

Using Python for Text Analysis in Accounting Research

Other titles in Foundations and Trends® in Accounting

Costing Systems

Eva Labro

ISBN:978-1-68083-568-7

Accounting Theory as a Bayesian Discipline

David Johnstone

ISBN:978-1-68083-530-4

Authority and Accountability in Hierarchies

Christian Hofmann and Raffi J. Indjejikian

ISBN: 978-1-68083-510-6

Dynamic Investment Models in Accounting Research

Alexander Nezlobin

ISBN: 978-1-68083-496-3

Financial Statement Analysis and Earnings Forecasting

Steven J. Monahan

ISBN: 978-1-68083-450-5

Executive Compensation, Corporate Governance, and Say on Pay

Fabrizio Ferri and Robert F. Gox

ISBN: 978-1-68083-420-8

Using Python for Text Analysis in Accounting Research

Vic Anand

University of Illinois at Urbana-Champaign
USA
vanand@illinois.edu

Khrystyna Bochkay

University of Miami
USA
kbochkay@bus.miami.edu

Roman Chychyla

University of Miami
USA
rchychyla@bus.miami.edu

Andrew Leone

Northwestern University
USA
andrew.leone@kellogg.northwestern.edu

now

the essence of knowledge

Boston — Delft

Foundations and Trends[®] in Accounting

Published, sold and distributed by:

now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:

now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

V. Anand, K. Bochkay, R. Chychyla and A. Leone. *Using Python for Text Analysis in Accounting Research*. Foundations and Trends[®] in Accounting, vol. 14, no. 3–4, pp. 128–359, 2020.

ISBN: 978-1-68083-761-2

© 2020 V. Anand, K. Bochkay, R. Chychyla and A. Leone

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, mechanical, photocopying, recording or otherwise, without prior written permission of the publishers.

Photocopying. In the USA: This journal is registered at the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923. Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by now Publishers Inc for users registered with the Copyright Clearance Center (CCC). The 'services' for users can be found on the internet at: www.copyright.com

For those organizations that have been granted a photocopy license, a separate system of payment has been arranged. Authorization does not extend to other kinds of copying, such as that for general distribution, for advertising or promotional purposes, for creating new collective works, or for resale. In the rest of the world: Permission to photocopy must be obtained from the copyright owner. Please apply to now Publishers Inc., PO Box 1024, Hanover, MA 02339, USA; Tel. +1 781 871 0245; www.nowpublishers.com; sales@nowpublishers.com

now Publishers Inc. has an exclusive license to publish this material worldwide. Permission to use this content must be obtained from the copyright license holder. Please apply to now Publishers, PO Box 179, 2600 AD Delft, The Netherlands, www.nowpublishers.com; e-mail: sales@nowpublishers.com

Foundations and Trends[®] in Accounting
Volume 14, Issue 3–4, 2020
Editorial Board

Executive Editors

Robert Bushman
The University of North Carolina at Chapel Hill

Sunil Dutta
University of California at Berkeley

Stephen Penman
Columbia University

Stefan J. Reichelstein, Managing editor
Stanford University

Editorial Scope

Topics

Foundations and Trends® in Accounting publishes survey and tutorial articles in the following topics:

- Auditing
- Corporate Governance
- Cost Management
- Disclosure
- Event Studies/Market Efficiency Studies
- Executive Compensation
- Financial Reporting
- Management Control
- Performance Measurement
- Taxation

Information for Librarians

Foundations and Trends® in Accounting, 2020, Volume 14, 4 issues. ISSN paper version 1554-0642. ISSN online version 1554-0650. Also available as a combined paper and online subscription.

Contents

1	Introduction	3
2	Installing Python on Your Computer	6
2.1	The Role of Packages in Python	6
2.2	The Anaconda Distribution of Python	7
2.3	Installing Anaconda	8
2.4	Launching and Using Anaconda	10
3	Jupyter Notebooks	12
3.1	Motivating Example	12
3.2	JupyterLab: A Development Environment for Jupyter Notebooks	14
3.3	How to Launch JupyterLab	17
3.4	Working in JupyterLab	18
3.5	The Markdown Language and Formatted Text Cells	24
4	A Brief Introduction to the Python Programming Language	28
4.1	Fundamentals	28
4.2	Variables and Data Types	30
4.3	Operators	38
4.4	The print Function	43

4.5	Control Flow	45
4.6	Functions	49
4.7	Collections—Lists, Tuples, and Dictionaries	58
4.8	Working with Strings	67
5	Working with Tabular Data: The Pandas Package	74
5.1	The Main Objects in Pandas	75
5.2	Required <code>import</code> Statements	76
5.3	Importing and Exporting Data	77
5.4	Viewing Data in Pandas	85
5.5	Selecting and Filtering Data	86
5.6	Creating New Columns	93
5.7	Dropping and Renaming Columns	97
5.8	Sorting Data	99
5.9	Merging Data	100
6	Introduction to Regular Expressions	102
6.1	Looking for Patterns in Text	102
6.2	Characters and Character Sets	105
6.3	Anchors and Boundaries in Regex	107
6.4	Quantifiers in Regex	108
6.5	Groups in Regex	109
6.6	Lookahead and Lookbehind in Regex	111
6.7	Examples of Regex for Different Textual Analysis Tasks	112
7	Dictionary-Based Textual Analysis	116
7.1	Advantages of Dictionary-Based Textual Analysis	116
7.2	Understanding Dictionaries	118
7.3	Identifying Words and Sentences in Text	120
7.4	Stemming and Lemmatization	124
7.5	Word Weighting	128
7.6	Dictionary-Based Word-Count Functions	129
8	Quantifying Text Complexity	138
8.1	Understanding Text Complexity	138
8.2	Calculating Text Length	139

8.3	Measuring Text Readability Using the Fog Index	141
8.4	Measuring Text Readability Using BOG Index	146
9	Sentence Structure and Classification	148
9.1	Identifying Forward-Looking Sentences	148
9.2	Dictionary Approach to Sentence Classification	153
9.3	Identifying Sentence Subjects and Objects	156
9.4	Identifying Named Entities	160
9.5	Using Stanford NLP for Part-of-Speech and Named Entity Recognition Tasks	163
10	Measuring Text Similarity	167
10.1	Comparing Text Using Similarity Measures	167
10.2	Text Similarity Measure for Long Text: Cosine Similarity	168
10.3	Text Similarity Measure for Short Text: Levenshtein Distance	176
10.4	Measuring Semantic Similarity Using Word2Vec Embedding Model	179
11	Identifying Specific Information in Text	185
11.1	Text Identification and Extraction Problem	185
11.2	Example: Extracting Management Discussion and Analysis Section from a Plain-Text 10-K Filing	187
11.3	Example: Extracting Management Discussion and Analysis Section from an HTML 10-K Filing	193
11.4	Extracting Text from XBRL Financial Reports	201
12	Collecting Data from the Internet	206
12.1	Accessing Data on the Web	206
12.2	EDGAR Data	206
12.3	Web Scraping	220
12.4	A Note on API's	225
	Acknowledgements	227
	References	228

Using Python for Text Analysis in Accounting Research

Vic Anand¹, Khrystyna Bochkay², Roman Chychyla³ and Andrew Leone⁴

¹*University of Illinois at Urbana-Champaign, USA; vanand@illinois.edu*

²*University of Miami, USA; kbochkay@bus.miami.edu*

³*University of Miami, USA; rchychyla@bus.miami.edu*

⁴*Northwestern University, USA; andrew.leone@kellogg.northwestern.edu*

ABSTRACT

The prominence of textual data in accounting research has increased dramatically. To assist researchers in understanding and using textual data, this monograph defines and describes common measures of textual data and then demonstrates the collection and processing of textual data using the Python programming language. The monograph is replete with sample code that replicates textual analysis tasks from recent research papers.

In the first part of the monograph, we provide guidance on getting started in Python. We first describe Anaconda, a distribution of Python that provides the requisite libraries for textual analysis, and its installation. We then introduce the Jupyter notebook, a programming environment that improves research workflows and promotes replicable research. Next, we teach the basics of Python programming and demonstrate the basics of working with tabular data in the Pandas package.

Vic Anand, Khrystyna Bochkay, Roman Chychyla and Andrew Leone (2020), "Using Python for Text Analysis in Accounting Research", *Foundations and Trends® in Accounting*: Vol. 14, No. 3–4, pp 128–359. DOI: 10.1561/1400000062. Supplementary material available from: http://dx.doi.org/10.1561/1400000062_supp.

The second part of the monograph focuses on specific textual analysis methods and techniques commonly used in accounting research. We first introduce regular expressions, a sophisticated language for finding patterns in text. We then show how to use regular expressions to extract specific parts from text. Next, we introduce the idea of transforming text data (unstructured data) into numerical measures representing variables of interest (structured data). Specifically, we introduce dictionary-based methods of (1) measuring document sentiment, (2) computing text complexity, (3) identifying forward-looking sentences and risk disclosures, (4) collecting informative numbers in text, and (5) computing the similarity of different pieces of text. For each of these tasks, we cite relevant papers and provide code snippets to implement the relevant metrics from these papers.

Finally, the third part of the monograph focuses on automating the collection of textual data. We introduce web scraping and provide code for downloading filings from EDGAR.

1

Introduction

Analyzing the textual content of corporate disclosures, contracts, analyst reports, news articles, and social media posts has gained an increased popularity among accounting and finance researchers and the investment community in general. Unlike numbers, which are often the outcome of formal accounting rules, trading activities, deal negotiations, etc., texts bring with them an infinite number of possibilities. Even when thinking about a single concept or thought, the number of ways in which that thought might be expressed is seemingly boundless, and this is no less true in the domain of corporate communications than in interpersonal communications.

In this monograph, we provide an interactive step-by-step framework for analyzing spoken or written language for faculty and PhD students in social sciences. Our goal is to demonstrate how textual analysis can enhance research by automatically extracting new and previously unknown information from voluminous disclosures, news articles, and social media posts. We present all materials in a way that allows the reader to learn about a textual analysis concept or technique and also replicate it by doing. Specifically, for each concept/technique, we cite relevant papers and provide reader-friendly code snippets, allowing

readers to execute our code on their own machines. We do not provide a comprehensive review of the textual analysis literature and refer our readers to Li (2010a), Loughran and McDonald (2016), and Henry and Leone (2016) that provide excellent surveys of the literature on the topic.

We begin by showing how to install and use Python. Python is a general purpose programming language that has been consistently ranked in the top ten most popular programming languages in the world. It is very efficient and intuitive in the areas of pattern matching and text analysis. We review Python's basic programming syntax, operators, data types, functions, etc., allowing the readers to familiarize themselves with the programming environment first. We also discuss the Jupyter notebook which is an open-source web application that allows creating, running, and testing your Python code interactively. We introduce the Pandas package for working with tabular data; this will aid researchers as they convert unstructured textual data into structured, tabular data.

Next, we introduce regular expressions which represent patterns for matching different elements in texts (e.g., individual words, variants of words, numbers, symbols, etc.). Regular expressions are the foundation of being able to calculate different textual analysis metrics. We then proceed with the discussion and coding of different textual analysis methods used in accounting and finance studies. These methods include parsing texts into individual words and/or sentences, measuring tone/sentiment of a document, identifying specific words or phrases in text, measuring text complexity, classifying sentences into categories, identifying linguistic structure of a sentence, and measuring textual similarity. To facilitate the exposition of our code, we cite relevant research studies that demonstrate specific uses of textual metrics.

Finally, we provide an overview of web scraping and file processing features in Python. Specifically, we focus on downloading EDGAR filings and identifying specific sections in them.

Taken together, the first five sections of this monograph will help readers get started with Python and prepare for writing their own code. The remaining sections will help the reader to learn various textual analysis methods and implement the coding of the methods in Python.

We make all our code (in Jupyter Notebooks) available as supplementary material. We kindly ask researchers who use our materials to cite this monograph.

References

- Bentley, J. W., T. E. Christensen, K. H. Gee, and B. C. Whipple (2018). “Disentangling managers’ and analysts’ non-GAAP reporting”. *Journal of Accounting Research*. 56(4): 1039–1081.
- Bentley, J. W., K. Stubbs, Y. Tian, and R. L. Whited (2019). “Manipulating the narrative: Managerial discretion in the emphasis of GAAP metrics in earnings announcement press releases”. Available at SSRN: URL: <https://ssrn.com/abstract=349773>.
- Blankespoor, E. (2019). “The impact of information processing costs on firm disclosure choice: Evidence from the XBRL mandate”. *Journal of Accounting Research*. 57(4): 919–967.
- Bochkay, K., R. Chychyla, and D. Nanda (2019). “Dynamics of CEO disclosure style”. *The Accounting Review*. 94(4): 103–140.
- Bochkay, K., J. Hales, and S. Chava (2020). “Hyperbole or reality? Investor response to extreme language in earnings conference calls”. *The Accounting Review*. 95(2): 31–60.
- Bochkay, K. and C. B. Levine (2019). “Using MD&A to improve earnings forecasts”. *Journal of Accounting, Auditing & Finance*. 34(3): 458–482.
- Bonsall, S. B., A. J. Leone, B. P. Miller, and K. Rennekamp (2017). “A plain English measure of financial reporting readability”. *Journal of Accounting and Economics*. 63(2): 329–357.

- Bozanic, Z., D. T. Roulstone, and A. Van Buskirk (2018). “Management earnings forecasts and other forward-looking statements”. *Journal of Accounting and Economics*. 65(1): 1–20.
- Brochet, F., K. Kolev, and A. Lerman (2018). “Information transfer and conference calls”. *Review of Accounting Studies*. 23(3): 907–957.
- Brown, S. V. and J. W. Tucker (2011). “Large-sample evidence on firms’ year-over-year MD&A modifications”. *Journal of Accounting Research*. 49(2): 309–346.
- Butler, M., A. J. Leone, and M. Willenborg (2004). “An empirical analysis of auditor reporting and its association with abnormal accruals”. *Journal of Accounting and Economics*. 37(2): 139–165.
- Campbell, J. L., H. Chen, D. S. Dhaliwal, H.-M. Lu, and L. B. Steele (2014). “The information content of mandatory risk factor disclosures in corporate filings”. *Review of Accounting Studies*. 19(1): 396–455.
- Cecchini, M., H. Aytug, G. J. Koehler, and P. Pathak (2010). “Making words work: Using financial text as a predictor of financial events”. *Decision Support Systems*. 50(1): 164–175.
- Chychyla, R., A. J. Leone, and M. Minutti-Meza (2019). “Complexity of financial reporting standards and accounting expertise”. *Journal of Accounting and Economics*. 67(1): 226–253.
- Dyer, T., M. Lang, and L. Stice-Lawrence (2017). “The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation”. *Journal of Accounting and Economics*. 64(2): 221–245.
- Filzen, J. J. and K. Peterson (2015). “Financial statement complexity and meeting analysts’ expectations”. *Contemporary Accounting Research*. 32(4): 1560–1594.
- Gow, I. D., D. F. Larcker, and A. A. Zakolyukina (2019). “Non-answers during conference calls”. *Chicago Booth Research Paper (19-01)*.
- Guay, W., D. Samuels, and D. Taylor (2016). “Guiding through the Fog: Financial statement complexity and voluntary disclosure”. *Journal of Accounting and Economics*. 62(2): 234–269.
- Gunning, R. (1952). *Technique of Clear Writing*. McGraw-Hill.
- Hanley, K. W. and G. Hoberg (2010). “The information content of IPO prospectuses”. *The Review of Financial Studies*. 23(7): 2821–2864.

- Heitmann, M., C. Siebert, J. Hartmann, and C. Schamp (2020). “More than a feeling: Benchmarks for sentiment analysis accuracy”. *Working Paper*. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3489963.
- Henry, E. and A. J. Leone (2016). “Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone”. *The Accounting Review*. 91(1): 153–178.
- Hoberg, G. and G. Phillips (2016). “Text-based network industries and endogenous product differentiation”. *Journal of Political Economy*. 124(5): 1423–1465.
- Hoitash, R. and U. Hoitash (2018). “Measuring accounting reporting complexity with XBRL”. *The Accounting Review*. 93(1): 259–287.
- Hope, O.-K., D. Hu, and H. Lu (2016). “The benefits of specific risk-factor disclosures”. *Review of Accounting Studies*. 21(4): 1005–1045.
- Huang, X., S. H. Teoh, and Y. Zhang (2014). “Tone management”. *The Accounting Review*. 89(3): 1083–1113.
- Jegadeesh, N. and D. Wu (2013). “Word power: A new approach for content analysis”. *Journal of Financial Economics*. 110(3): 712–729.
- Kincaid, J. P., R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom (1975). “Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel”, Naval Technical Training Command Millington TN Research Branch.
- Kravet, T. and V. Muslu (2013). “Textual risk disclosures and investors’ risk perceptions”. *Review of Accounting Studies*. 18(4): 1088–1122.
- Lang, M. and L. Stice-Lawrence (2015). “Textual analysis and international financial reporting: Large sample evidence”. *Journal of Accounting and Economics*. 60(2–3): 110–135.
- Larcker, D. F. and A. A. Zakolyukina (2012). “Detecting deceptive discussions in conference calls”. *Journal of Accounting Research*. 50(2): 495–540.
- Lehavy, R., F. Li, and K. Merkley (2011). “The effect of annual report readability on analyst following and the properties of their earnings forecasts”. *The Accounting Review*. 86(3): 1087–1115.

- Li, F. (2008). “Annual report readability, current earnings, and earnings persistence”. *Journal of Accounting and Economics*. 45(2–3): 221–247.
- Li, F. (2010a). “Survey of the literature”. *Journal of Accounting Literature*. 29: 143–165.
- Li, F. (2010b). “The information content of forward-looking statements in corporate filings—A Naïve Bayesian machine learning approach”. *Journal of Accounting Research*. 48(5): 1049–1102.
- Li, F., M. Minnis, V. Nagar, and M. Rajan (2014). “Knowledge, compensation, and firm value: An empirical analysis of firm communication”. *Journal of Accounting and Economics*. 58(1): 96–116.
- Lo, K., F. Ramos, and R. Rogo (2017). “Earnings management and annual report readability”. *Journal of Accounting and Economics*. 63(1): 1–25.
- Loughran, T. and B. McDonald (2011). “When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks”. *The Journal of Finance*. 66(1): 35–65.
- Loughran, T. and B. McDonald (2013). “IPO first-day returns, offer price revisions, volatility, and form S-1 language”. *Journal of Financial Economics*. 109(2): 307–326.
- Loughran, T. and B. McDonald (2016). “Textual analysis in accounting and finance: A survey”. *Journal of Accounting Research*. 54(4): 1187–1230.
- Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013a). “Efficient estimation of word representations in vector space”. *ICLR Workshop*, arXiv preprint arXiv:1301.3781.
- Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013b). “Distributed representations of words and phrases and their compositionality”. In: *Advances in Neural Information Processing Systems*. 3111–3119.
- Muslu, V., S. Radhakrishnan, K. Subramanyam, and D. Lim (2015). “Forward-looking MD&A disclosures and the information environment”. *Management Science*. 61(5): 931–948.

- Price, S. M., J. S. Doran, D. R. Peterson, and B. A. Bliss (2012). “Earnings conference calls and stock returns: The incremental informativeness of textual tone”. *Journal of Banking & Finance*. 36(4): 992–1011.
- Project Jupyter (2018). “JupyterLab documentation”. <https://jupyterlab.readthedocs.io/en/stable/> (accessed: 22 Mar 2020).
- Project Jupyter (2020). “About us”. <https://jupyter.org/about> (accessed: 22 Mar 2020).
- Securities and Exchange Commission (1999). “A plain English handbook: How to create clear SEC disclosure”, <https://www.sec.gov/reportspubs/investor-publications/newsextrahandbookhtm.html>.
- Tetlock, P. C. (2007). “Giving content to investor sentiment: The role of media in the stock market”. *The Journal of Finance*. 62(3): 1139–1168.
- Tetlock, P. C., M. Saar-Tsechansky, and S. Macskassy (2008). “More than words: Quantifying language to measure firms’ fundamentals”. *The Journal of Finance*. 63(3): 1437–1467.
- You, H. and X.-J. J. Zhang (2009). “Financial reporting complexity and investor underreaction to 10-K information”. *Review of Accounting Studies*. 14(4): 559–586.