# XSEDE Text Analytics Gateway

**Mike Black**[1], **Drew Schmidt**[2]

1. National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign
2.National Institute for Computational Sciences, University of Tennessee
[1]mlblac02@gmail.com
[2]schmidt@math.utk.edu

Increasingly, researchers in the humanities and social sciences are becoming more interested in utilizing computational resources to answer their questions. This is particularly relevant for text analysis, which has historically been dominated in these fields by qualitative methods. We believe the resistance to quantitative methods is, in part, due to a "technological familiarity gap" in these fields.

The XSEDE Text Analytics Gateway hopes to be transformational in this regard. The goal of the project is to provide researchers with a user-friendly text analytics workflow system. Unlike other web-based text mining portals in these disciplines such as TaPoR and HTRC, the XSEDE Text Analytics Gateway will not merely present users with a list of data retrieval operations. Instead, this project implements interactive documentation to walk users through the decisions needed to combine data retrieval, preprocessing, analysis, and postprocessing into a single, complete workflow. In this respect, the gateway is both a research and pedagogical tool.

The initial audience for the gateway will be researchers with little background in computational methods; however, a simpler interface will also be available for those with more experience in text analytics who would like to design their own workflows independent of the examples provided by the interactive documentation. Ideally, these more experienced researchers will be able to use this gateway as a place to test and design methodologies in pursuit of larger projects. Sample data will be provided for novice users, and the team is currently exploring the possibility of integrating access to web APIs from relevant data sources for more advanced users.

The core of the gateways text analytics functionality will be implemented in the popular programming language and analysis package R. The interface is being designed with shiny, an R package that connects R to the web and allows for the easy creation of interactive webapps. The "interactive documentation" will be created using RMarkdown, which is an enhancement to the simple document system markdown. Elements of the gateway will be embedded directly into these markdown documents.

The XSEDE Text Analytics Gateway is currently in the prototyping stage, with basic preprocessing and analytics implemented in the non-documentation driven interface. Currently, several example case study workflows have been drafted for implemented in Markdown. The immediate next steps will be to implement one or two sets of interactive documentation, move the framework onto the new XSEDE resource Comet, and begin small scale user testing with a sample data set. In this poster, we will share the early developments of the project, including technical issues and community building challenges. The poster should appeal to practitioners and gateway developers alike.

## References

[1] Allaire, J., J. Cheng, Y. Xie, J. McPherson, W. Chang, J. Allen, H. Wickham, and R. Hyndman (2015). *rmarkdown: Dynamic Documents for R*. R package version 0.5.1.

[2] Chang, W., J. Cheng, J. Allaire, Y. Xie, and J. McPherson (2015). *shiny: Web Application Framework for R*. R package version 0.11.1.

[3] R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.