Towards a reproducible framework for undergraduate data science research

Frederick J. Boehm and Bret M. Hanlon

8/20/2019

Contents

Last modified: 2019-12-02 11:23:57

1 Abstract

We develop a mentoring framework to guide undergraduate researchers through individualized research projects in social media data science. Our framework involves research question formulation, data acquisition, data analysis and visualization, and presentation and communication of results. Our two honors students, whose projects serve as case studies for our framework, completed all components of the individualized research projects. We found that data science research skills, self-confidence in research ability, and professional interest in data science increased for both students. We describe our successes, lessons learned, and ideas for others to build similar frameworks.

2 Introduction

- big data
- new areas of application of statistical methods - social media
- need for students to work with real data

- motivate statistical analysis and statistics research by real world scientific research questions (Box 1976)
- need for mentoring

The need to analyze unprecedentedly large volumes of information combined with the development of faster and more powerful computers has fueled advances in data science methods for big data. We present below a framework for mentored undergraduate data science research and our findings from its initial implementation for two statistics students. We conclude with lessons learned and suggestions for those who wish to reproduce and refine our framework.

Social media data, including tweets from Twitter and posts from Facebook, are available through website application product interfaces (APIs). Twitter shares, via a streaming API, a sample of approximately one percent of all tweets during an API query time period ("Sampled Stream," n.d.). Researchers have studied tweets for a variety of purposes, including inference of relationships and social networks among users (Lin et al. 2011); determination of authorship of specific tweets when multiple persons share a single account (Robinson, n.d.); and study of rhetoric in recruiting political supporters (Pelled et al. 2018,@wells2016trump). Recognizing the potential utility of tweets for data science research, we created a collection of tweets over time by repeated querying of the Twitter streaming API.

Nolan and Temple Lang (2010) argue for students to work with real data. Working with real data allows them to develop skill not only in statistical analysis, but also in data transfer from online sources, in data storage, and in using data from multiple file formats. In the case of Twitter data, tweets are stored in Javascript Object Notation (JSON) ("Consuming Streaming Data," n.d.,@json).

Mentoring in the work place and in higher education can have many benefits, including improving students' development as thinkers and scholars, confidence in their own abilities, integration into the campus community, and interest in graduate training (Baker and Griffin 2010,@higgins2001reconceptualizing). A key component of our data science mentoring framework is the emphasis on using real data to answer real scientific questions. We believe that this process develops problem-solving skills that students will need in their future careers in data science. We encouraged the student to articulate a scientific research question, translate that question into quantitative and statistical terms, determine which data could be used to address the question, acquire the data, analyze data, visualize results, and communicate what they learned.

We provide guidance regarding selection of

Need to explain above paragraph with our two student examples. What types of questions did they articulate? How did they translate to quant terms? How did they determine data availability? etc.

2.1 Our backgrounds

During the time when we first implemented our framework, we served as early-career instructors in the statistics department at the University of Wisconsin-Madison. One of us (Hanlon) had prior experience in mentoring students, while the other (Boehm) had none. Our initial conceptualization of mentoring drew heavily on ideas we first encountered in professional development activities, including the Delta Program's mentoring class (https://delta.wisc.edu) and Handelsman et al. (2005). Professor Erik Nordheim heavily influenced our approach to and philosophy of teaching statistics. We studied with Professor Nordheim early in our teaching careers, and his emphasis on active learning continues to influence our teaching practices.

We both have experience in teaching undergraduate introductory statistics courses with enrollments over 100 students. Through our interactions with students in these classes, we've grown to value not only the ideas in a traditional introductory course, but also the need to prepare students with the essential skills needed for success in data science. Nolan and Temple Lang (2010) summarizes these skill sets in the following three goals:

1. broaden statistical computing to include emerging areas
2. deepen computational reasoning skills
3. combine computational topics with data analysis in the practice of statistics

To these three praiseworthy goals, we add a fourth:

4. develop skills in reproducible research to promote open science practices

We see the fourth goal as an equal with the first three from Nolan and Temple Lang (2010). Data scientists are uniquely positioned to promote open science practices, including the free sharing of data, code, and instructions for their use. The need for science to be more transparent and more reproducible elevate this goal to the level of the first three.

Below, we detail our methods for creating a reproducible framework for undergraduate data science research. We describe our results before concluding with lessons learned, things we could have done differently, and recommendations for future mentors who may use and extend our framework.

3 Methods

We designed and used a framework for mentored undergraduate data science research projects with big data. Below, we describe our initial implementation of the framework before relating it to three major ideas from Nolan and Temple Lang (2010).

3.1 Framework implementation

Our mentored research framework begins with brainstorming scientific research ideas based on the student's interests. This enables us to craft a project that excites the student. With the results of brainstorming sessions, we (mentors and student together) formulate the most promising ideas into scientific hypotheses.

For the most appealing scientific hypotheses, we encourage the student to translate the scientific question into a statistical question that may be addressed with data. This is a crucial step in data science research question formulation. Skill in translating in both directions between scientific and statistical questions is a key communication skill that data science researchers offer.

We also incorporated data availability into our question formulation. We limited questions to those that could be studied with publicly available data. This practice also enabled reproducibility of our analyses, since students could share the URL from which they accessed data.

After identifying research questions and publicly available data, the next step is to decide on informative data visualizations and quantitative analyses. Because both projects primarily involved exploratory analyses of times series, we encouraged students to think about visualizations that might reveal relationships over time.

3.1.1 Examples

Examples may help to demonstrate our approach to identifying a statistical research question. One of our students had interests in acquiring and using social media posts. We helped her in brainstorming ideas for research involving social media sources like Facebook and Twitter. Through this brainstorming, we recognized that she had a parallel interest in financial markets. Our student hypothesized that sentiment analysis of finance-related tweets might reflect trends in financial market index prices. On days when the market index prices increase, sentiment analysis of finance-related tweets might reveal more use of positive words, while days with decreasing prices might have more negative words in finance-related tweets.

A second student wanted to study tweets over time and entertainment events that garner lots of attention in social media. We encouraged this student to develop a strategy for event detection from tweets over time. The rationale is that a big entertainment event, such as the National Football League's Super Bowl game, might generate enough tweets that Super Bowl-related words would appear with high weights in results from latent Dirichlet allocation modeling of collections of tweets at distinct time points. We reasoned that Super Bowl-related topics might appear during the Super Bowl and vanish soon after the game's conclusion.

3.2 Relating to three ideas from Nolan and Temple Lang (2010)

We incorporate three key aspects that Nolan and Temple Lang (2010) identified:

1. broaden statistical computing to include emerging areas
2. deepen computational reasoning skills
3. combine computational topics with data analysis in the practice of statistics

Additionally, our projects gave students opportunities to develop and to practice skills in reproducible research. Given the growing imperative to document and share code to promote open science, we feel that this skill set equals in importance the three points above.

4. develop and practice skills in reproducible research to promote open science

Below, we describe how our framework enabled students to achieve competence in the four areas listed above.

3.3 Broaden statistical computing to include emerging areas

Our framework broadens statistical computing by including the emerging areas of social media data analysis, sentiment analysis, and topic modeling. Both students used Twitter tweets, which we accessed through a Twitter streaming API.

Our computational system for acquiring tweets involved several steps. We interacted with the API via the R package `twitteR` (Gentry 2015). We used the free Twitter streaming API that gave us access to approximately one percent stream of all tweets during the specified query time period. To ensure that we collected tweets continuously, we used the linux tool `crontab` to execute our R script every five minutes. Each execution of the R script performed a single streaming API query for five minutes. Twitter's streaming API, at the time of our data collection, enforced rate limits on the frequency and duration of queries. With the above settings, we continuously collected tweets.

3.4 Deepen computational reasoning skills

Our framework encourages students to deepen computational reasoning skills in several ways. First, they work with a variety of internet-based data to answer research questions. In the two example cases, our students collected tweets over time and gathered complementary data from other resources, including daily closing prices of stock market indexes. This gave students opportunities to think creatively about what data to acquire and how to use multiple data sources in a single cohesive project.

Second, the students worked with a variety of data structures. The Twitter streaming API returns tweets as JSON (Javascript Object Notation). Because distinct Twitter users may provide different pieces of profile information, there is variability in the structure of each tweet's JSON. Additionally, tweet metadata fields may appear in any order (https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data). Students needed to recognize this and to write code that accommodated these variations in tweet data structure. Additional variability in tweet structure arose due to changes in the API. The evolving nature of JSON tweet structure (https://developer.twitter.com/en/docs/tweets/data-dictionary/guides/tweet-timeline) required students to write flexible code that could incorporate newly introduced or deprecated metadata.

Students wrote R code to parse and organize tweet JSON []. They organized their R code into a package, and shared it on Github (https://github.com/rturn/parseTweetFiles). Each tweet's JSON included required fields, and, possibly, some optional fields.

3.5 Combine computational topics with data analysis in the practice of statistics

computational topics: LDA & topic modeling; time series analysis; sentiment analysis.

Both mentored students combined computing with data analysis in the practice of statistics. They used a combination of latent dirichlet allocation topic modeling, sentiment analysis, and time series analysis to reach conclusions about real world data.

Both drew heavily on the collection of tweets. One student examined Standard and Poor's 500 index daily closing prices over time. She also analyzed sentiments from each day's stock market-related tweets to look for relationships between tweet sentiment and stock market prices.

Our other student focused on developing detection methods for social media events through topic modeling of tweets at different time periods. As a proof of principle, he fitted topic models to collections of tweets preceding, during, and following the National Football League's Super Bowl game. He hypothesized that topics would evolve over time, with football-related tweets appearing during the football game and disappearing soon after conclusion of the game.

Both students analyzed tweets as texts. This first required them to write code to parse the JSON that the API returns. Once they had isolated the tweet text from its metadata, they parsed the tweet text into words for use in sentiment analysis and topic modeling. For the stock market project, they analyzed only those tweets that contained finance-related keywords. Sentiment analysis involved comparisons of tweet words to a dictionary that mapped words to sentiments. This yielded a net sentiment score for each tweet. They then treated tweet sentiment scores as a time series and compared them with daily stock market index closing prices.

The second student project involved latent Dirichlet allocation modeling of tweet words at distinct time points to detect social media events (Blei, Ng, and Jordan 2003). Latent Dirichlet allocation is a bayesian nonparametric method for modeling text corpora as the result of words chosen from topics.

justify each step

Maybe use/cite Box's science and statistics article from... 1976?

1. Overview of our framework

a. student-led, hypothesis-driven research
b. BH and FB helped with data acquisition
c. helped with research question formulation
d. how to describe roles for students v mentors?

Our framework for mentored student research projects involves these components:

1. State the questions and research hypotheses that the two students chose

2. Data description

- how did we/they conceive of projects?
- how did we balance our input with that of the students?
- What data did we use?
- How did we collect data? (which packages, code to sample twitter). How did Jinyu get her stock market data?
- What is the sampling scheme - for the 1% sample - from the entirety of Twitter? (Cite the webpage that documents the API)

4 Results

We applied the project framework to our mentoring of two students. Both engaged in 12 months of mentored research during their senior year of undergraduate studies in statistics in 2015 and 2016.

4.1 Student Outcomes

We subjectively assessed student outcomes through conversations in our weekly student research meetings. Both students showed increases in confidence and ability to do data science research. While the students benefited

Both students secured positions in data science after graduation. One student later enrolled in a statistics graduate program, while the other pursued employment in health care analytics.

4.2 Mentor Outcomes

4.3 Scholarly outcomes

Our scholarly contributions include the `parseTweetFiles` R package on Github (https://github.com/rturn/parseTweetFiles) and

5 Discussion

5.1 Benefits of our framework

1. build self-confidence, self-assessed proficiency in data analysis and statistics
2. greater interest in quantitative careers and/or grad school in quant fields
3. Working with real data to solve real scientific questions
4. Learn by example interplay between data science (statistics) and science research

Benefits of our framework include enhanced

5.2 How does the framework relate to ideas in Nolan and Temple Lang 2010??

5.3 Critiques of our framework

Our measures of students' self-confidence in research ability was merely subjective. In future iterations of our framework, we would like to measure systematic and objective outcomes. One strategy for implementing this is to administer a survey, such as _____, both before and after the mentored research project. We would use validated survey questions that focused on student beliefs about themselves, their skills, and their future careers. **maybe list resources for such survey questions**

5.4 Areas for improvement

1. reproducible research best practices

One shortcoming of our initial framework was the relative lack of emphasis on best practices for computational reproducibility. This is one area that we would like to rectify in future iterations. The University of Wisconsin-Madison has periodically offered a semester course in best practices for computationally reproducible research. We especially see version control systems, such as Git and Github, as essential tools for the modern data scientist.

1. assessment of attitudes (pre and post survey??)

2. survey questions might include those from Vance et al. (2017). I have the survey as a pdf in mendeley. It is part of the supp info for Vance et al. (2017).

References

Baker, Vicki L, and Kimberly A Griffin. 2010. "Beyond Mentoring and Advising: Toward Understanding the Role of Faculty 'Developers' in Student Success." *About Campus* 14 (6). Wiley Online Library: 2–8.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

"Consuming Streaming Data." n.d. https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data.

Gentry, Jeff. 2015. *TwitteR: R Based Twitter Client*. https://CRAN.R-project.org/package=twitteR.

Handelsman, Jo, Christine Pfund, Sarah Miller Lauffer, and Christine Maidl Pribbenow. 2005. *Entering Mentoring*.

Higgins, Monica C, and Kathy E Kram. 2001. "Reconceptualizing Mentoring at Work: A Developmental Network Perspective." *Academy of Management Review* 26 (2). Academy of Management Briarcliff Manor, NY 10510: 264–88.

"Introducing Json." n.d. https://json.org.

Lin, Cindy Xide, Qiaozhu Mei, Jiawei Han, Yunliang Jiang, and Marina Danilevsky. 2011. "The Joint Inference of Topic Diffusion and Evolution in Social Communities." In *2011 Ieee 11th International Conference on Data Mining*, 378–87. IEEE.

Nolan, Deborah, and Duncan Temple Lang. 2010. "Computing in the Statistics Curricula." *The American Statistician* 64 (2). Taylor & Francis: 97–107.

Pelled, Ayellet, Josephine Lukito, Fred Boehm, JungHwan Yang, and Dhavan Shah. 2018. "'Little Marco,'Lyin'Ted,'Crooked Hillary,' and the 'Biased' Media: How Trump Used Twitter to Attack and Organize." In *Digital Discussions*, 176–96. Routledge.

Robinson, David. n.d. "Text Analysis of Trump's Tweets Confirms He Writes Only the (Angrier) Android Half." http://varianceexplained.org/r/trump-tweets/.

"Sampled Stream." n.d. https://developer.twitter.com/en/docs/labs/sampled-stream/overview.

Vance, Eric A, Erin Tanenbaum, Amarjot Kaur, Mark C Otto, and Richard Morris. 2017. "An Eight-Step Guide to Creating and Sustaining a Mentoring Program." *The American Statistician* 71 (1). Taylor & Francis: 23–29.

Wells, Chris, Dhavan V Shah, Jon C Pevehouse, JungHwan Yang, Ayellet Pelled, Frederick Boehm, Josephine Lukito, Shreenita Ghosh, and Jessica L Schmidt. 2016. "How Trump Drove Coverage to the Nomination: Hybrid Media Campaigning." *Political Communication* 33 (4). Taylor & Francis: 669–76.