Elements in Quantitative and Computational Methods for the Social Sciences
edited by
R. Michael Alvarez
California Institute of Technology
Nathaniel Beck
New York University

TEXT ANALYSIS IN PYTHON FOR SOCIAL SCIENTISTS

Discovery and Exploration

Dirk Hovy
Bocconi University
Text Analysis in Python for Social Scientists

Discovery and Exploration

Elements in Quantitative and Computational Methods for the Social Sciences

DOI: 10.1017/9781108873352
First published online: December 2020

Dirk Hovy
Bocconi University

Author for correspondence: Dirk Hovy, dirk.hovy@unibocconi.it

Abstract: Text is everywhere, and it is a fantastic resource for social scientists. However, because it is so abundant, and because language is so variable, it is often difficult to extract the information we want. A whole subfield of AI is concerned with text analysis (natural language processing). Many of the basic analysis methods developed are now readily available as Python implementations. This Element will teach you when to use which method, the mathematical background of how it works, and the Python code to implement it.

Keywords: text analysis; natural language processing; computational linguistics; exploration

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

© Dirk Hovy 2020

ISBNs: 9781108819824 (PB), 9781108873352 (OC)
ISSNs: 2398-4023 (online), 2514-3794 (print)
Contents

Introduction 1
Background 3
1 Prerequisites 3
2 What’s in a Word 5
3 Regular Expressions 19
4 Pointwise Mutual Information 24
5 Representing Text 27
Exploration: Finding Structure in the Data 50
6 Matrix Factorization 51
7 Clustering 59
8 Language Models 65
9 Topic Models 72
Appendix A: English Stopwords 84
Appendix B: Probabilities 85
References 90