Interactive Data Visualization Tool to Analyze Word Count Frequencies Over Time

Aryn Grause

University of Puget Sound, agrause@pugetsound.edu

Follow this and additional works at: http://soundideas.pugetsound.edu/summer_research

Part of the Graphics and Human Computer Interfaces Commons

Recommended Citation


http://soundideas.pugetsound.edu/summer_research/69

This Presentation is brought to you for free and open access by Sound Ideas. It has been accepted for inclusion in Summer Research by an authorized administrator of Sound Ideas. For more information, please contact soundideas@pugetsound.edu.
Interactive Data Visualization Tool to Analyze Word Count Frequencies Over Time

Aryn Grause

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.

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.

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.

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 visualization 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.

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, etc. 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