Big Data and Social Media Research

Pablo Barberá
London School of Economics

Course description

Citizens across the globe spend an increasing proportion of their daily lives online. Their activities leave behind granular, time-stamped footprints of human behavior and personal interactions that represent a new and exciting source of data to study standing questions about political and social behavior. At the same time, the volume and heterogeneity of web data present unprecedented methodological challenges. The goal of this course is to introduce participants to new computational methods and tools required to explore and analyze Big Data from online sources using the R programming language. We will focus in particular on data collected from social networking sites, such as Facebook and Twitter, whose use is becoming widespread in the social sciences.

Each session will provide an overview of the literature and research methods on a particular theme to then dive into a specific application, documenting each step from data collection to the analysis required to test hypotheses related to core social science questions. Code and data for all the applications will be provided. The course will follow a "learning by doing" approach, and participants will be asked to complete a series of coding challenges.

The first session will begin with a discussion of the definition of "Big Data" and the research opportunities and challenges of the use of massive-scale datasets in the social sciences. We will then focus on how social media sites represent a new source of data to study human behavior, and also how its use raises a whole new series of questions that are relevant to social scientists. The applied part of this session will provide a foundation of R coding skills upon which we will rely during the rest of the course. Here, we will go over existing packages to efficiently analyze large-scale datasets in R, how to parallelize for loops, and how to read and write large files.

The second session will focus on the most common application of Big Data in the social sciences: large-scale text classification. After a quick overview of the basics of machine learning, we will discuss specific details of the implementation of supervised learning algorithms in massive-scale datasets. Our emphasis will lie on the practical aspects: we will study these methods in the context of two applications: sentiment analysis of social media posts and ideological scaling of party manifestos. We will go through the entire research process, from the creation of a training dataset labeled by humans using crowd-sourcing platforms, to the application and validation of the machine learning algorithm, and passing through all the intermediate steps, such as cleaning and preprocessing the corpus of documents.

Exploratory data analysis can be a powerful tool for social scientists when they are interested in analyzing a new dataset. The third session will cover the existing tools for large-scale discovery in "Big Data" using R, applied to textual datasets. We will start with different techniques, such as collocation analysis, keyness, and readability, which will allow us to identify salient themes and ideas across documents. Then, we will move

to topic models, which allow researchers to automatically identify latent classes of documents in a corpus, with an application to the classification of Facebook posts by politicians into relevant political issues. This session will also cover other dimensionality reduction techniques that are commonly used in the social sciences to visualize large-scale datasets.

After the course, students will have an advanced understanding of the opportunities of big data and social media mining for social science studies, and will be equipped with the technical skills necessary to conduct their own research.

Software

The course will use the open-source software R, which is freely available for download at https://www.r-project.org/ . We will interact with R through RStudio Server. Each participant will be given access to the server used for the course, where all packages will be already installed, and which can be accessed through any web browser. Therefore, participants are not required to install any software before the course.

Prerequisites

The course will assume familiarity with the R statistical programming language at the level of the introductory course offered by the summer school. Participants should be able to know how to read datasets in R, work with vectors and data frames, write functions and loops, and run basic statistical analyses, such as linear regression.

Schedule

July 2nd, 2018

Time	Topic
9.00-9.30	Introductions and course overview
9.30-10.00	What is Big Data? The 4 V's of Big Data. Research opportunities and challenges in the social sciences.
10.00-10.45	What can we learn from web and social media data? Overview of social media research: theories, methods, and data.
10.45-11.15	Break
11.15-11.45	Good practices in R
11.45-12.30	Introduction to parallel computing with R
12.30-13.00	Challenge 1: writing efficient R code

July 3rd, 2018

Time	Topic
9.00-9.45	Introduction to automated text analysis.
9.45-10.15	Application: Sentiment analysis
10.15-10.45	Challenge 2: Sentiment analysis of political tweets
10.45-11.15	Break
11.15-12.00	Text classification at scale using machine learning

12.00-12.30	Application: Measuring incivility in online communication
12.30-13.00	Challenge 3: Automatic classification of newspaper articles

July 4th, 2018

Time	Topic
9.00-9.30	Exploratory analysis of large-scale text datasets
9.30-10.15	Application: Longitudinal changes in presidential rhetoric
10.15-10.45	Challenge 4: Exploratory analysis of tweets by Members of Congress
10.45-11.15	Break
11.15-11.45	Unsupervised machine learning: topic models
11.45-12.15	Application: Media coverage of the state of the economy
12.15-13.00	Challenge 5: tracking agenda-setting on Facebook using topic models

References

Course largely based on:

Day 1

Main readings:

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., … & Jebara, T. (2009). Life in the network: the coming age of computational social science. *Science*, 323(5915), 721-3.

Lazer, D., & Radford, J. (2016). Introduction to Big Data. *Annual Review of Sociology*, *43*(1).

Matloff, N. (2011). *The art of R programming: A tour of statistical software design*. No Starch Press.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063-1064.

Recommended readings:

Barberá, P., & Rivero, G. (2014). Understanding the political representativeness of Twitter users. *Social Science Computer Review*, 0894439314558836.

Golder, S. A., & Macy, M. W. (2014). Digital footprints: Opportunities and challenges for online social research. *Sociology*, *40*(1), 129.

Nagler, J. (1995). Coding style and good computing practices. *PS: Political Science & Politics*, 28(3), 488-492.

Nagler, J., & Tucker, J. A. (2015). Drawing inferences and testing theories with big data. *PS: Political Science & Politics*, *48*(01), 84-88.

Grimmer, J. (2015). We are all social scientists now: how big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, *48*(01), 80-83.

Monroe, B. L., Pan, J., Roberts, M. E., Sen, M., & Sinclair, B. (2015). No! Formal theory, causal inference, and big data are not contradictory trends in political science. *PS: Political Science & Politics*, *48*(01), 71-74.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... & Jebara, T. (2009). Life in the network: the coming age of computational social science. *Science (New York, NY)*, *323*(5915), 721.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063-1064.

Salganik, M. (2017). *Bit by Bit: Social Research in the Digital Age.* Princeton, NJ: Princeton University Press.

Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *ICWSM*, 14, 505-514.

Day 2

Main readings:

González-Bailón, S., & Paltoglou, G. (2015). Signals of public opinion in online communication: A comparison of methods and data sources. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 95-107.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, mps028.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). New York: springer.

Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2), 205-231.

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S., & Parnet, O. (2016) A Bad Workman Blames His Tweets? The Consequences of Citizens' Uncivil Twitter Use When Interacting with Party Candidates. *Journal of Communication*, forthcoming.

Recommended readings:

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2015). Crowd-sourced text analysis: reproducible and agile production of political data. *American Political Science Review*.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, *97*(02), 311-331.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1), 24-54.

Day 3

Main readings:

Bauer, P. C., Barberá, P., Ackermann, K., & Venetz, A. (2017). Is the Left-Right Scale a Valid Measure of Ideology?. *Political Behavior*, *39*(3), 553-583.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Albertson, B., ... & Rand, D. (2014). Topic models for open ended survey responses with applications to experiments. *American Journal of Political Science*, *58*, 1064-1082.

Recommended readings:

Gilardi, F., Shipan, C. R., & Wueest, B. (2017). Policy Diffusion: The Issue-Definition Stage. Working paper, University of Zurich.

Short bio

Pablo Barberá (PhD in Politics, New York University) is an Assistant Professor of Computational Social Science at the London School of Economics, and a former Moore-Sloan Fellow at the Center for Data Science in New York University. His primary research interests include quantitative political methodology and computational social science applied to the study of political and social behavior. He is an active contributor to the open source community and has authored several R packages to mine social media data. His research has been published in high-impact journals such as Political Analysis, International Studies Quarterly, Journal of Communication, PLOS ONE, Psychological Science, Political Science Research and Methods, the Journal of Computer-Mediated Communication, and Social Science Computer Review, among others. More information is available at www.pablobarbera.com