Quantitative Text Analysis II

IPSA-NUS Methods Summer School
July 2020

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

Course Description

The exponential growth in computing power, and the simultaneous decrease in computing costs, that started in the mid 2000s has given way to the development of complex algorithms that are able to classify, scale and map textual data with limited input from humans. These algorithms are often grouped under the label of unsupervised machine learning, and will be the focus of this course.

In the first half of Quantitative Text Analysis II, you will move from studying supervised methods for quantitative text analysis (such as dictionary approaches and Naïve Bayes) to unsupervised methods (such as Wordfish, Latent Dirichlet Allocation and Structural Topic Modelling). During the second half of the course, we will explore different applications of network analysis to QTA, and you will be introduced to some of the most recent developments in the field, such as machine translation and word embeddings.

Quantitative Text Analysis II adopts a project-based approach. Because students are assumed to have some basic prior knowledge on computational methods for the analysis of text, this course will provide the opportunity to implement those skills in the collection, analysis and interpretation of textual data of each students' choice. At the beginning of the week students will be asked to build (collect, clean and process) a large dataset of texts. After learning how to describe it and visualize some of the main metrics, students will use unsupervised methods to classify, scale and map the content with the goal of testing hypotheses, or answering research questions of interest. Students will be able to work individually, or in groups of 2 or 3.

Prerequisites

Students taking Quantitative Text Analysis II should be familiar with simple computational text analysis methods such as dictionary-based approaches (e.g. sentiment analysis), as well as supervised machine learning (e.g. Naïve Bayes classifiers) and document scaling (e.g. Wordscores). More specifically, students should have some familiarity with the R programming language and the text analysis packages quanteda and tidytext. Some knowledge of web scraping is preferred.

Software

All analyses in this course will be implemented in R, a free open-source programming language commonly used in data science. You will need to have R and RStudio installed in your computer to take this course. Instructions on how to install the software will be

provided before the start of the course. For students without admin privileges on their computer (i.e. those unable to install any software), the instructor will provide access to RStudio Cloud.

Starting on the first day of the course, students will receive 6 months 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." A selection of courses and exercises will be recommended at the end of each lecture to reinforce the material learned in class.

Recommended Texts

There is no one single textbook that covers all the content that you will learn in Quantitative Text Analysis II. However, the following books can be useful, as they cover in detail some important components of the course. Jurafsky & Martin (2018) is rich in theoretical and mathematical descriptions of methods, while Lantz (2015), and Silge & Robinson (2017) offer applied examples using R.

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

Lantz, B. (2015). *Machine learning with R: Discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R* (Second edition). Birmingham: Packt Publishing.

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

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 at the end of the syllabus.

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 10:00am (SG time)**. I will record the labs so that, those who cannot make it to the live session, can still watch them at a later time.

Expect to spend between 50 and 70 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 2.5 hours, including time for individual/group project time.

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.

On the last day of the course, each group of students (or individual students) will present the results of their week-long project to the rest of the group.

Schedule & Readings

This is a tentative schedule for Quantitative Text Analysis II. 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 | Data wrangling and web scraping

After a quick review of the different ways in which supervised quantitative text analysis methods can be of use to social scientists, we will explore some ways in which text pre-processing can help us improve the accuracy of QTA models (e.g. concordance analysis, bigrams, KWIC…). The largest component of the lab will be focused on learning how to crawl websites and scrape content from them.

Recommended readings:

Freelon, D. (2018). Computational Research in the Post-API Age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Trilling, D., & Jonkman, J. G. F. (2018). Scaling up Content Analysis. *Communication Methods and Measures*, *12*(2–3), 158–174. https://doi.org/10.1080/19312458.2018.1447655

Wesslen, R. (2018). Computer-Assisted Text Analysis for Social Science: Topic Models and Beyond. *ArXiv:1803.11045 [Cs]*. Retrieved from http://arxiv.org/abs/1803.11045

Recommended DataCamp course:

Introduction to Text Analysis in R

Day 2 | From Supervised to Unsupervised Text Analysis

In this session you will be introduced to two of the most frequently used unsupervised machine learning algorithm for text analysis: Latent Dirichlet Allocation (LDA) and Structural Topic Modeling (STM). We will discuss how to determine parameters in the algorithms, how to analyze the results, and how to validate the findings. The benefits of using this approach will be introduced and, again, you will have the opportunity to test the algorithms, using R's quanteda and stm packages

Recommended readings:

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., … Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, *58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Silge, J., & Robinson, D. (2017). Topic Modeling. In *Text Mining with R* (pp. 89–108). Sebastopol: O'Reilly Media.

Recommended DataCamp course:

Supervised Learning in R: Classification

Day 3 | Text as Networks and Visualization

The relationship between words in a text can be often best understood visually. In this session, you will learn some ways in which network analysis, from semantic networks to networks of co-occurrence, has been applied to the study of textual data. After experimenting with different R packages for network analysis, the project section for today will focus on applying these packages to your own text dataset.

Recommended readings:

Atteveldt, W. van. (2008). Knowledge Representation and the Semantic Web. In *Semantic network analysis: techniques for extracting, representing and querying media content* (pp. 51–62). Charleston, SC: BookSurge.

Bail, C. A. (2016). Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(42), 11823–11828. https://doi.org/10.1073/pnas.1607151113

Silge, J., & Robinson, D. (2017). Relationships between words: n-grams and correlations. In *Text Mining with R* (pp. 45–68). Sebastopol: O'Reilly Media.

Recommended DataCamp course:

Unsupervised Learning in R

Day 4 | Beyond the Bag of Words Approach: Word Embeddings

In this session, you will be introduced to the most common alternative to the "bag of words" approach to text analysis: word embeddings, which use vectorized representations

of texts. While computationally rather intensive, word embeddings widen the scope of questions that other QTA methods are able to answer. In the project section of the class, each student will wrap up their analysis and briefly present their journey (and results) to the group.

Recommended readings:

Jurafsky, D., & Martin, J. H. (2018). Vector Semantics. In *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Retrieved from https://web.stanford.edu/~jurafsky/slp3/

Rheault, L., & Cochrane, C. (2019). Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora. *Political Analysis*, 1–22. https://doi.org/10.1017/pan.2019.26

Spirling, A., & Rodriguez, P. L. (2019). *Word Embeddings What works, what doesn't, and how to tell the difference for applied research*. New York.

Recommended DataCamp course:

Text Mining: Bag of Words

Day 5 | Working with Multilingual Data

Building on the study of word embeddings from the previous day, we will spend time understanding how word embeddings can be used to improve the analysis of corpora that include data in multiple languages. We will also test the accuracy of using machine translation to in text analysis. During the last part of the session, each student/group will present the results of their week-long project using the QTA skills learned in this class.

Recommended readings:

de Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications. *Political Analysis*, *26*(4), 417–430. https://doi.org/10.1017/pan.2018.26

Lind, F., Eberl, J.-M., Heidenreich, T., & Boomgaarden, H. G. (2019). When the Journey Is as Important as the Goal: A Roadmap to Multilingual Dictionary Construction. *International Journal of Communication*, *13*(0), 21.

Proksch, S.-O., Lowe, W., Wäckerle, J., & Soroka, S. (2019). Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches. *Legislative Studies Quarterly*, *44*(1), 97–131. https://doi.org/10.1111/lsq.12218