Elements in Corpus Linguistics
edited by
Susan Hunston
University of Birmingham

NATURAL LANGUAGE PROCESSING FOR CORPUS LINGUISTICS

Jonathan Dunn
University of Canterbury
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.

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)
## Contents

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

References ........................................................................... 81
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
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