2011

Interactive Data Visualization Tool to Analyze Word Count Frequencies Over Time

Aryn Grause
University of Puget Sound, agrause@pugetsound.edu

Interactive Data Visualization Tool to Analyze Word Count Frequencies Over Time

Aryn Grause

Introduction

Data visualization is a visual interpretation of data which can be used when analyzing data collections. It can be challenging to build complex visualizations that are meaningful yet accurately display the data. Current available tools are limited into the types of datasets they can be applied. I built the Frequency Over Time Interactive Visualization (FOTIV) tool that creates an interactive data visualization of word frequencies from electronic text documents. FOTIV is the first tool available that uses interactive visualization to assist in the analysis of word usage frequencies.

Data Visualization

Simple examples of data visualizations include tables and graphs. The fundamentals of data visualizations have been formally established overtime by researchers in the field. There are currently two widely accepted paradigms on how data visualizations should be organized: The first indicates that the visualizations should contain copious amounts of information, to the extent that viewers have to seek for specific facts [4]. The second paradigm emphasizes simplicity to exaggerate facts in the data [2]. Both styles of formatting attempt to reduce user errors and shorten time required to analyze the data. Interactive data visualization may have the capabilities of merging these two approaches. Interaction allows the viewer to dynamically alter the information being displayed by the visualization [1]. This interaction enhances the viewer's ability to understand trends and discover anomalies within the data set. FOTIV allows the user to access more information by rebuilding a series of simple graphs to the user's specifications.

Document Analysis

FOTIV is comprised of three distinct components. It first conducts an analysis of the inputted text documents. FOTIV begins the analysis stage by distributing the text document to a Java library. The Java library simplifies the document (e.g. removes punctuation). Once the document is rendered, FOTIV parses the text into individual words. This library then calculates the frequencies of each individual word used in the text.

Fig.1

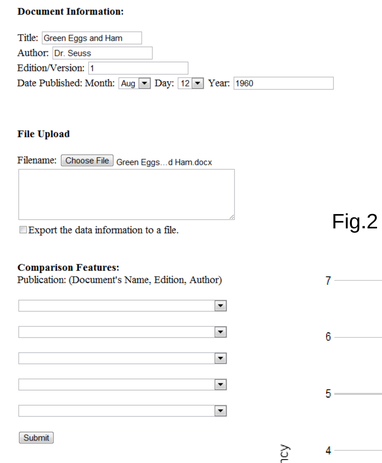
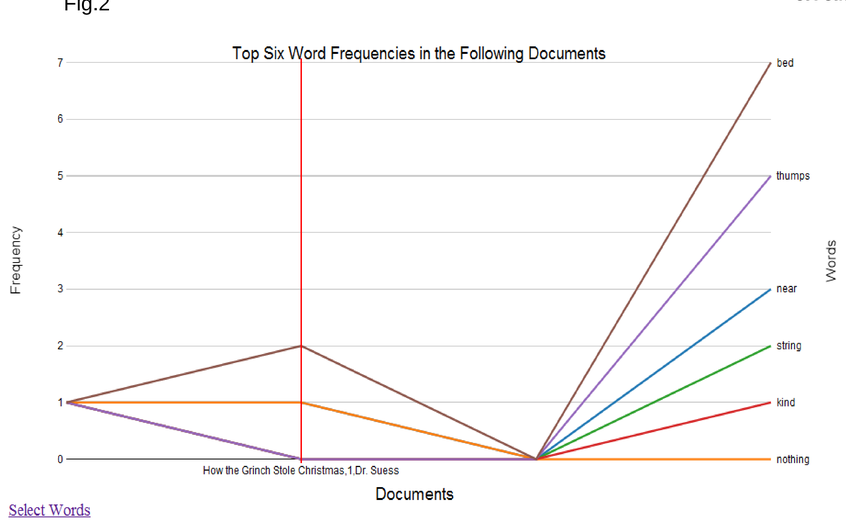


Fig.2



The Database

To aid in flexibility and speed, FOTIV utilizes a SQL Database. All results generated by the Java program are stored in the database creating a persistent repository. This allows FOTIV to create multiple different visualizations of a document. The database allows the user to input multiple editions of documents. Another key feature of the database is to allow users to create a second repository of common words or specific words the user does not want the visualization to display.

The Visualization

Using the web interface, FOTIV accesses the database and retrieves the frequencies for the documents that the user requested. Using the program, Protovis [3], FOTIV builds a visualization of the six most frequently used words in the document added by the viewer and shows their frequencies compared to the other documents. Then FOTIV allows the user to select from all of the words used. FOTIV will then render a graph in real time that contains the selected words. The interaction occurs when the user gets to continually select words of interest while FOTIV continues to create visualizations. This allows the user to narrow down the data they are viewing while having all the data available.

Fig.3

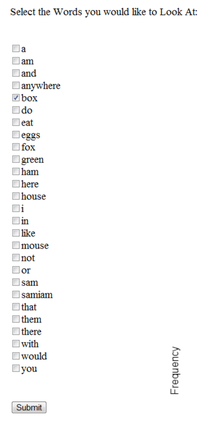
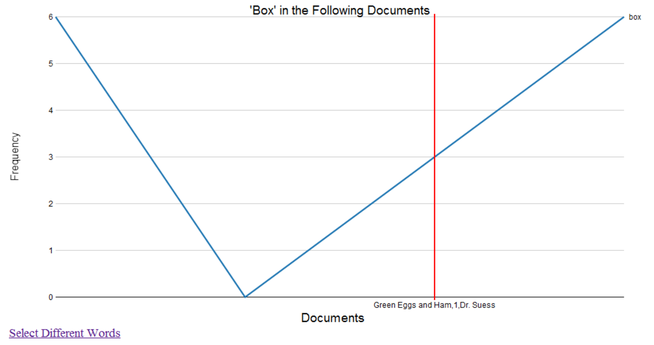


Fig.4



Future Work

In the future, I plan to add more functionality to the program, such as: the ability to remove words, tense changes, more detailed graphs, ect. I also plan to conduct a human subjects test to determine if FOTIV reduces errors or the amount of time for the subjects to interpret the data.

Acknowledgements

This research was made possible by the University of Puget Sound Summer Scholar Grant and aided by my advisor Jason Sawin, my five Lab Mates, Amazon Web Services Cloud Computing Grant.

References

[1] H. Akiba and K-L. Ma. An An Interactive Interface for Visualizing Time-varying multivariate Volume Data, *Proc*. *EuroVis* 2007.
[2] W. Cleveland. The Elements of Graphing Data. Monterey: Wadsworth Advance Books and Software, 1985.
[3] *Protovis*. Stanford University, 2010. Web. 27 January 2011.
[4] E. Tufte. Envisioning Information. Cheshire: Graphics Press, 1998.

Fig.5

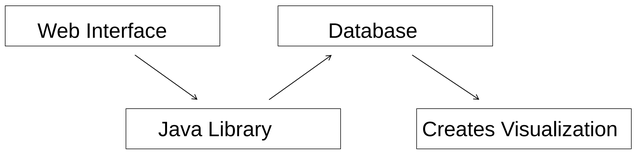


Fig. 5 FOTIV collects Dr. Suess' "Green Eggs and Ham" through the web interface and then passes the document to the Java library where the words are parsed and counted. The word frequencies are stored into the Database. The web interface then requests the frequency counts for "Green Eggs and Ham", "How the Grinch Stole Christmas", "The Cat in the Hat", and "One Fish, Two Fish, Red Fish, Blue Fish". The visualization is then displayed for the viewer to analyze.

Fig. 1 Is the web interface that allows the user to input a new text document and select other text documents to compare to. The user inserted "Green Eggs and Ham" by Dr. Suess and wants to compare it to three of Dr. Suess' other works: How the Grinch Stole Christmas", "The Cat in the Hat", and "One Fish, Two Fish, Red Fish, Blue Fish".

Fig. 2 This is the first visualization FOTIV draws. As the viewer's mouse floats over the visualization the red line will appear to show the points for that document. The six most frequently used words in "Green Eggs and Ham" are: near, nothing, string, kind, thumps, bed. We can notice that "How the Grinch Stole Christmas" uses bed more frequently.

Fig. 3 This page allows the viewer to specify which words in "Green Eggs and Ham" they would like to compare across the documents. The viewer has selected 'bed'.

Fig. 4 This is the visualization FOTIV has created for the viewer based on their specifications of 'bed'. The red line represents the current document the viewer is looking at, in this case: "Green Eggs and Ham". The viewer can move the mouse to look at the others.