Automated Textual Analyses Syllabus

Instructor

Harm H. Schütt
Assistant Professor
Tilburg University

Email

h.h.schutt@tilburguniversity.edu

Course dates

tbd

Course location

tbd

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 tool set which enables researchers to measure previously hard to measure latent concepts using text data.

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

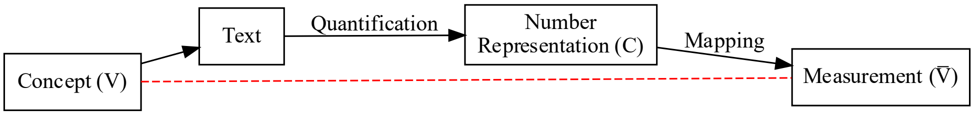

Figure 1: Text as data framework (Gentzkow, Kelly, Taddy, 2019)

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

Participants will be introduced to commonly applied approaches for both steps and will learn to reason about which approaches are advisable given the text at the 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.

Course Schedule

Day	Subject	Practice
Day 1	Measuring concepts with text data - Overview of methods - Overview and rationale of the course content - Lookahead to interesting 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 modelling library	Explore upcoming trends in scientific articles
	Automating large language models and its use as measurement tools	

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 analyze them every statement made there. I just want you to have heard the terms and be familiar with some ideas before coming to 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 roughly divided into two-thirds lecture and one-third coding and 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. To make this work, you need to bring your own laptop with the required software installed and ready to go. A short guideline to installing the right Python distribution and packages is attached below. *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 to always ask clarifying questions. This is a methods course, so the onus is on making sure you understood 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 with you 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 give an introduction that will cover 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 on your own problems.

If you can, I highly 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 maybe three hours into such an online tutorial beforehand.

Required Software

Please bring your own laptop and install the latest Anaconda installation before coming to class (Python 3.xx 64bit). Anaconda is a very convenient scientific Python distribution. That means it automates installing Python for scientific work and gives you simple tools to keep everything up to date (which is not necessarily simple if you would do that manually, especially if you are not on Linux). It makes working with Python very easy, even for people without a programming background. Make sure you install the latest Anaconda Python 3.x distribution and not the old 2.7 distribution. You need new versions of spacy, so if you have a python installation installed, you might need to re-install or create a new conda environment (instructions can be found on the anaconda website).

Go into your start menu and opening the anaconda prompt **with administrator rights**. On windows, this sometimes involves right-clicking on the anaconda prompt and choosing (more)/"open in administrator mode". (For Mac users, open a terminal. It should be tucked away in your "Utilities" folder). Once you have the prompt (terminal) open, type (type each step in one line , or copy and paste) and execute the first line, then type and execute the second line, the third, the fourth, and finally the fivth:

```
conda update –-all

conda install -c conda-forge spacy

conda install -c conda-forge genism

conda install -c conda-forge textstat

      (Alternatively: pip install --upgrade gensim

                      pip install textstat

      )

python -m spacy download en_core_web_sm
```

If all of this was successful, you should be able to execute the following code in any python prompt without any (error) message:

```
import pandas as pd
import spacy
import gensim
import textstat
nlp = spacy.load("en_core_web_sm")
```

Spacy install errors often arise, if you did not install it using administrator rights. If this does not solve the problem, check the docs at spaCy · Industrial-strength Natural Language Processing in Python

Course Policies

Grading Policy

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

E-mail Policy

You can always write me an e-mail (h.h.schutt@tilburguniversity.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 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.