Introducing TAG: The Text Analytics Gateway

What is TAG?

- A "gateway".
- Technological service primarily for HASS researchers.

Humanities, Arts, and Social Sciences

- University departments, libraries, labs, ...
- No access to technical training.
- Lack of institutional rewards.
- Lots of text!

Computing in HASS

What they're used to

- GUI/Web service.
- Generally like local installs.
- Interactive.

Programming

- Not in their wheelhouse.
- So why expect users to start with this?

```
corpus <- tm::tm_map(corpus, tm::content_transformer(base::tolower))
corpus <- tm::tm_map(corpus, tm::removePunctuation)
corpus <- tm::tm_map(corpus, tm::removeNumbers)
corpus <- tm::tm_map(corpus, tm::stripWhitespace)
corpus <- tm::tm_map(corpus, tm::removeWords, tm::stopwords("english"))
tdm <- tm::TermDocumentMatrix(corpus)
wordcount_table <- sort(rowSums(as.matrix(tdm)), decreasing=TRUE)
```

TAG: Text Analytics Gateway

The TAG Team

- PI Mike Black
- Drew Schmidt
- New ECSS Staff soon...

Timeline

- November 12, 2014: XSEDE Allocation awarded.
- March 20, 2015: Developent begins.
- July 7, 2015: Stood up on AWS, small usability studies begin.
- Soon: Live on comet, workshops, ...

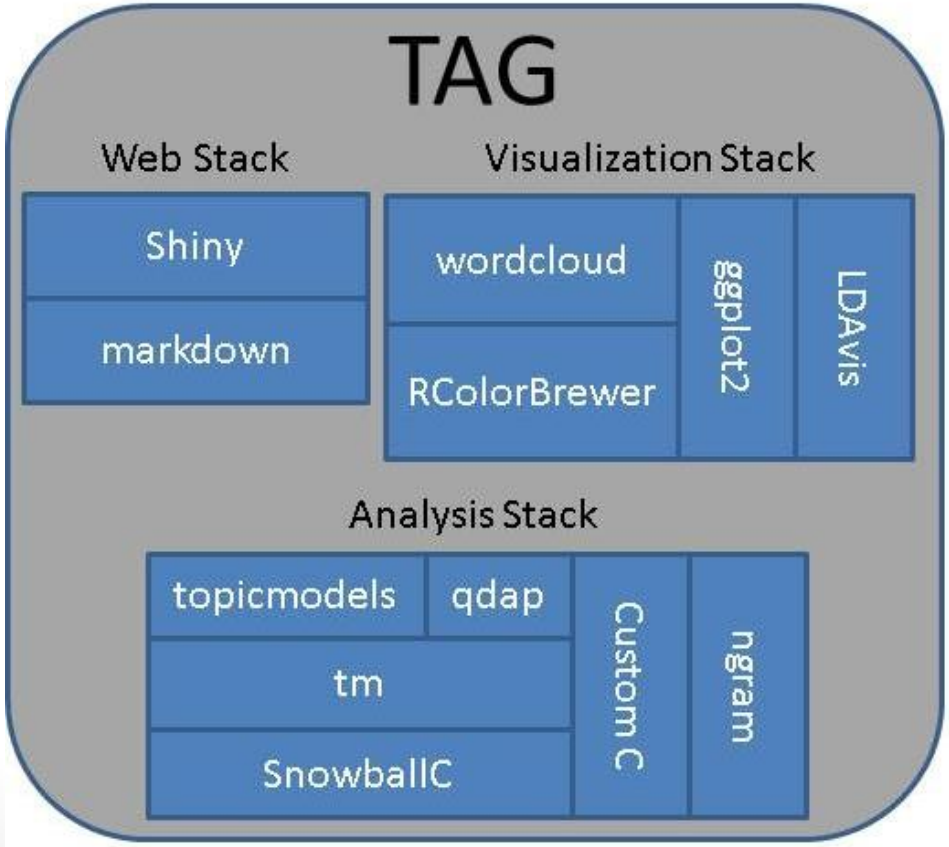
XSEDE

- eXtreme Science and Engineering Discovery Environment
- NSF cyberinfrastructure organization
- "supercomputing"
- Growing interest in data.

The TAG Philosophy

- Easy to use.
- Reproducibility should not be an afterthought.
- Interactivity is good!
- Community contributions, comments, suggestions should be encouraged.
- We're not chasing exascale.

TAG Design



- Written mostly in R.
- Some custom high-performance C.
- Extensively uses existing R packages.
- UI served via shiny.
- Help files written in Markdown.

TAG Installation

- Easy to install and run locally.
- Relatively easy to stand up in a vm.
- Installation and updating:

  R> devtools::install_github('XsedeScienceGateways/TAG')
- Starting:

  TAG::runTAG()
- Deployment script for Ubuntu:
  https://raw.githubusercontent.com/XSEDEScienceGateways/TAG/master/inst/deploy.sh

Challenges and Future Work

- Integration within XSEDE.
- Add multidocument support.
- Better I/O.
- Add sentiment analysis.
- Develop userbase and community.
- Legal issues?
- Sustainability.
- Proper UI/UX.

Community Building

- Encourage usage and contributions.
- GitHub: https://github.com/XSEDEScienceGateways/TAG
- Help written in Markdown.
- Plans to include reporting system that doesn't require GitHub account.

Thanks

- Questions?