Cambridge University Press 978-1-009-07443-8 — Natural Language Processing for Corpus Linguistics Jonathan Dunn Frontmatter <u>More Information</u>

Cambridge Elements^{Ξ}

Elements in Corpus Linguistics edited by Susan Hunston University of Birmingham

NATURAL LANGUAGE PROCESSING FOR CORPUS LINGUISTICS

Jonathan Dunn University of Canterbury



© in this web service 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

© Jonathan Dunn 2022

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

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

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

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

© Jonathan Dunn 2022 ISBNs: 9781009074438 (PB), 9781009070447 (OC) ISSNs: 2632-8097 (online), 2632-8089 (print) Cambridge University Press 978-1-009-07443-8 — Natural Language Processing for Corpus Linguistics Jonathan Dunn Frontmatter <u>More Information</u>

Contents

1	Computational Linguistic Analysis	1
2	Text Classification	13
3	Text Similarity	39
4	Validation and Visualization	62
5	Conclusions	79
	References	81

Cambridge University Press 978-1-009-07443-8 — Natural Language Processing for Corpus Linguistics Jonathan Dunn Frontmatter <u>More Information</u>

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:

+ 13 3 0	≅ Metadata 🛛 🗙			
(itelies a la l			iii -	or launch a cloud workstation
→ Erretadata ==== 8	Metadata			0 2 0
• 9 environment				
* D code 201100 0	Social Memory	Natural Language Processing for Corpus Linguistics		lingue
Hab 1.3 Young Con. (1777) 0	Encol Street	I Natural Language Processing for Corpus Linguistics	100	C donathas Date rat 00 000000 - 2
Gab 1.4. Using Groupby L		These labs accompany the book Natural Language Processing for Corpus Linguistics. The labs show		fine (1, 303)
List 15 Vectorizing Text. Income @	6.0	how to do computational impactic analysis for usage-based grammar, comput-based social regulation.		* 🗇 Run 4573685
Hab 21, Dettings and y_ 94.5114 @		styletics, and computational social science.		C) onderse intra
Hab 2.2. Classifying Chies 10.0111 @		(overstational legiting) (comparational initial science) (corporational and science)		COLUMN HINT
Hab 2.3 Classifying Authors 11110 @				Published Version 1.1
🗏 Lab 2:4. Using Posttorial 15.1719 💩		(Rate & Lago grap processing)		COLORY COMPANY
H Lab 25: Classifying Hotels				

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

C		Cop Isula Fila		iputational L	inguisti	ts for Corpu	s Analysis (Jonathan Dunn)		the states of th
Panel Ries	+ • 10 metad • (0 araino • () code		1	10 1010 0 1010 0 1010 0	12 nm 1 2 3 4 1	alling/bin set es a this is file sells nythen -u	the marter script for the capsale. When you click "Reproducible	Work in real time on a powerful Cade Grown Could muchine using your Favorite addor that you con- unit from this between. Give it a tryi	C: Reproducible Run er launch a doud workstation

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

CAMBRIDGE

Cambridge University Press 978-1-009-07443-8 — Natural Language Processing for Corpus Linguistics Jonathan Dunn Frontmatter <u>More Information</u>

Lab 2.5. Classifying Hotels by Quality Using Sentiment Analysis

Lab 2.7. Classifying Cities Using MLPs

Lab 3.2. Register and Corpus Similarity

Lab 3.3. Finding Similar Documents

Lab 3.4. Finding Associated Words

Lab 3.5. Working with Word Embeddings

Lab 3.6. Clustering Word Embeddings

Lab 4.1. Baselines for Classifying Political Speeches

Lab 4.2. Ensuring Validity Using Cross-Validation

Lab 4.3. Unmasking Authorship

Lab 4.4. Comparing Word Embeddings

Lab 4.5. Making Maps for Linguistic Diversity