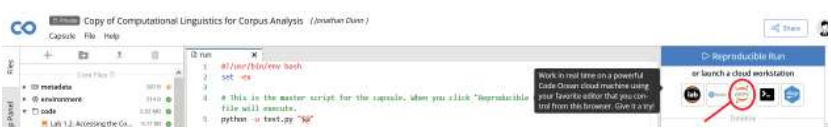
Contents

Accessing the Code Notebooks

https://doi.org/10.24433/CO.3402613.v1
https://github.com/jonathandunn/text_analytics
https://github.com/jonathandunn/corpus_analysis

To run the notebooks through Code Ocean, you will need to click the command that says "Edit Your Copy" in the top right-hand corner, as shown in the first screenshot:

The "Jupyter" command will now be available under the heading "Reproducible Run" as shown in the second screenshot:



This will start up the interactive notebook container. You can now find the notebooks within the "code" folder.

The following is a list of interactive notebooks together with the section of the Element which they accompany:

Lab 1.2. Accessing the Corpora

Lab 1.3. Visualizing Categories

Lab 1.4. Using Groupby to Explore Categories

Lab 1.5. Vectorizing Texts

Lab 2.1. Getting *x* and *y* Arrays for Dialects

Lab 2.2. Classifying Cities with TF-IDF and PMI

Lab 2.3. Classifying Authors with Function Word N-Grams

Lab 2.4. Using Positional Vectors for Parts of Speech