Web-Based Document Visualization with DocuBurst

Bradley Chicoine ■ Christopher Collins ■ Visualization for Information Analysis Lab ■ Faculty of Science, UOIT ■ Ontario, Canada

Research Goals

Create a web-based implementation of DocuBurst which will support:

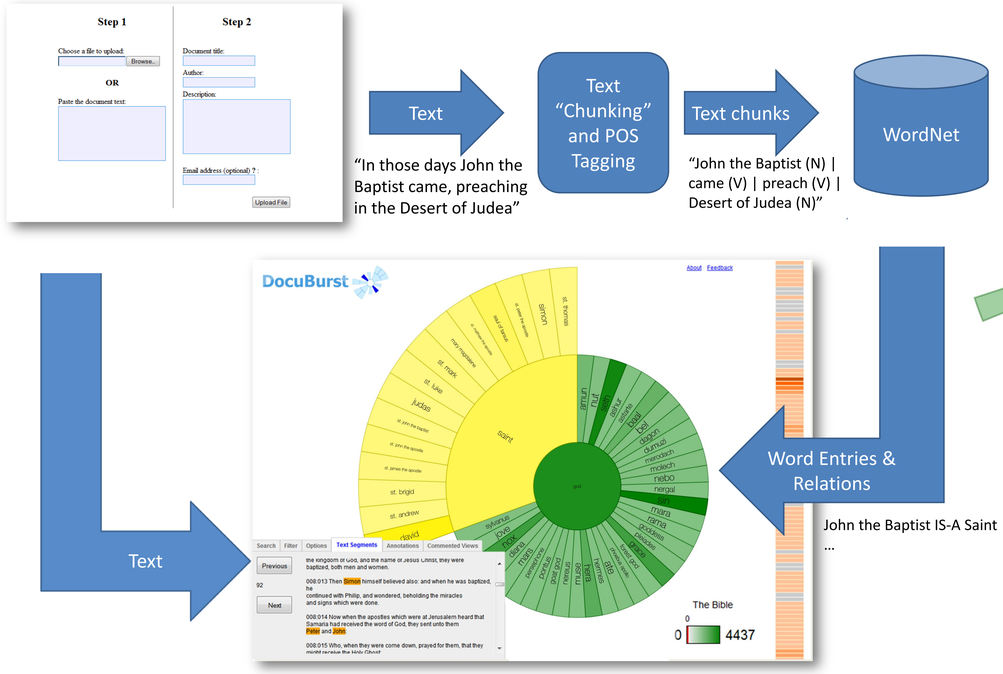
- Quick perusal of a text for key concepts
- Exploration of patterns of concept usage
- Comparisons between multiple documents
- Ability to explore different word scoring methods
- Usage logging and user feedback to evaluate the visualization

DocuBurst Visualization

DocuBurst is a visualization [1] which represents the hierarchal structure of the relationships between English nouns under the IS-A relationship (*table* IS-A *furniture*). The most general concept occurs at the center, with more specific concepts towards the edges. Word counts are used to colour single nodes or can be propagated to the center, aggregating related concepts into more general themes.

Text Uploading and Processing

Anyone may upload data to be visualized using DocuBurst. Visitors can upload a text file for processing, or simply paste the text directly into a box on the upload page. Users also enter meta-data for the document such as the title, author, and description. The processing is complete and the visualization is ready in about 10 seconds.

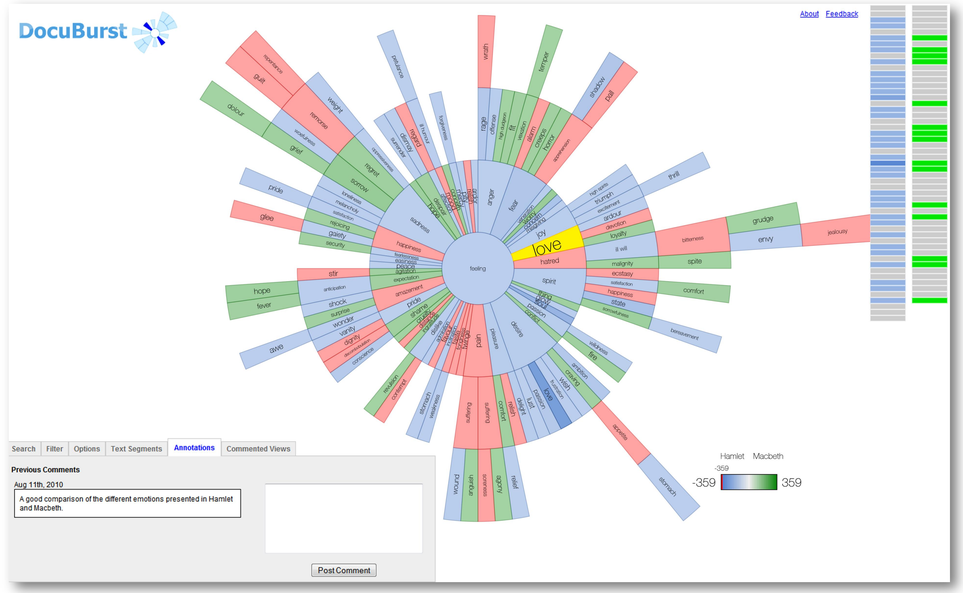


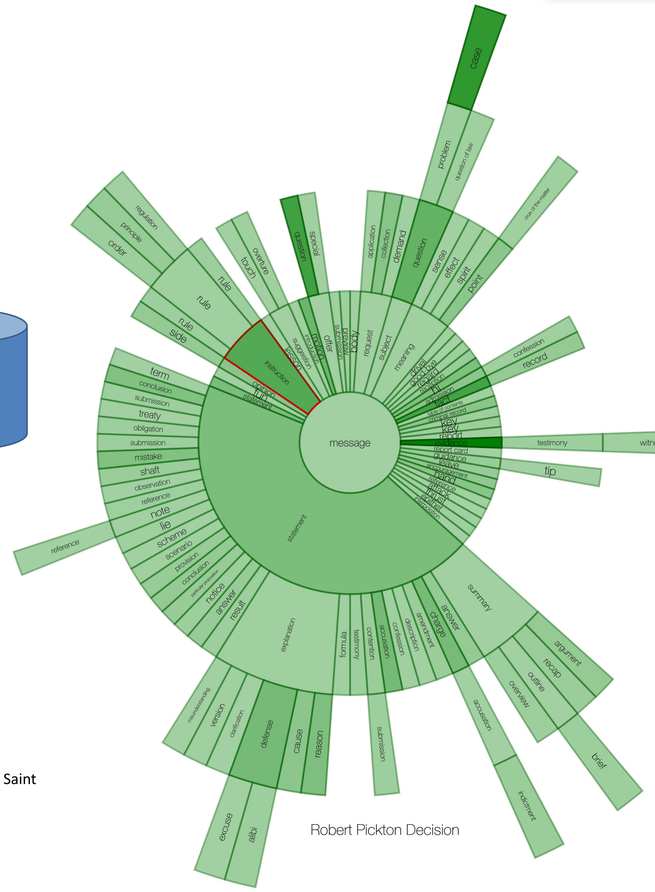
Document Comparison

Documents are compared by taking the difference in word scores for two documents. The graphic then displays three hues based on the resulting node differences: green, blue and red representing positive, negative and equal values respectively.

Collaborative Analysis

DocuBurst views can be bookmarked and annotated. Pages that contain annotations are listed both in the search tab of the DocuBurst homepage and in the Commented Views tab for the DocuBurst visualization page for quick reference and sharing.





Rich Interaction

DocuBurst-web is fully interactive, implemented with the ProtoVis [2] toolkit. Areas of interest on the visualization can be enlarged using the mouse wheel. The nodes in DocuBurst support single and multiple selections (in yellow) which present information such as definitions and synonyms. The text browser on the right side of the display reveals the paragraphs in which the selected concepts are more prevalent. Clicking a text segment shows the term usage in the context of the original text.

Application Domains

- Replace traditional keywords for e-book and online articles
- Legal applications: assist in cataloguing for e-discovery
- Humanities research: authorship analysis, plagiarism detection
- Business analytics: explore company repositories, patent collections, emails, customer feedback

Ongoing Research

Future work includes public deployment to analyze usage logs as well as elicit feedback. Additionally, we plan to tune the text processing program to improve word sense disambiguation between multiple senses of a word, and to explore variations on the current word scoring method, for example, to give additional weight to words that are rare in a reference corpus.

[1] Collins, Christopher; Carpendale, Sheelagh; and Penn, Gerald. "DocuBurst: Visualizing Document Content using Language Structure." Computer Graphics Forum, 28(3): pp. 1039-1046, June, 2009.
[2] Bostock, Michael and Heer, Jeffrey. "Protovis: A Graphical Toolkit for Visualization," IEEE Transactions on Visualization and Computer Graphics, pp. 1121-1128, November/December, 2009

http://vialab.science.uoit.ca/docuburst ■ christopher.collins@uoit.ca

UOIT CHALLENGE INNOVATE CONNECT
NSERC CRSNG
vialab visualization for information analysis