Title
Bringing Data Science to Qualitative Analysis

ISBN
9781538691564

Authors
Cheah, You-Wei
Paine, Drew
Ghoshal, Devarshi
et al.

Publication Date
2018-10-01

DOI
10.1109/escience.2018.00076

Peer reviewed

Bringing Data Science to Qualitative Analysis

You-Wei Cheah, Drew Paine, Devarshi Ghoshal, Lavanya Ramakrishnan
Lawrence Berkeley National Laboratory, Berkeley, CA
Email: {ycheah, pained, dghoshal, lramakrishnan}@lbl.gov

***Abstract*—Qualitative user research is a human-intensive approach that draws upon ethnographic methods from social sciences to develop insights about work practices to inform software design and development. Recent advances in data science, and in particular, natural language processing (NLP), enables the derivation of machine-generated insights to augment existing techniques. Our work describes our prototype framework based in Jupyter, a software tool that supports interactive data science and scientific computing, that leverages NLP techniques to make sense of transcribed texts from user interviews. This work also serves as a starting point for incorporating data science techniques in the qualitative analyses process.**

I. Introduction

Today, qualitative user research is a human-intensive approach that draws upon ethnographic methods from social sciences to develop nuanced insights about work practices to inform the design and development of software tools. More recently, data science has evolved as a paradigm employing techniques from disciplines such as statistics, machine learning, pattern recognition, data processing, and visualization. Data Science approaches enable us to derive machine-generated insights that can support and augment human interactions for more in-depth exploration and analyses.

In the Usable Data Abstractions project [1], we are using qualitative user research to study and understand work practices around managing data and workflows on HPC systems to shape the design of next-generation software tools. Recently, we have augmented our qualitative research with data science methods to derive additional insight from semi-structured interviews. Specifically, we have built a prototype tool that uses Natural Language Processing (NLP) techniques to make sense of transcribed texts. Our framework is based in Jupyter [2], a software tool that supports interactive data science and scientific computing. Our work provides a foundation to derive insights using data science to supplement human-intensive analyses since qualitative data analysis work has similarities to NLP data exploration process [3].

II. Background & Related Work

Our qualitative research focuses on the study of scientific researchers' work processes. The work is motivated by related threads of work in Computer Supported Cooperative Work, eScience, and Human-Computer Interaction that have investigated the use of workflow and provenance tools to address the needs of scientific data exploration processes [4], [5], [6], [7]. Developing comprehensive understanding and insights into actual scientific work practices is vital for building next-generation eScience tools. This can be addressed through qualitative studies of science that leverage a variety of data collection methods to understand the scientific process.

Semi-structured user interviews are a common approach used in qualitative user research. Analyses of text artifacts from these interviews is commonly a human-driven process, where people systematically develop insights through "coding" of the artifacts [8]. The coding process can be completed on a range of units of analysis—from single words to entire paragraphs capturing streams of thought. This labor-intensive process is exposed to practices and biases of any humans doing the work, presenting the opportunity to augment these practices with additional tools for developing insights.

The artifacts collected during qualitative investigations are well suited for the application of data science methods and techniques. These techniques have been applied to social science investigations of Twitter users [9]. Prior work by Muller et. al. [3] explores the connections between grounded-theory development and machine learning to posit that these seemingly disparate techniques share some common underlying strategies for making meaning out of data. Such work motivates our use of NLP techniques to aid and augment qualitative data analysis in user research.

We have developed a framework that aims to augment existing methods by providing more insight into the data from natural language processing. For example, a qualitative user researcher trying to understand the challenges faced by scientific users might code the document looking for specific insights on the problems faced by the users. A qualitative researcher using our framework would be able to see additional insights (e.g., sentiment analyses) or context that provides a different perspective on the data.

III. Approach

Our work brings a data science approach to qualitative user research. Specifically, a) we explore NLP techniques to analyze transcripts and b) we provide a user interface to explore the NLP results for further systematic qualitative data analyses.

We have taken an approach to extract key topics or themes that are central to the interview transcripts by leveraging a number of linguistic features such as part-of-speech tags and named-entity recognition. Common NLP techniques such as tokenization, lemmatization, and use of stop-words are used to segment and aggregate texts for improved textual comprehension. Our framework also supports using N-grams and dependency parsing labels to extract more complex or compound keywords. We use the spaCy [10] and NLTK [11]

59	0.7506	If I just take this A goes to B, B goes to C, there's too many holes in the network for it to achieve successful growth in silico, so they do this process called gap fill, where it adds a whole bunch of reactions to the model.
60	0.296	Yeah, because we don't know the function of most of these genes.
61	-0.4767	And many of the functions that we think we know are wrong.
62	0.7717	This is the minimum number of reactions needed to achieve the successful growth on this medium, so therefore it's most likely right.
63	0	Media would be the compounds that you said your organism could grow on.

Fig. 1. Sentiment analysis heatmap with corresponding scores and text.

libraries for these NLP tasks along with algorithms like TextRank [12] to summarize and reduce the amount of text that needs to be processed. Through various visualizations such as Word Cloud and term frequency plots, we are able to visualize and understand the topics and themes that are present in a single transcript, and over a group of transcripts, from different vantage points.

We examine a few sentiment analyses tools, such as TextBlob [13] and Vader [14] algorithms to flag and present alternate sentiment perspectives in our transcripts. Our investigations are tested on a data set of ≈50 semi-structured interviews with scientists to understand their work. These interviews were recorded as audio files, professionally transcribed, then cleaned by the team. The transcripts were open coded multiple times for emerging ideas to develop themes [15].

The intent of our framework is to help qualitative researchers develop different perspectives on their interview data to complement their open coding. The Jupyter notebook framework incorporates the NLP tools noted above so that different ways of looking at this data can be produced. The initial version allows users to parse and tokenize text transcripts into sentences, condenses information through summarization for easier analysis, analyzes sentiment per sentence, and provides a way to visualize and present the data (e.g., see Figure 1). A qualitative researcher can iterate using both their open coding and data science analyses to develop insights about common topics in the interviews and differing sentiments over the course of interviews. For example, this toolset provides an additional perspective for a solo qualitative researcher just beginning to analyze their data. As they open code interviews they can iterate between these analysis approaches when developing themes and insights.

IV. Summary

The work in this paper provides a foundation for incorporating data science techniques for qualitative analyses, and describes our framework for allowing users to augment existing qualitative methods with data science methods. Our early experiences indicate that this provides a valuable tool to augment current qualitative data analyses.

Acknowledgment

This work is supported by the U.S. Department of Energy, Office of Science and Office of Advanced Scientific Computing Research (ASCR) under Contract No. DE-AC02-05CH11231, program manager Lucy Nowell. The authors would also like to thank the various scientists that interviewed, contributing to the dataset that we used for this study.

References

[1] Usable Data Abstractions. [Online]. Available: http://dst.lbl.gov/projects/uda/

[2] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. B. Hamrick, J. Grout, S. Corlay *et al.*, "Jupyter notebooks-a publishing format for reproducible computational workflows." 2016.

[3] M. Muller, S. Guha, E. P. Baumer, D. Mimno, and N. S. Shami, "Machine learning and grounded theory method: Convergence, divergence, and combination," in *Proceedings of the 19th International Conference on Supporting Group Work*, ser. GROUP '16. New York, NY, USA: ACM, 2016, pp. 3–8. [Online]. Available: http://doi.acm.org/10.1145/2957276.2957280

[4] M. Jirotka, C. P. Lee, and G. M. Olson, "Supporting scientific collaboration: Methods, tools and concepts," *Computer Supported Cooperative Work (CSCW)*, vol. 22, no. 4-6, pp. 667–715, 2013. [Online]. Available: http://dx.doi.org/10.1007/s10606-012-9184-0

[5] V. Pipek, H. Karasti, and G. C. Bowker, "A preface to ?infrastructuring and collaborative design?" *Computer Supported Cooperative Work (CSCW)*, vol. 26, no. 1, pp. 1–5, 2017. [Online]. Available: https://doi.org/10.1007/s10606-017-9271-3

[6] S. B. Davidson, A. Eyal, B. Ludäscher, T. M. McPhillips, S. Bowers, M. K. Anand, and J. Freire, "Provenance in scientific workflow systems." 2007.

[7] I. J. Taylor, E. Deelman, D. B. Gannon, and M. Shields, *Workflows for e-Science: Scientific Workflows for Grids*. Springer Publishing Company, Incorporated, 2014.

[8] R. M. Emerson, R. I. Fretz, and L. L. Shaw, *Writing Ethnographic Fieldnotes*. Chicago, IL: The University of Chicago Press, 1995.

[9] D. Karamshuk, F. Shaw, J. Brownlie, and N. Sastry, "Bridging big data and qualitative methods in the social sciences: A case study of twitter responses to high profile deaths by suicide," *Online Social Networks and Media*, vol. 1, pp. 33–43, 2017.

[10] M. Honnibal and I. Montani, "spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing," *To appear*, 2017.

[11] E. Loper and S. Bird, "Nltk: The natural language toolkit," in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ser. ETMTNLP '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 63–70. [Online]. Available: https://doi.org/10.3115/1118108.1118117

[12] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004.

[13] TextBlob: Simplified Text Processing. [Online]. Available: https://textblob.readthedocs.io/

[14] C. H. E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*, 2014.

[15] K. Charmaz, *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*, 2nd ed. Sage, 2014.