A Survey of Data and Encodings in Word Clouds

Muyang Shi[1], Danielle Albers Szafir[2], Eric Alexander[1]
[1]Carleton College , [2]University of Colorado Boulder

Background & Aim

Word clouds are an increasingly popular means of presenting statistical summaries of document collections, appearing frequently in digital humanities literature, newspaper articles, and social media. Word clouds typically use font size to illustrate the frequency of word use: commonly-used words are larger and are therefore more likely to draw a reader's attention. Despite their ubiquity and intuitive appeal, our ability to read such visualizations accurately is not yet fully understood. Some past work has shown that compared to a visualization of a traditional word list, people do poorly at lookup and recall tasks in word clouds; the shape of a word as well as placement of the word in the word cloud might also bias the perception of the word. [1,2]

We conducted a survey on the word cloud usage in digital humanities contexts compared to the New York Times articles. Specifically, we focused on how people are using word cloud visualizations differently in terms of **the intended task** for the word clouds (to provide summary/presentation or to attract attention), the **source of the corpus** of the word clouds (whether it is generated from a single or multiple documents, whether the corpus comes from topic modeling, and how the data sources are shown), and the **visual design and features** of the word clouds (interactive or static design, as well as the data encoding used such as color and word positioning besides the traditional font size).

Main Take-Away

Word clouds in digital humanities contexts are more often used to visualize the various topics contained in multiple documents whose sources are more likely to be explicitly and concisely laid out compared to word clouds being used in the New York Times. They are used to illustrate data, rather than a simple colorful "eye-grabber" only for engagement purposes, which is the case in many of the New York Times articles. In this survey, we found that word clouds used by DH researches are often contained in integrative data visualization tools (existing ones or proposed by the authors); there are usually more than one type of data encoding used to build the word cloud in such tools. The word clouds are being used in an intermediate stage of data processing, and some other data encodings such as color and word positioning are often used to enrich and strengthen the data encoded in the word clouds.

As a conclusion, word cloud's audience and authors should pay attention to the above-mentioned aspects of word clouds, especially **the use of various visual features that enrich the information contained in the word clouds**. Because we see from the survey of people's needs from these word clouds, designers and developers of word cloud tools may consider to **develop generic tools to incorporate topic models for word clouds on multiple documents, adopt interactive designs, and enable users to specify various kinds of visual features on word clouds**. From the survey, we have seen that people have been trying to improve word clouds in different ways to better visualize the data; however, the effectiveness of these designs and features haven't been fully understood. For example, how color and word positioning can be used together to make a better readability of a word cloud. We encourage future researchers to **study the efficacy of the different designs and visual features.**

Acknowledgements, References, & Contacts

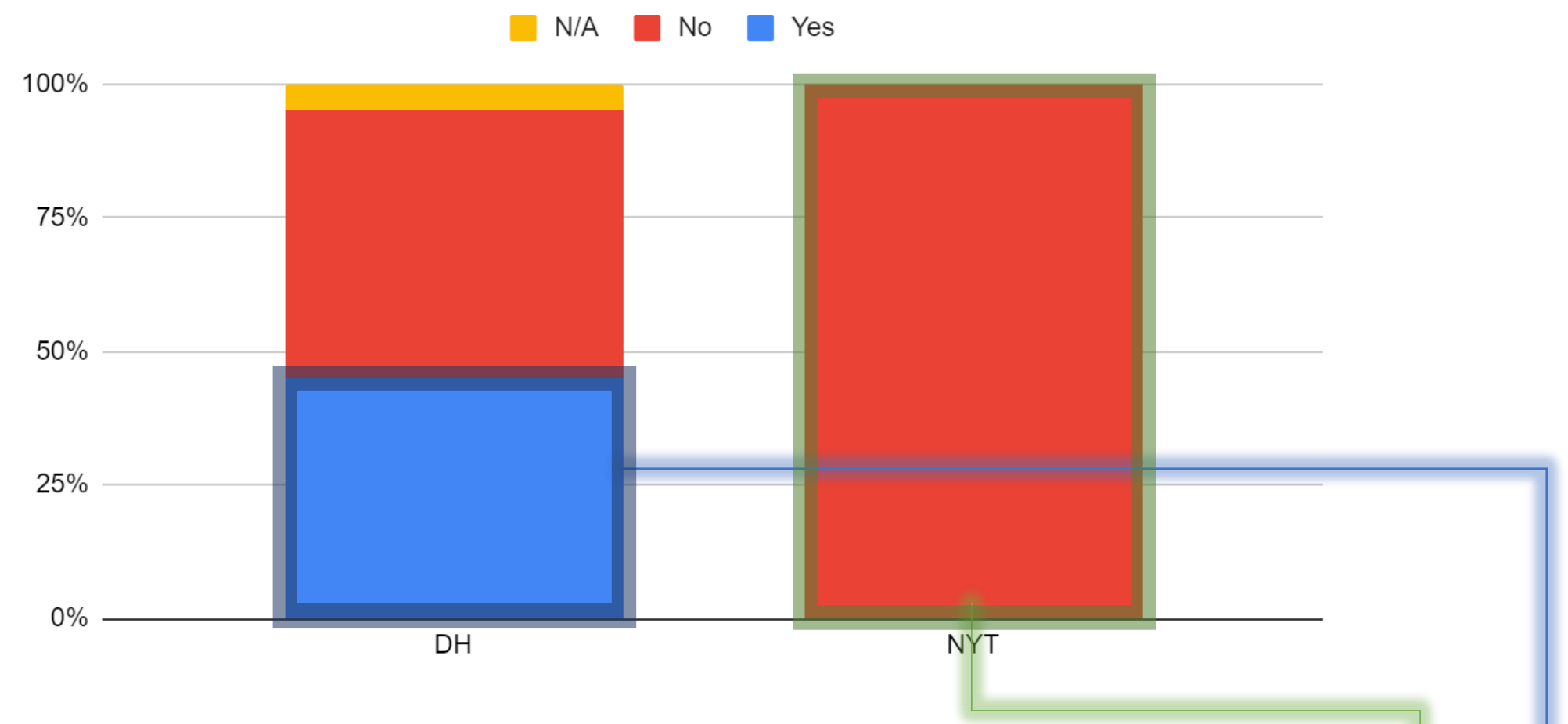
We thank the Carleton College and Gould Librarians for their support.

[1] Rivadeneira, Anna W., et al. "Getting our head in the clouds: toward evaluation studies of tagclouds." *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*. ACM, 2007.

[2] Alexander, Eric, et al. "Perceptual biases in font size as a data encoding." *IEEE Transactions on Visualization and Computer Graphics* 24.8 (2017): 2397-2410.

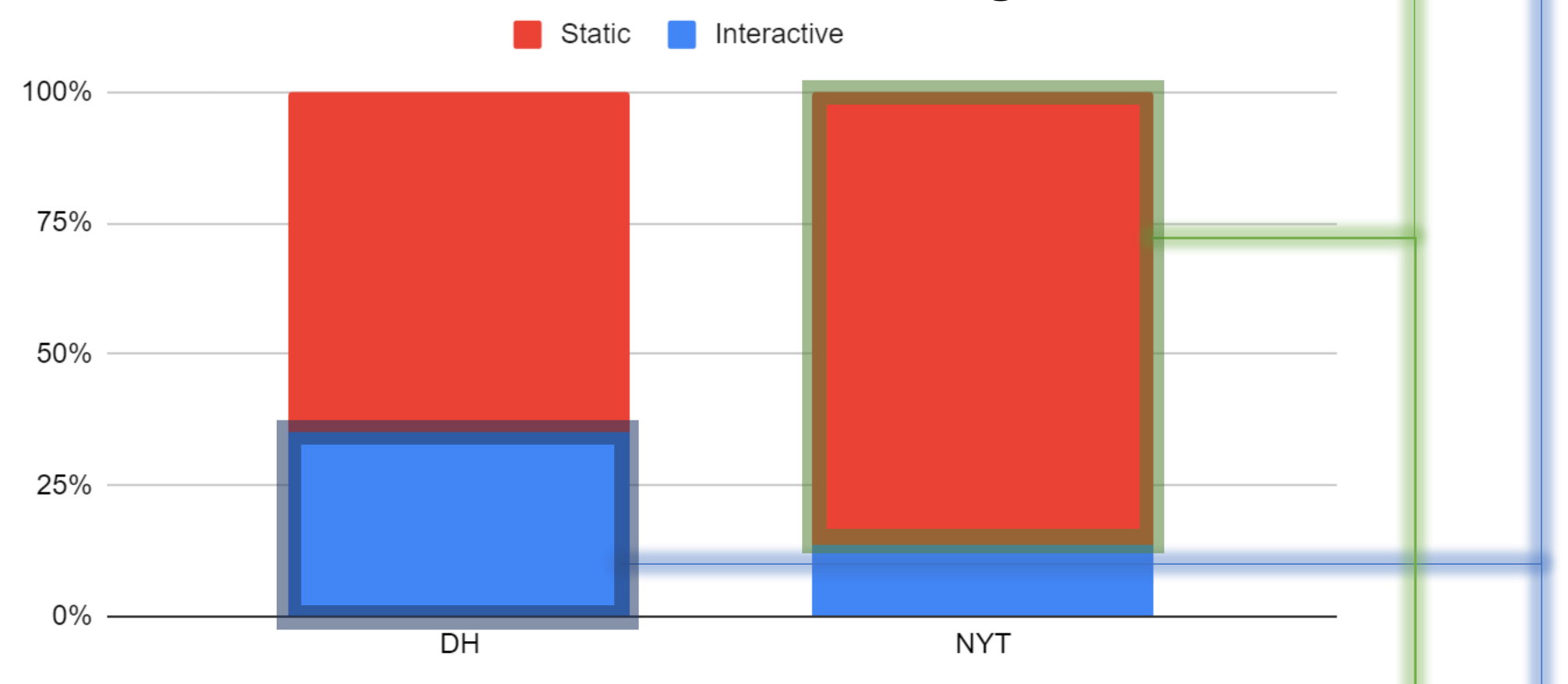
Muyang Shi **Email:** shim@carleton.edu
Danielle Albers Szafir **Email:** Danielle.Szafir@colorado.edu
Eric Alexander **Email:** ealexander@carleton.edu

Words Generated From Topic Modeling



A higher proportion of the DH word clouds are visualizations generated from the result of topic modeling algorithms. They are used to visualize the different topics of a multiple of documents. In contrast, word clouds from NYT mostly feature the frequency of words contained in the source corpus. None of them uses topic modeling

Interactive versus Static Design

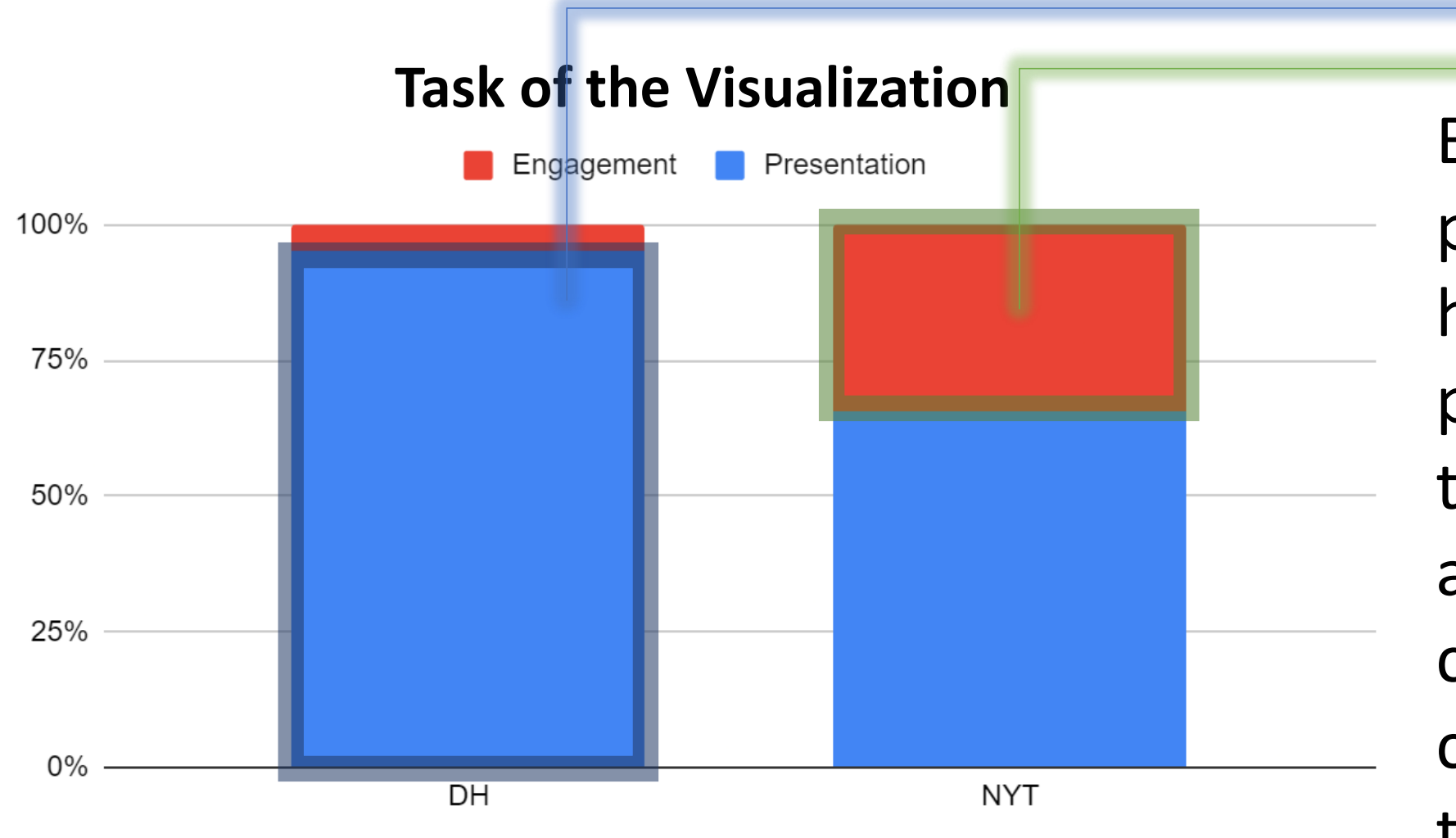


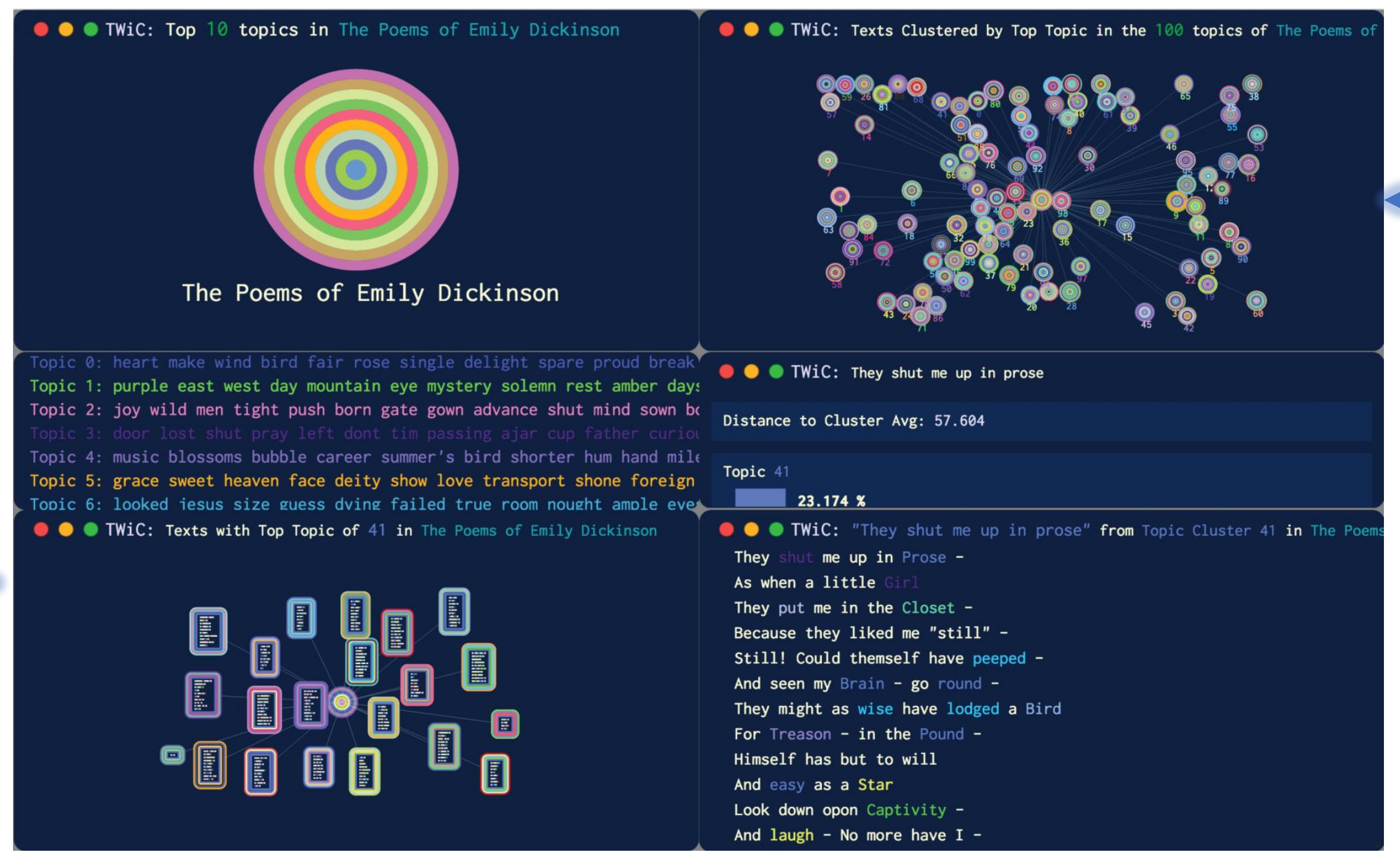
Both DH and NYT contain interactive as well as static word clouds. However, they differ in the proportion of word clouds being of an interactive versus static design. We think this is also related to the purpose/goal/task of the word clouds.

In DH, we see a high proportion of interactive design of word clouds, especially if the word cloud is part of the tool proposed by the paper. We think this is because word clouds come up as an internal/intermediate process for a user in processing a chain of events (in the middle of processing data).

In NYT, most of the word clouds are static. We think this is because the word cloud is "the ultimate presentation of the data" -- either for engagement/presentation/summarization, this is the "visual view" of the data, and the reader should refer to the raw text (or read the journal) to gain more detailed information.

Task of the Visualization





```
TWiC: Top 10 topics in The Poems of Emily Dickinson
Topic 0: heart make wind bird fair rose single delight spare proud break
Topic 1: purple east west day mountain eye mystery solemn rest amber days
Topic 2: joy wild men tight push born gate gown advance shut mind sown bo
Topic 3: door lost shut pray left dont tim passing ajar cup father curiou
Topic 4: music blossoms bubble career summer's bird shorter hum hand mile
Topic 5: grace sweet heaven face deity show love transport shone foreign
Topic 6: looked jesus size guess dying failed true room nought ample eve

TWiC: Texts Clustered by Top Topic in the 100 topics of The Poems of

TWiC: They shut me up in prose
Distance to Cluster Avg: 57.604
Topic 41
23.174 %

TWiC: Texts with Top Topic of 41 in The Poems of Emily Dickinson

TWiC: "They shut me up in prose" from Topic Cluster 41 in The Poems
They shut me up in Prose -
As when a little Girl
They put me in the Closet -
Because they liked me "still" -
Still! Could themself have peeped -
And seen my Brain - go round -
They might as wise have lodged a Bird
For Treason - in the Pound -
Himself has but to will
And easy as a Star
Look down opon Captivity -
And laugh - No more have I -
```

"Topic Words in Context": an in-browser tool for exploring the scales of data in a topic model. Armoza, Jonathan. How to Close Read a Topic Model: TWiC Reads Emily Dickinson's Fascicles, **DH 2017**.

Flash
Apple
Adobe
mobile
platform
developers
devices
web
open
apps
HTML5
third
products
video
many
party
available
standards
H.264
iPhones
iPods
websites

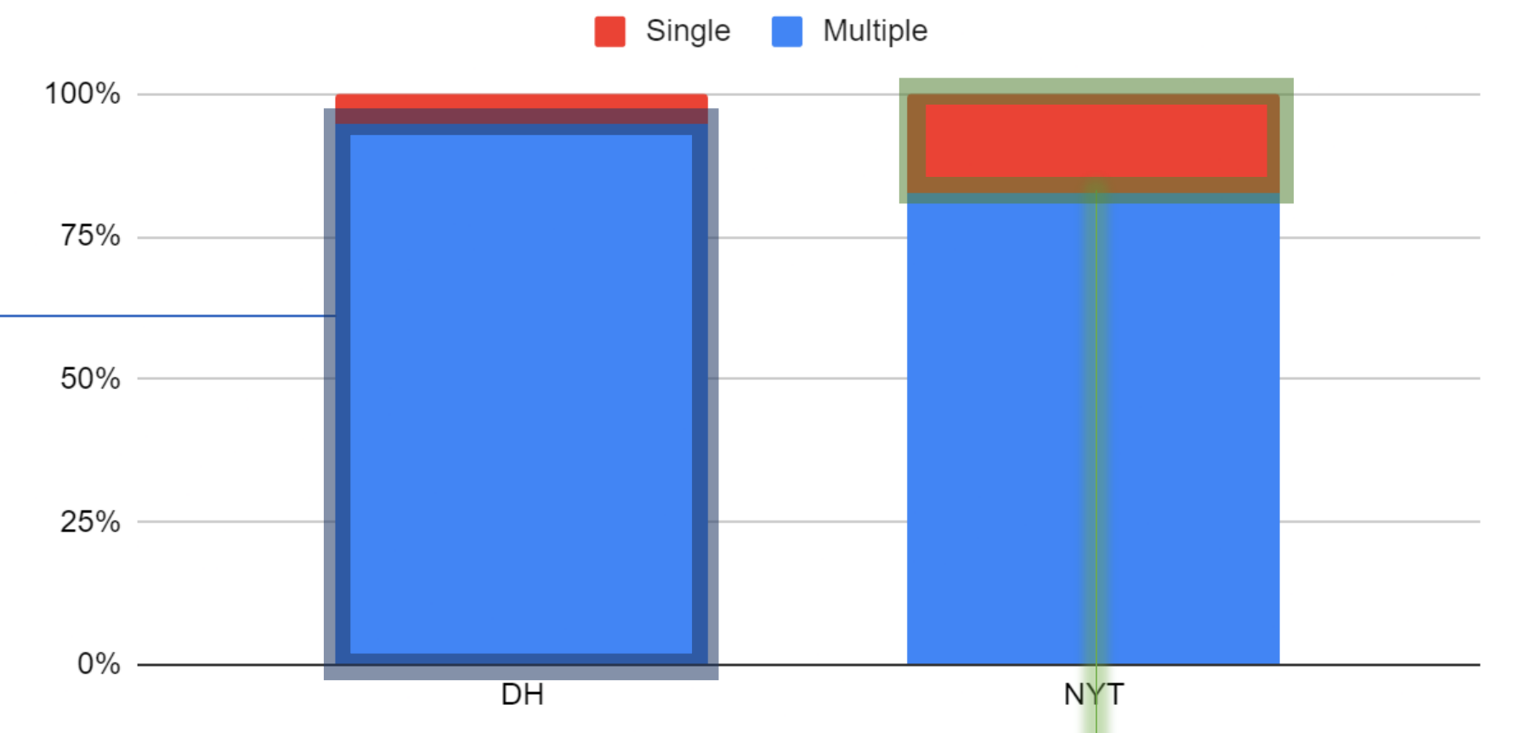
Flash
Player
Adobe
video
content
open
technologies
devices
H.264
source
mobile
multimedia
well
developed
Security
performance
including
better
complete
support

Apple's letter to Adobe and customers and Adobe's response to Apple and developers. Bilton, Nick. Word For Word: The Apple & Adobe Letters, **The New York Times** Bits.

Both DH and NYT use word clouds for summary purposes. In the DH word clouds, we see a higher proportion of them used for presentation purposes -- for example, many of them are used for illustrations of a tool built and proposed by the paper. In NYT, many word clouds are used for engagement -- they are colorful, visually appealing, placed at the top of the webpage to "grab attention".

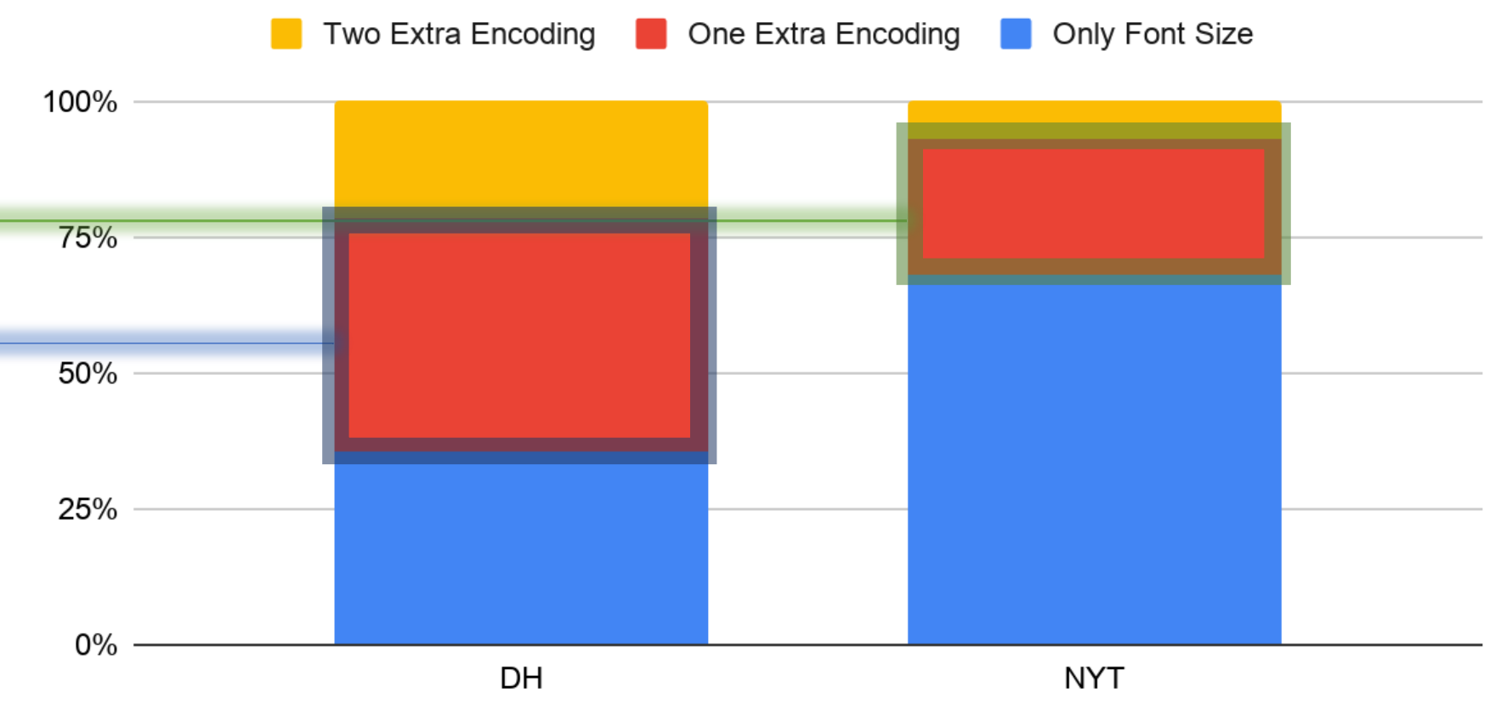
When a word cloud image is shown, it is more likely that the academic papers from DH include specific descriptions of where the document is coming from (their sources). This is done typically by explicitly mentioning the excerpt from some chapters of someone's work. However, NYT journals that include word cloud images usually point to some "interviews" or "speeches", without giving a clear clue as to the raw corpus the word cloud is made from.

Corpus Composition: Single or Multiple Documents



Both DH and NYT have a high proportion of word clouds being made from multiple documents. We do see a few more NYT word clouds generated from a single document.

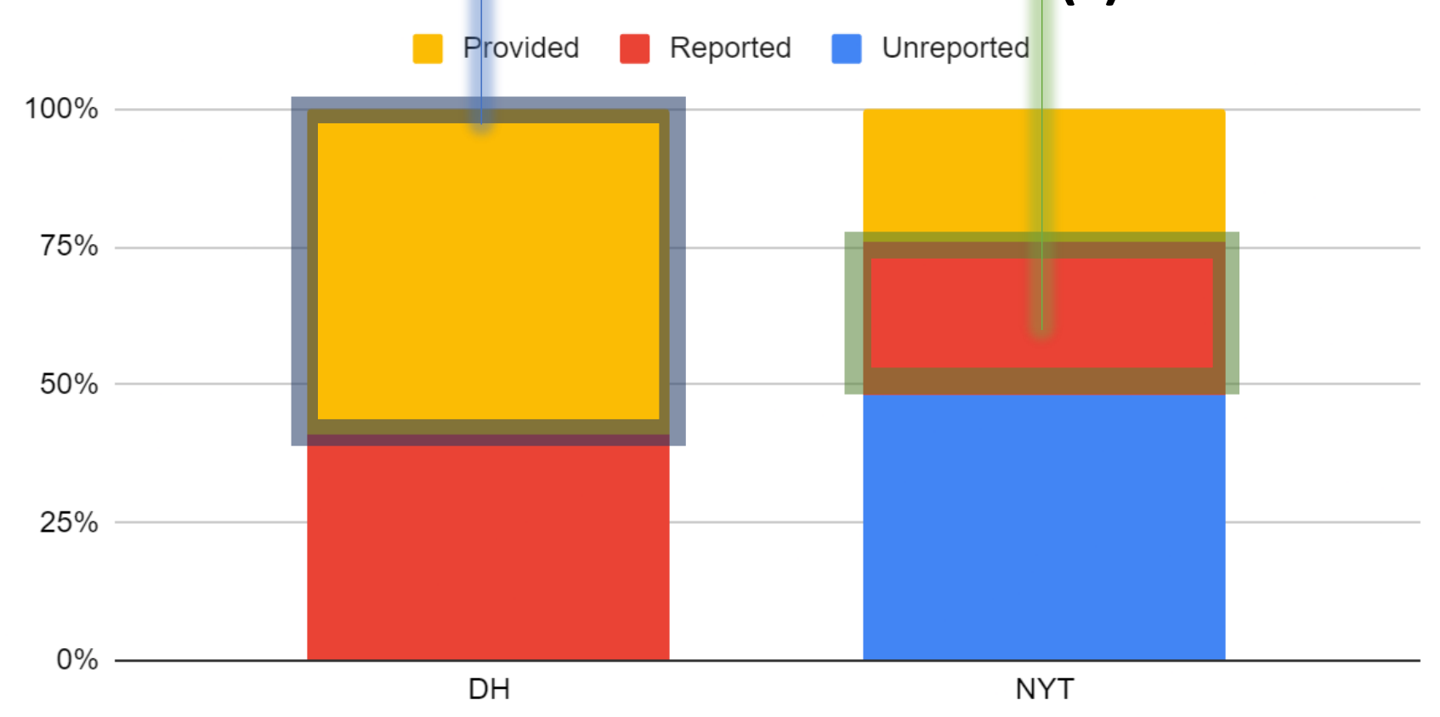
Number of Visual Feature(s)



Word clouds in both DH and NYT use font size as encoding, and basically none of them uses font or orientation as encoding. In DH word clouds, we see more word clouds using features such as position and color as "redundant encoding" (to reinforce the font size encoded data) or simply to encode more information into the words.

In NYT word clouds, there does not appear to be a clear usage of redundant encoding nor extra layer of information encoded using a feature other than font size.

References of Data Source(s)


Provided
Reported
Unreported
DH
NYT