



XSEDE Text Analytics Gateway

Mike Black¹ and Drew Schmidt²
¹University of Illinois, ²University 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?

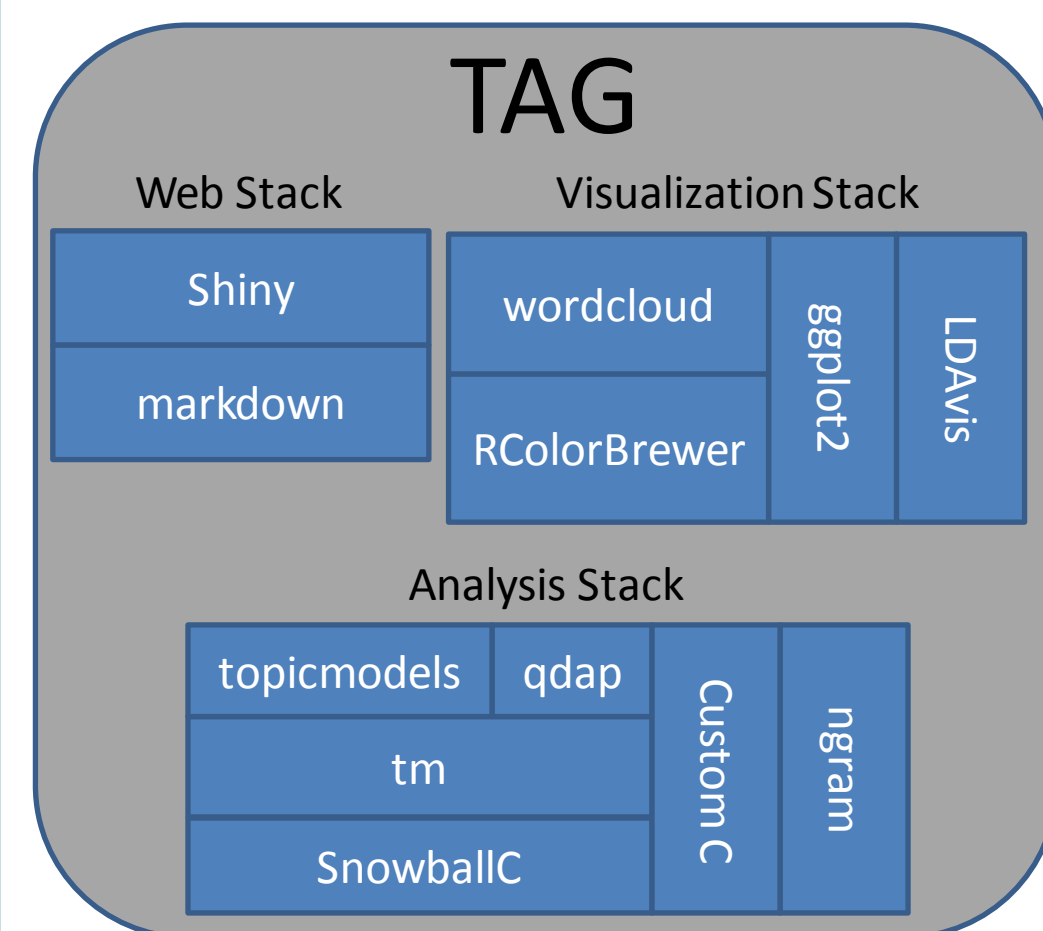
```
1 corpus <- tm::tm_map(corpus, tm::content_transformer(base::tolower))
2 corpus <- tm::tm_map(corpus, tm::removePunctuation)
3 corpus <- tm::tm_map(corpus, tm::removeNumbers)
4 corpus <- tm::tm_map(corpus, tm::stripWhitespace)
5 corpus <- tm::tm_map(corpus, tm::removeWords,
6   tm::stopwords(input$data_filter_stopwords_lang))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

Programming is great! But...

- ▶ Users often lack the background and have little access to training.
- ▶ No institutional rewards for developing technical skills.
- ▶ Programming vs GUI: recall vs. recognition.

TAG: The Text Analytics Gateway

Basic Text Analysis Tools Without the Programming



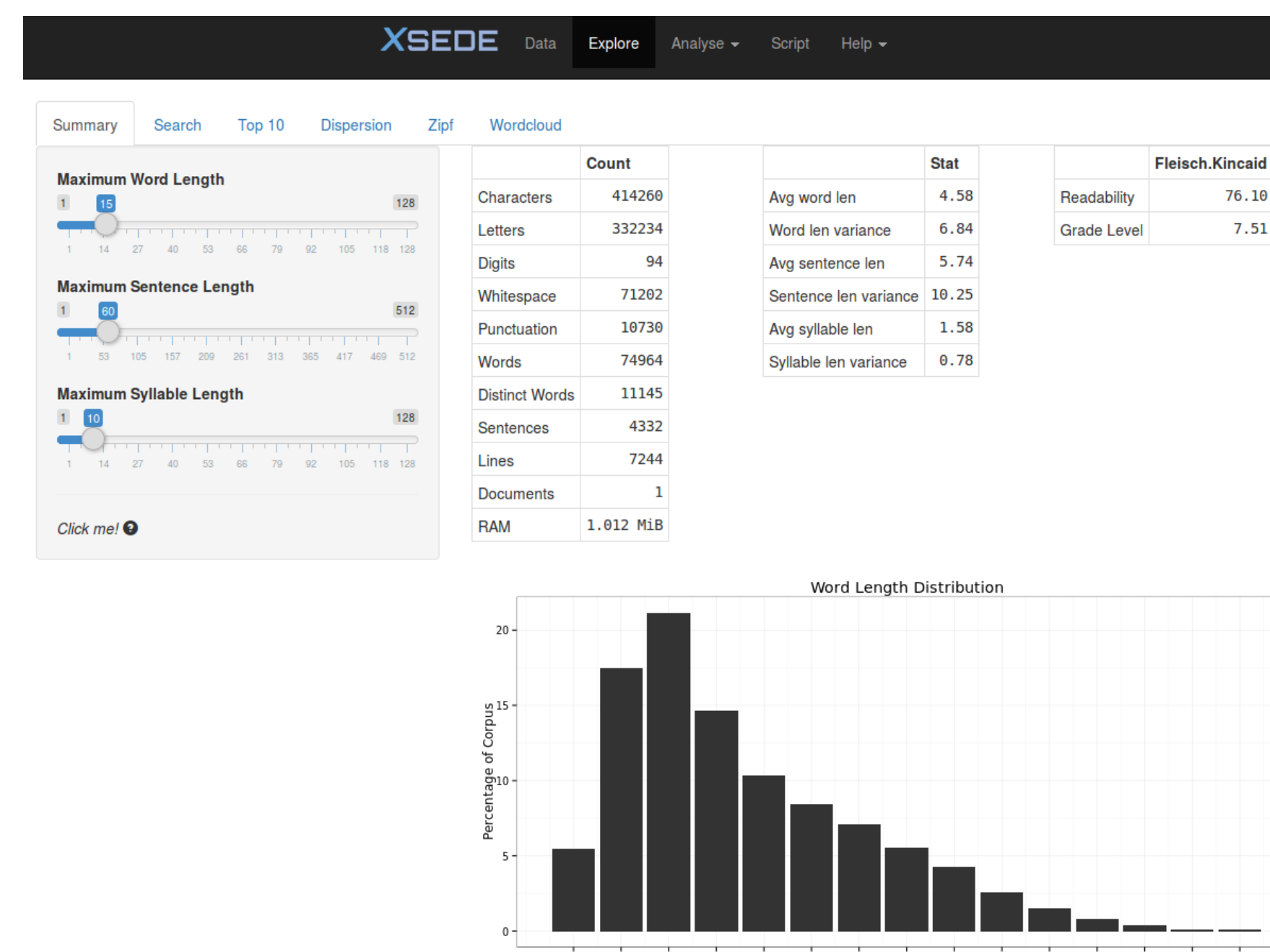
- ▶ Written mostly in R.
- ▶ Extensive re-use of existing R packages.
- ▶ UI served via shiny [1] and shiny server.
- ▶ Some custom high-performance C.
- ▶ Help files written in markdown, rendered on the fly.

Offer full analysis pipeline, by combining:

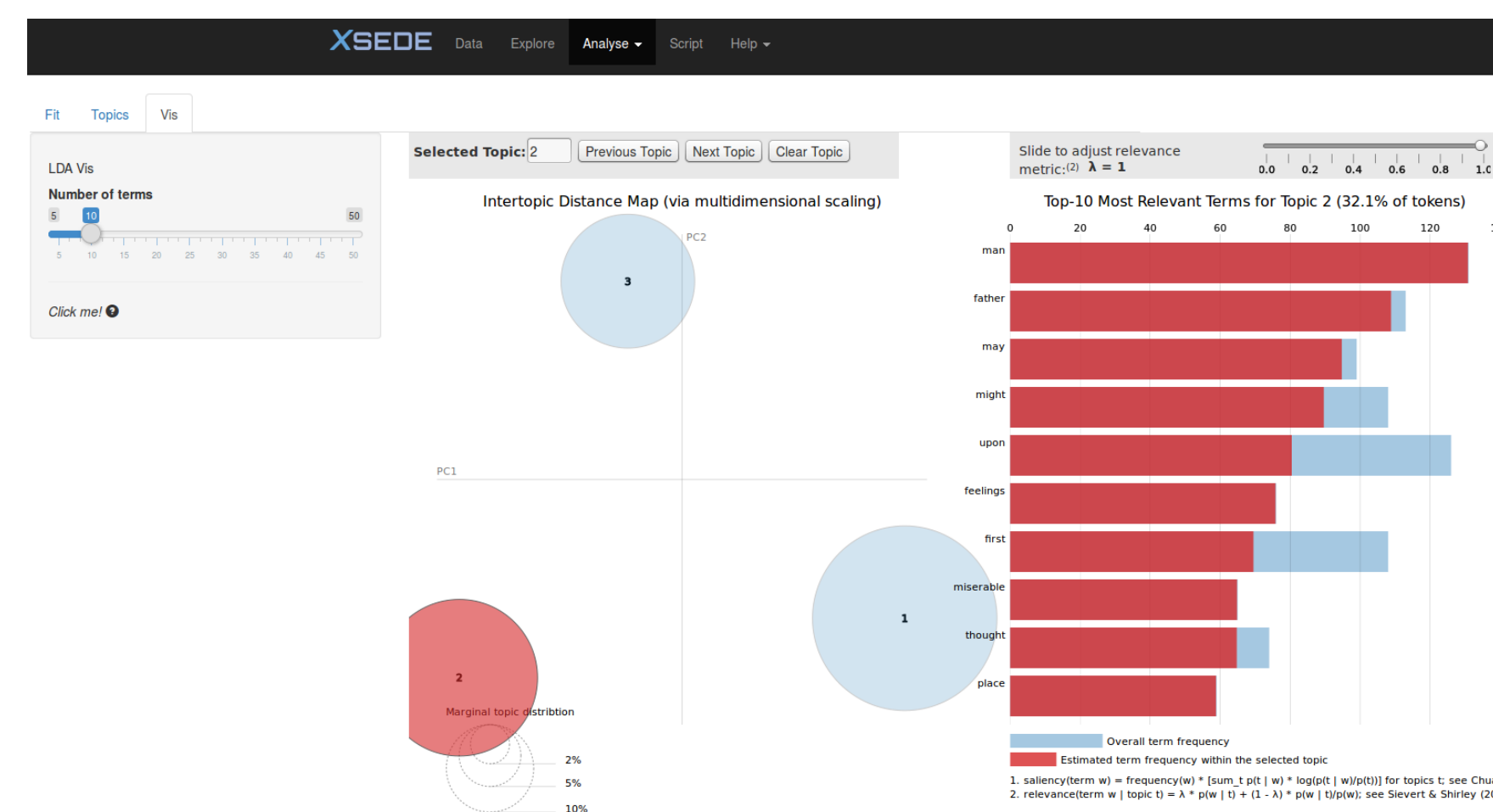
- ▶ Data retrieval
 - ▶ Preprocessing
 - ▶ Data retrieval
 - ▶ Preprocessing
- into a single, complete workflow.

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.



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™ 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.