Accounting Department

School of Economics and Management
Tilburg University
Warandelaan 2
5037 AB Tilburg
+31-013 466 8288
s.hollander@tilburguniversity.edu

Stephan Hollander
Professor

Ph.D. Seminar
Accounting Text as Data
July 1-3, 2026
University of Zurich

COURSE OVERVIEW AND OBJECTIVES

This course covers the fast-evolving field of computational analysis of accounting text, teaching students how to collect, process, and analyze textual content from regulatory filings (e.g., SEC Form 10-K), earnings conference call transcripts, and related documents. Drawing on natural language processing, it combines conceptual foundations with hands-on experience in building corpora, applying analytical tools, and interpreting results. By the end of the course, students will be equipped to use accounting text as data for research across a range of academic fields, including but not limited to accounting.

COURSE STRUCTURE

Textual analysis has become a core empirical method across accounting, economics, finance, law, management, and beyond. Its prominence lies in its ability to transform unstructured non-numerical data into quantitative measures of constructs that were previously difficult or impossible to observe directly, such as sentiment, discourse topics, and entrepreneurial style. This course introduces the main methodological approaches, emphasizing their theoretical foundations and the conditions under which they are most effective.

To support both the application of existing techniques and their adaptation to new constructs, the course is organized around a unifying three-part framework: an introduction to Python for textual analysis, followed by two sections devoted to the core workflow of quantification and mapping:

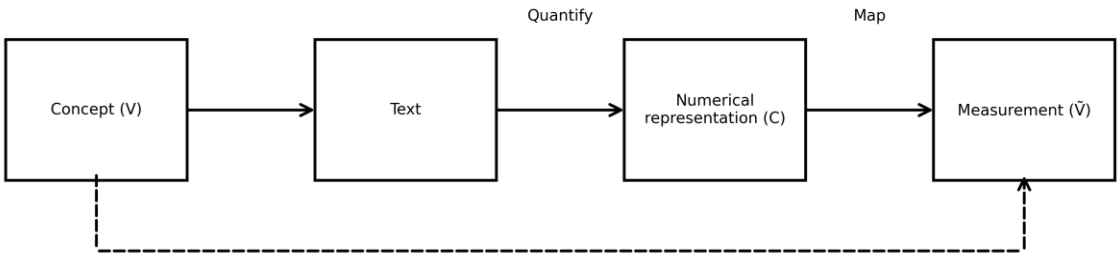

Figure 1: Text-as-data framework (Gentzkow, Kelly and Taddy 2019)

Within this framework, quantification involves transforming text into numerical, machine-readable representations, such as a bag-of-words vectors. Mapping encompasses techniques—ranging from word lists to supervised and unsupervised methods—that convert

these numerical representations into operational measures of the construct under study.

COURSE REQUIREMENTS

In preparation for the course, it would be helpful to briefly review the required readings. The main objective is to familiarize yourself with the lingua franca: to have encountered the core terminology and key ideas ahead of class. Most, if not all, of these will be explored in greater depth and contextualized during class discussions.

In addition, you are asked to skim the literature—spanning accounting, economics, finance, law, and management—and select a "text as data" (= turning unstructured text data into structured numerical data) paper that you find most interesting. If you already know a paper, you may skip this step.

> My top three, listed in alphabetical order (by first author's last name), text-as-data picks—subject to change:
>
> Baker, Scott, Nicholas Bloom, and Steven J. Davis, 2016. "Measuring economic policy uncertainty." *The Quarterly Journal of Economics* 131: 1593-1636
>
> Cohen, Lauren, Christopher Malloy, and Quoc Nguyen, 2020. "Lazy price." *Journal of Finance* 75: 1371-1415
>
> Hoberg, Gerard and Gordon Phillips, 2016. "Text-based network industries and endogenous product differentiation." *Journal of Political Economy* 124: 1423-1465
>
> What I like most about them is their simplicity (don't get me wrong, they're very thoughtfully executed)—both in idea and methodology.

Please prepare a brief two-sentence explanation of your choice, focusing on what you find most compelling about the paper. In the first session, you will be asked to share your selection in class. This helps me get a sense of your interests and tailor the course discussions accordingly.

No prior experience with Python is necessary. This course provides a step-by-step introduction to the fundamentals of programming in Python. We will use Google Colab to run Python code: an interactive coding environment built on the Jupyter Notebook framework. Colab is selected to ensure all students have equal access to computing power through Google's servers and to make setup as simple as possible for the course. A Google account is required to use Google Colab.

COURSE LOGISTICS

Time:	July 1 (beginning at 10:30 AM) to July 3, 2026 (ending at 3:00 PM)
Location:	TBA
Participants:	Max. 20
Cost:	No charge (participants will have to cover their travel expenses)

TARGET AUDIENCE

Ph.D. students

REQUIRED READINGS

Matthew Gentzkow, Bryan Kelly, and Matt Taddy (2019). "Text as Data." *Journal of Economic Literature* 57(3), 535-574. DOI: 10.1257/jel.20181020

Tarek A. Hassan, Stephan Hollander, Aakash Kalyani, Laurence van Lent, Markus Schwedeler, and Ahmed Tahoun (2025). "Text as Data in Economic Analysis." *Journal of Economic Perspectives* 39 (3), 193-220. DOI: 10.1257/jep.20231365

COURSE SCHEDULE

Session 0 (July 1, 10:30-12:00 AM): Overview and Foundation

Topic	Key Concepts & Activities
Lecture on "Accounting Text as Data"	Deep dive into the central theme: How unstructured text is transformed into quantitative measures of otherwise unobservable constructs. Discussing the econometric and statistical foundations and their role in enabling credible empirical inference from textual data. Introduction of the core three-part workflow that organizes the course

Session 1 (July 1, 1:00-3:30 PM): Introduction to Python for Textual Analysis

Topic	Key Concepts & Activities
Setting up the Environment	Setting up the computational environment: Python, Jupyter/Colab, essential libraries like pandas, nltk, scikit-learn
Python Fundamentals for Text Data	Data structures (lists, dictionaries, strings). Basic file I/O (reading and writing text files). Introduction to the pandas library for handling tabular data
Retrieving Accounting Text	Discussion on sources (SEC EDGAR, earnings call transcripts). Hands-on exercise: Retrieving and loading a small sample document

Session 2 (July 2, 9:30-12:00 AM): Quantification—Transforming Text into Data

Topic	Key Concepts & Activities
Corpus Construction and Pre-processing	Defining a corpus. Core pre-processing steps: tokenization, lowercasing, stop word removal, punctuation stripping, stemming vs. lemmatization
Numerical Representation: Bag-of-Words (BoW)	The concept of a document-term matrix (DTM). Creating BoW vectors (count-based). Introduction to scikit-learn's CountVectorizer
Advanced Quantification: Term Frequency-Inverse Document Frequency (TF-IDF)	Understanding and calculating TF-IDF vectors as a refinement over raw counts

Session 3 (July 2, 1:00-3:30 PM): Mapping I—Dictionary and Rule-Based Methods

Topic	Key Concepts & Activities
Conceptualizing Text-Based Constructs	How numerical vector representations are mapped onto substantive concepts such as sentiment, readability, and pairwise document similarity
Dictionary-Based Analysis	Introduction to established accounting/finance dictionaries (e.g., Loughran-McDonald). Applying dictionaries to BoW/TF-IDF vectors to measure sentiment. Calculating dictionary scores (e.g., % of negative words)
Rule-Based Methods	Applying rule-based methods to compute readability scores (e.g., Flesch-Kincaid) using libraries like textstat and measuring pairwise document cosine similarity

Session 4 (July 3, 9:30-12:00 AM): Mapping II— Supervised Learning for Classification

Topic	Key Concepts & Activities
Introduction to Supervised Methods	The requirement for labeled data and the core machine learning workflow: involving training, validation, and testing phases; including how models are developed, tuned, and evaluated for out-of-sample performance
Text Classification Models	Applying traditional machine learning models (e.g., Naive Bayes, Logistic Regression) to the DTM/TF-IDF matrix for classification tasks
Model Evaluation	Key metrics for text classification: Accuracy, Precision, Recall, F1-Score. Interpreting model coefficients to understand which words drive the classification

Session 5 (July 3, 1:00-3:30 PM): Unsupervised Learning and Next Steps

Topic	Key Concepts & Activities
Introduction to Unsupervised Methods	When ground-truth labels are unavailable, focus shifts to uncovering latent structure in the data. The goal of discovering hidden structures or topics
Word Embeddings	Introduction to word embeddings (e.g., Word2Vec, BERT) as a modern quantification technique that preserves semantic meaning. Discussion on future research directions
Synthesis and Project Planning	Review of the entire workflow. Leveraging course techniques for research across various fields