Data Visualization Professor Harrison
The Word Whales (Elaine Chen, Jules Cazaubiel, Nicholas Tourtillott, Johvanni Perez)3/30/2022

Final Project Prospectus

Abstract/Introduction:

Word clouds are a visualization strategy used to represent data that can be pulled from articles, websites, or any kind of input data. The purpose of this study is to analyze which elements of a word cloud contribute to its effectiveness. Due to previous studies, we are able to look into completed experiments and test elements like shape interpolation, color, size, and bar length. Studies have already been completed across various age demographics, ranging from elementary school children to college students. We are looking to test a random group of college students, and have them read a short excerpt. Then, they will be presented a variety of word clouds with different characteristics that are made from the except, and rank which word clouds are the most effective. We are looking to find the most impactful element of a word cloud by finding the characteristic that gets ranked the highest based on our results.

While there are a multitude of visualization methods available to represent numeric data that help create an effective narrative, there are limited methods that are effective in portraying relevant written information. Word clouds are a method that are variable enough to be used across all age groups, easy to understand, and quick to decipher. However, they are a controversial concept in the data visualization community, and it is worthwhile to explore methods for visualizing textual content as opposed to numerical information. After finding a variety of studies that look into testing the effectiveness of word clouds across different types of audiences, it provides a backing that they are able to be easily understood, and have an impact on audiences in some regard. Other studies look into methods that make word clouds more or less comprehensive, and our study is planning on examining these methods to draw our own conclusions. We are hoping to collect quantitative data that enables us to conclude on the most effective traits of a word cloud and identify why they are helpful. With the rise of data visualization growing rapidly, it is important to establish the most effective and fastest methods of portraying all types of information, such as the text that is portrayed in word clouds.

One-sentence Description

Participants of this study will be given one document and then will be asked to rank five word clouds, each one with a different design trait, to determine which trait is the most effective for comprehension.

Project Type:

Experiment utilizing quantitative surveys

Audience

The results of this experiment will have important implications for many groups. Most directly, this project will affect researchers who focus their work on document summarization and data visualization, as it will ideally lay the groundwork for the best practices when generating word clouds. It will highlight a currently underutilized method of visualizing textual data and provide context for which aspects of these visualizations are the most important for the efficacy.

Additionally, this project will have practical applications for those less concerned with the theory of data visualization and more concerned with actual uses of document summarization. As mentioned before, we aim to provide best practices to generate effective word clouds so those who need methods of document summarization for their work, regardless of the field, should likely take note. Honestly the uses could range from a variety of fields like: education, law, news/journalism, and any other sector which involves the spread and consumption of a mass amount of textual documents and information.

Approach Details

We will begin by selecting a text to create our word cloud, and that people taking the questionnaire will have to read. In our selection, we will have to keep in mind the length and approachability of the text. Indeed, a text that is too long or too complicated to read is likely going to discourage people from participating in the study. Then, we will need to identify which aspects of the word cloud will change before creating our word cloud variations. Finally, after enough people participate in our experiment, we will analyze and visualize the results to identify which word clouds were rated the highest, and which feature contributed the most to high rankings.

Evidence for Success

A previous experiment involving a similar approach in Assignment 3 highlighted the need for the texts read by people to be short, in order to not discourage them from taking the survey. Additionally, a previous study that largely inspired this project involved changing aspects of word clouds to identify their importance in conveying information (Cristian, 2017). This approach was successful and partially used to guide our own.

Best-case Impact Statement

In the best case scenario, the impact statement of our project would allow us to recommend the best practices for word cloud generation. It would also allow us to point to the most important attributes within a word cloud in regards to document summarization.

Major Milestones

1. Create the questionnaire

    a. Select the text
    b. Create the word cloud variants
2. Gather sufficient responses
3. Analyze and present the results in a clear manner

Obstacles

Major obstacles

- Our biggest obstacle is getting enough people to participate in our survey to generate an adequate amount of data. To be direct about it, most individuals do not care enough to participate nor do they have the time to sit down and take a long survey. A lack of responses will significantly hinder the conclusions we can draw from our data, so finding a way to combat this obstacle early will be crucial to our project.

Minor obstacles

- We must establish what we assume is the baseline method of generating a word cloud. The way our survey is structured we will need this baseline word cloud to then manipulate other versions of said word cloud by tweaking different elements. Essentially, this baseline would then be modified in coloring in one version, word size in another, shape in another, etc. However, the way in which we establish a baseline word cloud could alter our results by affecting the resulting children of that word cloud so to speak.

Resources Needed

- A text processing tool to gather and process text from the selected documents
- A D3 script to generate the word clouds
- Qualtrics to create the questionnaire for our survey
- Python to help analyze and manipulate the data collected from the survey and to create data visualizations.

5 Related Publications

Aldalalah provides a study with a similar purpose of ours, by trying to find the effectiveness of word clouds made by college students. While the purpose of the word clouds in this study differs from our intended purposes, the study looks into the qualities that make an effective word cloud and explores what these traits are and how capable the students are at making them.

Hearst et al.'s study explores the grouping of word clouds and their effectiveness. This allows our study to grow off of their findings and identify which qualities we should be testing in our experiment.

The shape of word clouds was originally proposed to be observed. Because this study provides a strong argument and methodology to understand and test the

effectiveness of how shape interpolation influences the effectiveness of word clouds, it will allow this experiment to build off of Chi et al.'s results.

The main usage of word clouds can be very versatile, but an interesting target audience for them could be young children's education. This study looks at the effectiveness of them on elementary school children, and identifies them as a tool in their primary education. Based on this information it will be interesting to see how the results differ from a pool of college students. It will be interesting to see if results differ if they are too easy to identify, or if there are useful qualities that would benefit all levels of education.

Creating word clouds relies on different relationships between design and network science, which Rodighiero and Romele explore. They look into how alternative paths led to exploring innovative ways to present network diagrams. It discusses adding contour lines and how it might impact word clouds, which we can explore to see their effectiveness in our study.

Define Success

This project will be considered a success if we are able to determine which element of a word cloud is best for comprehending the contents of text documents. While we cannot say which element we believe to be the most significant, we hope to have a substantial amount of data to demonstrate the best element. We will define the best element as the factor that facilitates the participants' ability to identify the correct documents provided and has the best success rate.

References

Aldalalah, O. M. A. A. (2022). Employment the Word Cloud in Brainstorming via the Web and Its Effectiveness in Developing the Design Thinking Skill. International Journal of Instruction, 15(1).

Cristian Felix, Steven Franconeri, Enrico Bertini (2017). Taking Word Clouds Apart: An Empirical Investigation of the Design Space for Keyword Summaries. IEEE Transactions on Visualization and Computer Graphics (Proc. of InfoVis).

Hearst, M. A., Pedersen, E., Patil, L., Lee, E., Laskowski, P., & Franconeri, S. (2019). An evaluation of semantically grouped word cloud designs. IEEE transactions on visualization and computer graphics, 26(9), 2748-2761.

Nezhyva, L. L., Palamar, S. P., & Marienko, M. V. (2022). Clouds of words as a didactic tool in literary education of primary school children. In CEUR Workshop Proceedings (pp. 381-393).

Rodighiero, D., & Romele, A. (2022). Reading Network Diagrams by Using Contour Lines and Word Clouds.

Saranya, M. S., & Geetha, P. (2020, July). Word Cloud Generation on Clothing Reviews using Topic Model. In 2020 International Conference on Communication and Signal Processing (ICCSP) (pp. 0177-0180). IEEE.