Syllabus

Automated Textual Analyses: Text as Data

Instructor

Harm H. Schütt
Professor of Financial Accounting
WHU – Otto Beisheim School of Management

Email

harm.schuett@whu.edu

Course dates

18.12.2024 – 20.12.2024

Course location

Room 202, Kaulbachstr. 45

Course hours

08:45 – 16:45

(4 blocks of 90min each per day)

Course Overview

The course is aimed at doctoral students and teaches current textual analysis methods used in Accounting, Management, and Finance research. It introduces a framework and a toolset that enables researchers to measure previously hard-to-measure latent concepts using text data.

The course is roughly divided into three parts of unequal length. The first part is an introduction to the Python programming language for the purposes of textual analysis. The other two parts divide textual analysis into two connected steps: quantification and mapping

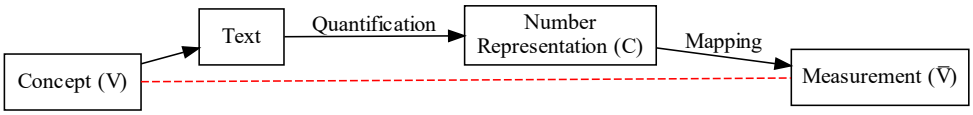

Figure 1: Text as data framework (Gentzkow, Kelly, Taddy, 2019)

Quantification concerns quantifying text into machine-readable form, such as the bag-of-words representation. Mapping encompasses methods, such as word lists and supervised or unsupervised methods, that turn numerical representations into the measure of interest.

Participants will be introduced to commonly applied approaches for both steps. They will learn to reason about which approaches are advisable given the text at hand and the concept to be measured. We will see multiple examples of how the concept to be measured influences certain texts and suggests particular quantification and mapping steps. Finally, we will discuss and analyze how generative AI can be used in this context.

Course Schedule

Day	Subject	Practice
Day 1	Measuring concepts with text data - Overview of methods - Overview and rationale of the course content - Illustration of famous use cases	
	A guiding framework for deciding which method to use - Introduction to Gentzkow, Kelly, Taddy (2019) - The importance of the signal-to-noise ratio	
	Short introduction to Python - Introduction to the Python language	Preparing and structuring financial documents
	Parsing text data - Introduction to regex and parsing using Python	Extracting which numbers refer to in conference calls
Day2	Turning text into numbers - Pre-processing text - Bag-of-words representation - Introduction to Spacy NLP tools - Introduction to the SKlearn ML library - Word-embeddings representation - Finding informative features - best-practices	
	Measuring concepts by word classification – Dictionary approaches - Mapping word counts to concepts - When does it work well? - Which words? – Designing word lists	Computing sentiment scores Exploring token occurrences
Day 3	Measuring concepts by document similarity – Cosine similarity - Similarity as a powerful analogy concept - Cosine similarity from bag-of-words	Sorting firms into business models
	Measuring concepts by document classification – Supervised approaches - Prediction approaches based on training data - Simple regression versus more flexible ML approaches	Classifying speech, Named-entity recognition

Day	Subject	Practice
	- Naïve Bayes, vs penalized regressions, vs. SVM - Multilabel and Multiclass problems	
	Measuring concepts by document classification – Unsupervised approaches - Clustering approaches: One topic per document - Model-based approaches and Latent Dirichlet Allocation: multiple topics per document - Introduction to the genism topic modeling library	Explore upcoming trends in scientific articles
	Measuring constructs using large language models. Examples and critique	

Before class

Do three things before class. First, browse the two papers cited in the references at the end of the syllabus. You do not have to internalize every statement made there. I just want you to have heard the terms and be familiar with some ideas before class. Second, install the software (see below). Third, are you a complete Python beginner? If so, see whether you can still take a quick Datacamp beginner course. It is not required but can help you a lot.

Teaching Mode

This course is divided into two-thirds lectures and one-third practicing the methods under my guidance. As a result, the number of seats is limited. During the sessions, you will be asked to apply the methods and coding patterns you learned in selected exercise sessions and answer certain questions. You must bring your laptop with the required software installed and ready to go to make this work. If you are a novice in coding and things go too fast, do not worry—just team up with someone in class. Below is a short guideline for installing the right Python distribution and packages. *Important:* I will not have the time to troubleshoot installation problems during class. If there are errors, it is a good exercise to try to solve the problem using stackoverflow.com or the package documentation. If you cannot get the software to run after honest effort, please contact me before the course starts.

We all want the sessions to be interesting and enjoyable, with lots of discussion. Therefore, feel free always to ask clarifying questions. This is a methods course, so the onus is ensuring you understand and digest the ideas and can apply the methods. We will look at a couple of use cases, and I will ask you to critique the use of methods and critically discuss the limitations of inference and possible extensions.

Required Level of Python Skills

Given the time we have, I cannot give a full introduction to everything Python can do. But I will introduce everything you need to know to apply textual analysis methods. Learning and internalizing these methods takes

practice! The course is designed to give you code and pointers to help you practice and apply the learned material to your own problems.

If you are a coding novice, I encourage you to take one of the many excellent free introduction tutorials that can be found online (see, for instance, the Datacamp course catalog). You will likely have a much smoother ride if you invest three hours into such an online tutorial beforehand.

If my in-class live coding is a bit too fast for you, do not worry. This course is based a lot on group work—team up with one or two of your fellow class mates.

Required Software

Please bring your own laptop. If you do not have one, create a GitHub account at https://github.com/ and send me your user ID. I will then grant you access to a GitHub repository. It contains all further software installation instructions. It also contains all the teaching material and papers.

Course Policies

Grading Policy

This is a pass/fail course that depends on your participation. The deliverable is a short 2-page research note afterward. I require you to be present 80% of the days to pass the course successfully.

E-mail Policy

You can always write me an e-mail (harm.schuett@whu.edu); I am sometimes a bit slow to respond, but I will respond. Generally, if you have an administrative question, please look at the syllabus first. Emails with questions that should be clear from just looking at the syllabus will be a very low priority.

References

Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy. (2019) "Text as Data." *Journal of Economic Literature* 57(3).

Dikolli, Shane S., Thomas Keusch, William J. Mayew, and Thomas D. Steffen (2020) "CEO behavioral integrity, auditor responses, and firm outcomes." *The Accounting Review* 95(2): 61-88.