Quantitative Text Analysis I

IPSA-NUS Methods Summer School
July 2020

Dr. Dani Madrid-Morales | University of Houston | dmmorales2@uh.edu

Course Description

Over the last two decades, during which the world has seen the spread of digital technologies to almost ubiquity, and the sprawl of communication networks worldwide, humans have generated more textual data than in the previous 1,000 years combined. Not only has the production of text records grown exponentially in recent years, but also our ability to access, store and analyze them. Today we are able to seek answers to questions that we were unable to tackle previously; we can test hypotheses that require large amounts of computing power; and, we can revisit theories that were long considered settled science.

This course introduces you to some of these advances in quantitative text analysis methods used to systematically extract information from large amounts of texts. It starts with a very brief overview of traditional approaches to analyzing texts, such as manually-coded content analysis, before moving on to computational methods that treat text as data. You will learn different automated forms of text acquisition (e.g. web scraping, API…) and pre-processing techniques (e.g., tokenization, stemming, lemmatization); dictionary-based approaches, such as sentiment analysis, as well as scaling of political texts and supervised text classifiers. The course combines lectures with hands-on labs that will allow you to practice and apply newly acquired skills on a daily basis.

Prerequisites

While there are no formal prerequisites, it would be beneficial if you were familiar with basic statistical concepts (e.g. univariate and multivariate analysis), and had some experience with the statistical software R. However, even if you are unfamiliar with these concepts and tools, you will still be able to participate in the course.

In preparation for the course, particularly if you haven't used R for some time, I would strongly recommend that you take one of the many free introductions to R available online. Below are a few options:

- https://www.pluralsight.com/courses/r-programming-fundamentals
- https://datacarpentry.org/r-socialsci/
- https://rstudio.cloud/learn/primers
- https://www.dataquest.io/course/intro-to-r/

Software

All analyses in this course will be implemented in R, a free open-source programming language commonly used in data science. We will be using RStudio Cloud

(http://rstudio.cloud) to run the labs during this course. If you want to, you can also install R and RStudio on your computer

Starting on July 6 (the first day of the course), you will receive 6 months of free access to DataCamp, "the most intuitive learning platform for data science." You will be able to take 200+ courses in R and other programming languages and learn by using "a combination of short expert videos and hands-on-the-keyboard exercises."

Recommended Texts

There is no one single textbook that covers all the content that you will learn in Quantitative Text Analysis I. However, the following books can be useful, as they cover some important parts of the course.

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology* (4th Ed.). Thousand Oaks: Sage.

Silge, J., & Robinson, D. (2017). *Text Mining with R*. Sebastopol: O'Reilly Media. Available at: https://www.tidytextmining.com/.

Wickham, H., & Grolemund, G. (2016). *R for data science: Import, tidy, transform, visualize, and model data* (First edition). Sebastopol, CA: O'Reilly. Available at: https://r4ds.had.co.nz.

I have also curated a list of book chapters, academic journal articles, and other readings for each of the topics we will cover. You can find the list below.

Communication Tools

The main communication tool for this course will be slack, which is a widely used platform in the media and IT industries. You will able to communicate with the instructor and with your classmates through slack. It will also be a useful to work on group assignments, and to post questions about how to solve homework and in-class exercises. Make sure to install the slack desktop version on your computer. You will receive an invitation to join the slack workspace before the beginning of the course.

Evaluation

An optional written exam will be held on July 18. The exam will consist of ten multiple choice questions and ten short answer questions. You will have sixty minutes to complete the exam, which will be exclusively based on the contents covered in class.

Course structure

You will find all the course materials at https://danimadrid.net/teaching/qta. For each day, I will provide a series of pre-recorded videos that cover the context and the theoretical underpinnings of our labs. You should watch these videos before joining our daily synchronous lab on Zoom. Labs will start at 10:30am. I will record the labs so that, those who cannot make, can watch them at a later time.

Expect to spend between 70 to 100 minutes watching the pre-recorded videos, in which I will cover the theoretical and methodological foundations of one method, tool or technique. The lab session will last approximately 120 minutes. During the labs, you will have the chance to ask questions about the contents of the videos, and I will walk you through some demo code, and help you solve a series of coding problems and exercises.

At the end of each class you will be provided with some additional (and optional) practice material to be completed individually or in small groups. I encourage you to use Slack to communicate with each other as you work on these exercises.

Schedule & Readings

This is a tentative schedule for Quantitative Text Analysis I. Based on our progress, your feedback and your specific learning goals, I might make changes to it as we go. While we will not be specifically discussing the readings outlined below, it will be useful if you read them before coming to class.

Day 1 | Content Analysis Overview & Acquisition of Text Data

This session will cover the fundamentals of quantitative text analysis, with a special focus on the differences between non-computer aided content analysis, and computational approaches. We will review important concepts such as sampling, reliability, validity, and discuss seminal research projects that use content analysis as a method. In the hands-on-session, different data acquisition techniques online will be introduced.

Recommended readings:

Krippendorff, K. (2004). Reliability in Content Analysis.: Some Common Misconceptions and Recommendations. *Human Communication Research*, *30*(3), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

Krippendorff, K. (2019). Conceptual Foundation. In *Content analysis: An introduction to its methodology* (4th ed, pp. 24–46). Thousand Oaks: Sage.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., … Alstyne, M. V. (2009). Computational Social Science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742

Day 2 | Descriptive Computational Text Analysis Methods & Data Pre-processing

This session will focus on three topics. First, it will introduce computational text analysis in general, and discuss the advantages and disadvantages of using the "bag of words" approach to text analysis. Second, it will offer an overview of different quantitative methods to describe textual data, including some commonly used measures. And, third, it will discuss in detail how to process raw text data (e.g. stemming, removing punctuation, removing stop-words…) in order to build a corpus that can be used by most text analysis packages.

Recommended readings:

Banks, G. C., Woznyj, H. M., Wesslen, R. S., & Ross, R. L. (2018). A Review of Best Practice Recommendations for Text Analysis in R (and a User-Friendly App). *Journal of Business and Psychology*, *33*(4), 445–459. https://doi.org/10.1007/s10869-017-9528-3

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Lucas, C., Nielsen, R., Roberts, M., Stewart, B., Storer, A., & Tingley, D. (2015). Computer assisted text analysis for comparative politics. *Political Analysis*, *23*(2), 254–277. https://doi.org/10.1093/pan/mpu019.

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text Analysis in R. *Communication Methods and Measures*, *11*(4), 245–265. https://doi.org/10.1080/19312458.2017.1387238

Day 3 | Dictionary Based Approaches & Sentiment Analysis

This session will introduce one of the most commonly used techniques in computational text analysis. Dictionary-based methods use pre-defined lists of terms, each of which has a value assigned to it, to describe texts in one or more dimensions. Often-used dictionaries include those trained for sentiment analysis (e.g. LWIC, Harvard IV-4…). You will learn how to build a dictionary, how to test and refine it, and how to apply it to different types of text data using the quanteda and/or tidytext packages.

Recommended readings:

Silge, J., & Robinson, D. (2017). Sentiment analysis with tidy data. In *Text Mining with R* (pp. 13–29). Sebastopol: O'Reilly Media.

Soroka, S., Young, L., & Balmas, M. (2015). Bad News or Mad News? Sentiment Scoring of Negativity, Fear, and Anger in News Content. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 108–121. https://doi.org/10.1177/0002716215569217

Young, L., & Soroka, S. (2012). Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication*, *29*(2), 205–231. https://doi.org/10.1080/10584609.2012.671234

Zhang, L., & Liu, B. (2017). Sentiment Analysis and Opinion Mining. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 1152–1161). https://doi.org/10.1007/978-1-4899-7687-1_907

Day 4 | Machine Learning for Text Analysis & Naïve Bayes Classifiers

This session will introduce one type of supervised machine learning approach to classify texts into pre-defined categories. Classification methods use manually-coded training sets to "learn" patterns, which are then applied to an unseen group of texts, that can be classified into groups automatically. You will learn the rationale behind one of the most popular classifiers, the Naïve Bayes classifier, and how its results can be validated and visualized.

Recommended readings:

Jurafsky, D., & Martin, J. H. (2018). Naive Bayes Classification and Sentiment. In *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Stanford. Available at: https://web.stanford.edu/~jurafsky/slp3/4.pdf

Lantz, B. (2015). Probabilistic Learning - Classification Using Naive Bayes. In *Community Experience Distilled*. *Machine learning with R: discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R* (2nd ed., pp. 89–124). Birmingham: Packt Publishing.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Text classification and Naive Bayes. In *Introduction to information retrieval* (pp. 234–265). New York: Cambridge University Press.

Day 5 | Scaling of Political Texts

This session builds on the study of classification methods, particularly the Naïve Bayes model, to introduce different forms of scaling of political texts. You will learn how to use Wordscores, an algorithm that uses a pre-trained set of documents to place an unseen set of texts along a continuous scale. If time permits, you will also be introduced to unsupervised methods for scaling such as Wordfish and Wordshoal, both of which assume Poisson distributions.

Recommended readings:

Bruinsma, B., & Gemenis, K. (2019). Validating Wordscores: The Promises and Pitfalls of Computational Text Scaling. *Communication Methods and Measures*, *13*(3), 212–227. https://doi.org/10.1080/19312458.2019.1594741

Laver, M., Benoit, K., & Garry, J. (2003). Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review*, *97*(2), 311–331. https://doi.org/10.1017/S0003055403000698

Lowe, W. (2008). Understanding Wordscores. *Political Analysis*, *16*(4), 356–371. https://doi.org/10.1093/pan/mpn004

Slapin, J. B., & Proksch, S.-O. (2008). A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, *52*(3), 705–722. https://doi.org/10.1111/j.1540-5907.2008.00338.x