Using Python for Text Analysis in Accounting Research
Other titles in Foundations and Trends® in Accounting

Costing Systems
Eva Labro

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

Financial Statement Analysis and Earnings Forecasting
Steven J. Monahan

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
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
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

¹University of Illinois at Urbana-Champaign, USA; vanand@illinois.edu
²University of Miami, USA; kbochkay@bus.miami.edu
³University of Miami, USA; rchychyla@bus.miami.edu
⁴Northwestern 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.
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.
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
Introduction

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.


228
References


