XSEDE Text Analytics Gateway

Mike Black[1] and Drew Schmidt[2]
[1]University of Illinois, [2]University of Tennessee

Background

- Humanities and social sciences historically dominated by qualitative methods.
- Researchers in these fields increasingly turning to quantitative methods and computational resources to answer larger questions.
- Still much resistance, often driven by a "technological familiarity gap".

So why expect users to start with this?

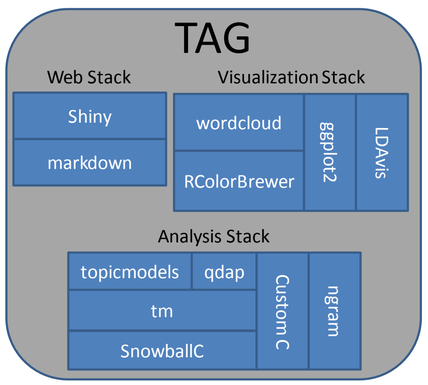
```
corpus <- tm::tm_map(corpus, tm::content_transformer(base::tolower))
corpus <- tm::tm_map(corpus, tm::removePunctuation)
corpus <- tm::tm_map(corpus, tm::removeNumbers)
corpus <- tm::tm_map(corpus, tm::stripWhitespace)
corpus <- tm::tm_map(corpus, tm::removeWords,
    tm::stopwords(input$data_filter_stopwords_lang))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

Programming is great! But...

- Users often lack the background...
- Or the desire! They just want to get to their science.
- Programming vs GUI: Recall vs. recognition.

TAG: The Text Analytics Gateway

Basic Text Analysis Tools Without the Programming



- Written mostly in R.
- Extensive re-use of existing R packages.
- UI served via shiny [1] and shiny server.
- Some custom high-performance C.
- Help files written in markdown, rendered on the fly.

Offer full analysis pipeline, by combining:

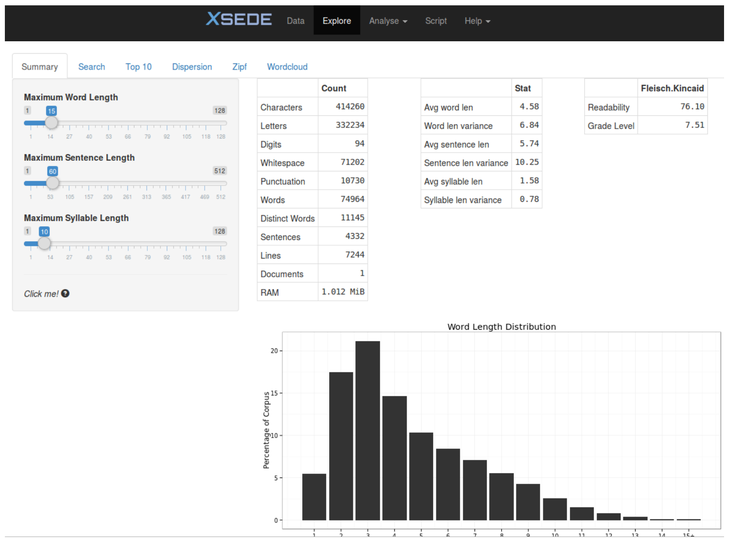
- Data retrieval
- Preprocessing

into a single, complete workflow.

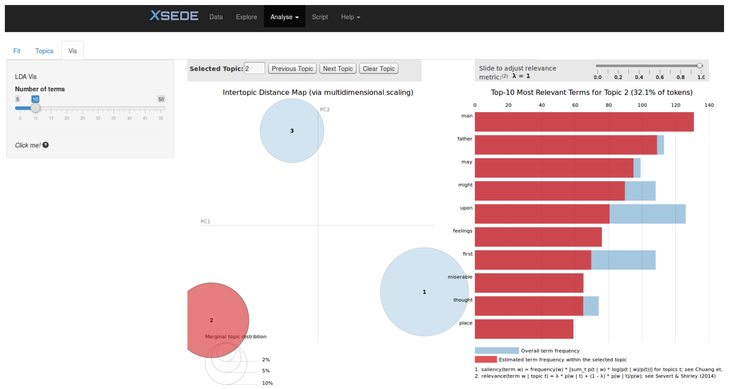
- Data retrieval
- Preprocessing

TAG Features

Detailed summary statistics



Includes LDAvis [2] for powerful topic modeling visualization:



And more...

- **Processing:** Text filtering and transformations.
- **Vis:** Dispersion plots and wordclouds.
- **Analysis:** LDA and n-gram modeling.
- **Reproducibility:** State management, script generation, 100% open source.

Timeline and Status

- November 12, 2014: Allocation awarded.
- April 6, 2015: PI connected to ECSS staff.
- April 6—March 15, 2015: Research technology stack.
- Mar 20, 2015: Development begins.
- July 7, 2015: Gateway stood up on AWS. Usability studies begin.
- Soon: live on Comet, tutorials, workshops...

: Coming soon^{TM} to Comet!

: Try now on our AWS hosting!

Future Work

- Stand up gateway on Comet.
- Incorporate multidocument support.
- Add sentiment analysis feature.
- Develop userbase and community.

References

[1] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2015). shiny: Web Application Framework for R. R package version 0.12.1.9000. http://shiny.rstudio.com

[2] Carson Sievert and Kenny Shirley. LDAvis: Interactive Visualization of Topic Models. R package version 0.3.1. https://github.com/cpsievert/LDAvis

Acknowledgements

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575 and the XSEDE Science Gateways Program.