Test Dataset for TextDNA

This is a practice dataset for users to gain an understanding of how TextDNA leverages color and position to explore statistical information based on a corpus. Furthermore, it is a demonstration to build user trust in the tool. Because big data can be overwhelming, this dataset is simple and artificial. The corpus contains seven text documents, each populated by four words: “the,” “example,” “test,” and “count.” More detailed information regarding the documents can be found in the testTextDNA_summary spreadsheet, which is meant to be read alongside this tutorial. The majority of the words in the test data set are “the.” “Example,” “test,” and “count” have been used sparingly to demonstrate TextDNA’s functions.

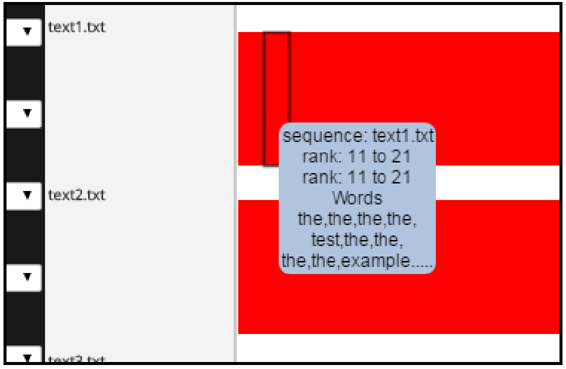
Word Rank

Word Rank refers to the position of a word in a sequence. TextDNA assigns a rank, or position, to every word with a text document, starting with 0 and ending with the number of words minus one (n-1). Words are displayed from left to right, the first word to the last word.

Table 1: Word Ranking in Raw Text

RANK	0	1	2	3	4	5	6
WORD	This	is	an	example	sentence	with	ranks.

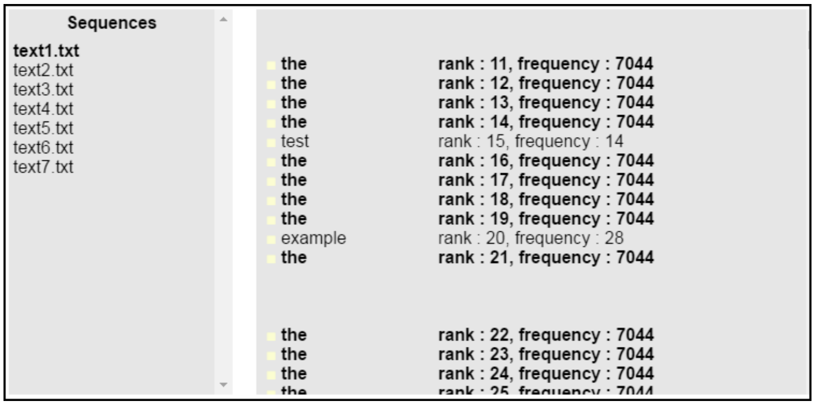
Figure 1: Tooltip



sequence: text1.txt
rank: 11 to 21
rank: 11 to 21
Words
the,the,the,the,
test,the,the,
the,the,example.....

Word rank can be accessed in TextDNA by two methods: through a tooltip that displays when hovering over a block in a sequence, and a Show Words module that provides detailed information about an entire block. Hovering over blocks in sequences will display the tool tip, which shows preliminary information such as rank in reference. The summary spreadsheet provides rank information for the sparingly used words in this artificial dataset, which you can use to help you search for them in the visualization. For example, the tooltip in Figure 1 shows information about a block in text1. The block contains words ranked from 11 to 21, and “test” appears in this block. The summary spreadsheet indicates that “example” occurs at rank 20 in text1.

Right-clicking on the block will provide a pop up list of options. Select “Show Words.” Figure 2 shows the Words module for the block in Figure 1. The module has two panels that display information about the words contained within a selected text sequence. The panel on the left lists the dataset’s text sequences and bolds the specific sequence you’re examining. The panel on the right lists words and related information, with blocks demonstrating the color mapped to the word. In Figure 2 you can see word rank and frequency are also listed. “Example” is at rank 20, and we can see that “test” appears at rank 15.



Sequences

text1.txt
text2.txt
text3.txt
text4.txt
text5.txt
text6.txt
text7.txt

the	**rank : 11, frequency : 7044**
the	**rank : 12, frequency : 7044**
the	**rank : 13, frequency : 7044**
the	**rank : 14, frequency : 7044**
test	rank : 15, frequency : 14
the	**rank : 16, frequency : 7044**
the	**rank : 17, frequency : 7044**
the	**rank : 18, frequency : 7044**
the	**rank : 19, frequency : 7044**
example	rank : 20, frequency : 28
the	**rank : 21, frequency : 7044**
the	**rank : 22, frequency : 7044**
the	**rank : 23, frequency : 7044**
the	**rank : 24, frequency : 7044**
the	**rank : 25, frequency : 7044**

Figure 2: Words Module

Word rank has different meanings for n-gram data and raw text datasets. For datasets built on n-grams, the word rank can correlate with popularity. Using one-grams from the Google Books Dataset as an example, the most popular word in the printed record from 2000 to 2009 is "the." The higher the word rank in an n-gram dataset, the less popular it is. Raw text is more challenging to think about. Word rank for raw texts is the ordinal position of a word within a text read sequentially, as demonstrated in Table 1.

Word Frequency

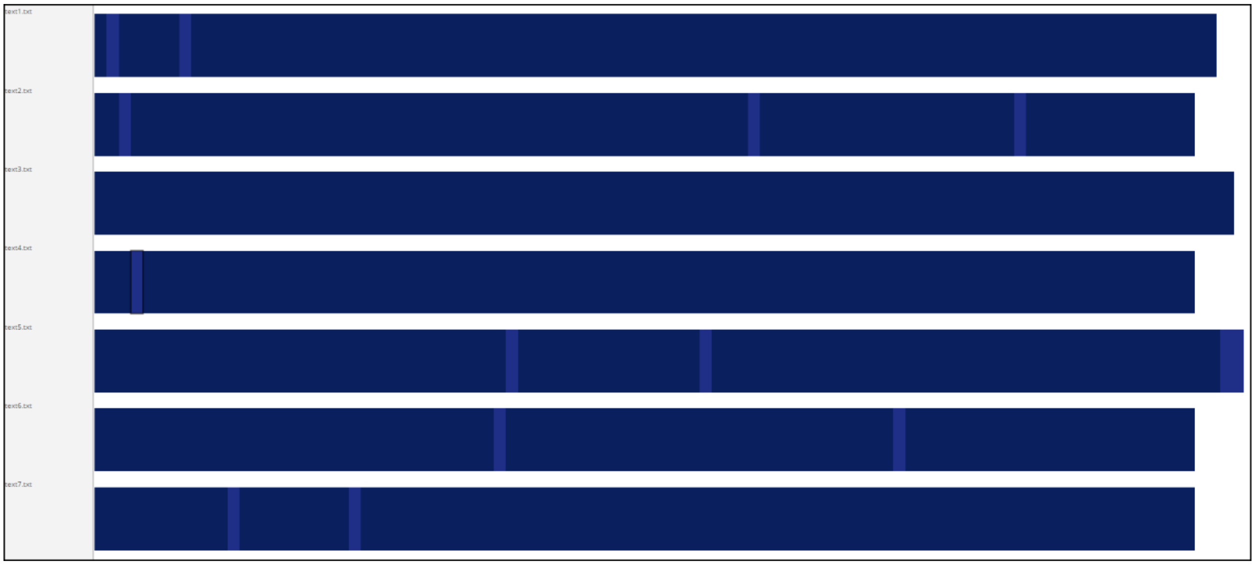


Figure 3: Word Frequency

Word Frequency refers to the number of times a word occurs within a corpus. Change "Color By" to word frequency and "Order By" to word rank. This combination allows you to see where words sequentially occur within a text and how many times they are used through color. TextDNA offers two color aggregations, averaging and color weaving, to highlight patterns of different scales. This example will rely on averaging, and color weaving will be explained in Sequence Co-Occurrence. As an aggregation type, averaging works by assigning a color value to sequence block that is average color mapped to the words within it.

With “Color Scheme” set to Blue-Yellow sequential and “Aggregation Type” to averaging, the horizontal bars that represent text sequences are mainly dark blue with bands of lighter blue, shown in Figure 3. In those bands of lighter blue there is word variation.

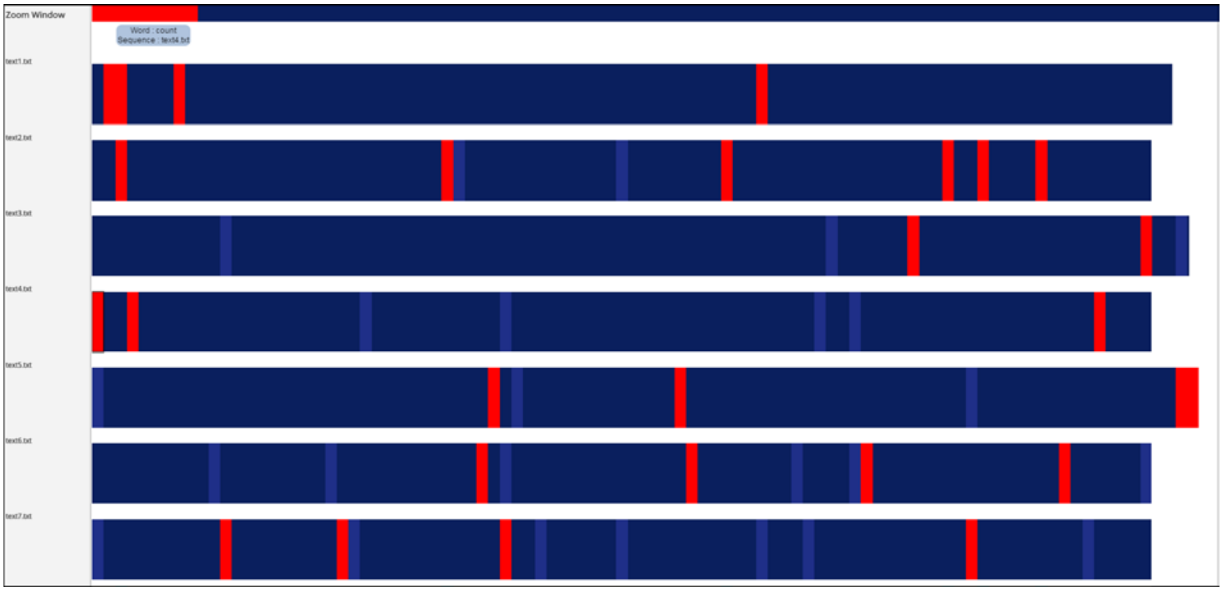


Figure 4: Matching On Highlighting

Take the first lighter blue block in text4, for example. Left-click on it to lock the information in the block to the zoom window at the top of the screen. The words that the block represents are mapped to individual blocks in the zoom window. You will see a yellow block with dark blue ones in the zoom window. If you hover over the yellow block, the tool tip will let you know which word it is. Here, it is “count.”

To explore words that occur equally as often, change “Match On” to frequency. “Match On” will highlight selected criteria across the sequences in bright red. Set it to Frequency and if you hover over “count” in the zoom window, it will highlight blocks in sequences that contain “count” and “test.” In the dataset, “test” and “count” have the same frequency (14).

Sequence Frequency

Sequence Frequency refers to the number of sequences in which a word occurs, allowing you to detect global- and sequence-level patterns. This underwhelming, artificial dataset is populated with four words to showcase how TextDNA’s functions work. The word “test,” which you can see on the summary sheet, is the only word that doesn’t appear in all sequences. It is absent from text3.txt. “Test”’s sequence frequency is six, unlike “example”’s and “count”’s, which are seven. Sequence frequency is another useful measure, like word rank, to gauge word popularity in datasets.

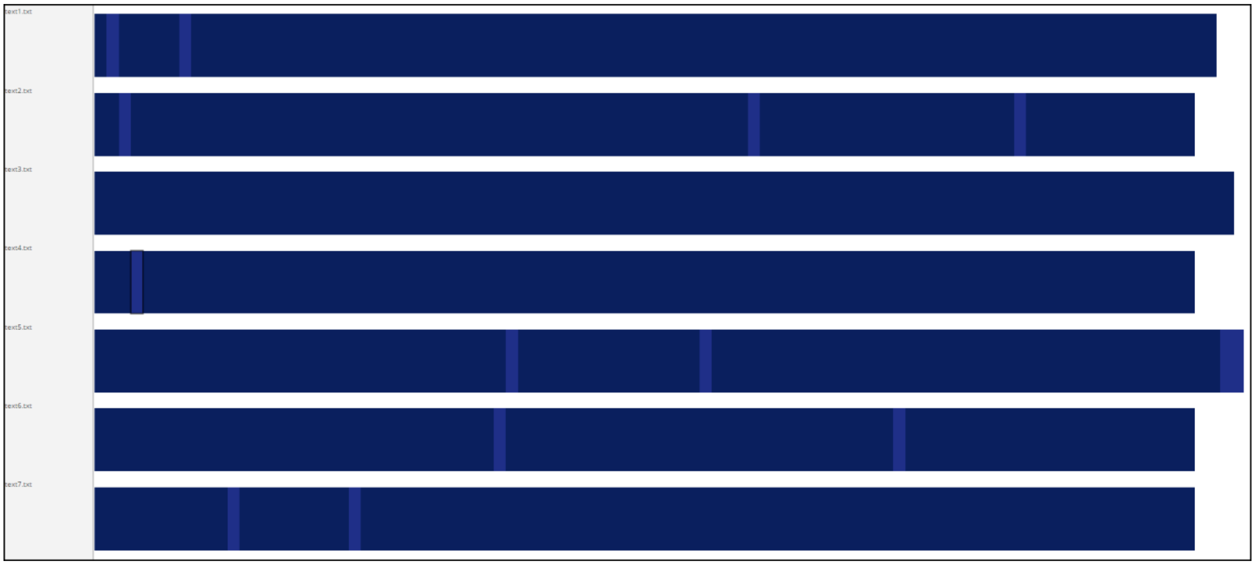
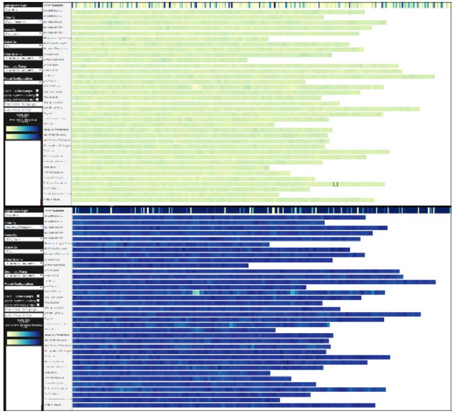


Figure 5: Sequence Frequency

Sequence frequency is only available as a "Color By" encoding. If you select it in the "Color By" dropdown menu, the blocks within sequences will be colored according to the popularity of the words within them. With the current color ramp (Blue-Yellow Sequential), the most popular words will be colored in dark blue, the least popular words yellow. You can see that sequence three has a uniform dark blue color, which means that the words in the blocks in sequence three can be found in all other sequences. With "Aggregation Type" set to Color Weaving, you can see blocks with yellow pixels in them. These blocks with yellow pixels contain words that aren't included in every sequence of the data set. Comparing the blocks with the input on testTextDNA_summary.csv you will see that these are the blocks that contain the word "test."

Figure 6: Shakespeare Raw Text



Given the simplicity of the test dataset, coloring by sequence frequency and by word frequency offer similar overviews. This is not the case with actual with actual datasets, especially raw text datasets. Figure 6 demonstrates the difference between the raw text Shakespeare dataset when coloring is mapped to word frequency (top) and sequence frequency (bottom) with "Aggregation Type" set to averaging. Unlike raw-text datasets, n-gram datasets that explore word usage over time offer potentially less data complexity due to integral components of language evolving slowly over time. For example, definite and indefinite articles (e.g., the, an) and prepositions (e.g., to, for, on) are among the most commonly used words in the English language. N-gram datasets can offer less data complexity due to the word rank corresponding to the metric (e.g., popularity by decade) used to generate a corpus.

Sequence Co-Occurrence

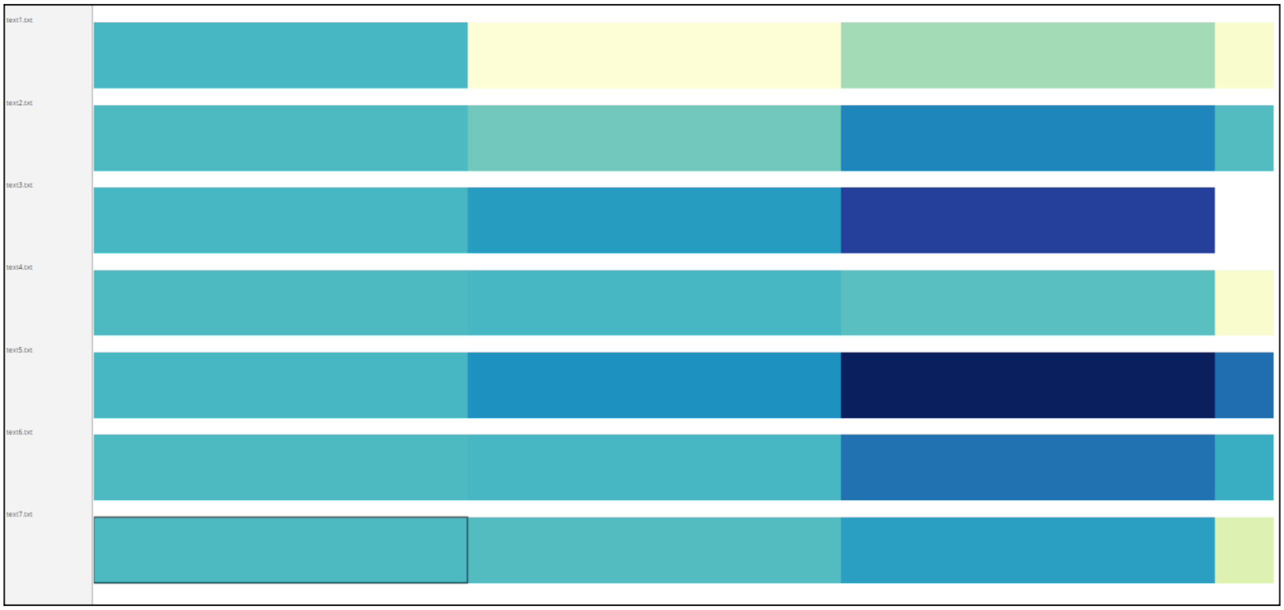
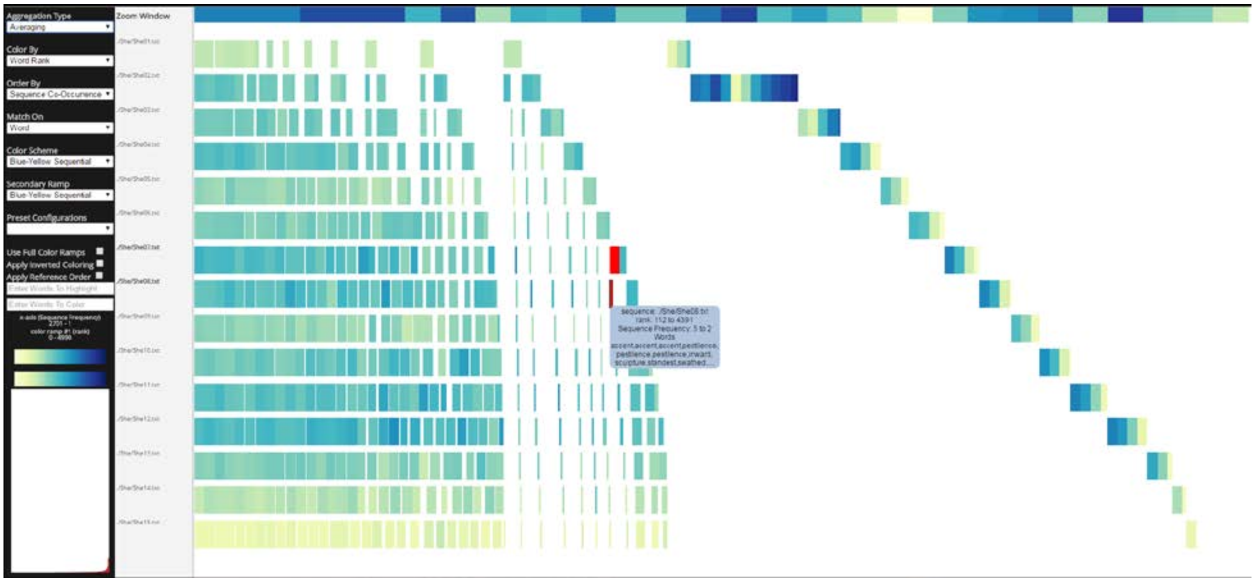


Figure 7: Sequence Co-Occurrence Ordering in Test Dataset

Sequence Co-Occurrence measures the co-occurrence of words within a dataset. It ranks words first by the number of sequences in which they occur, then by the set of sequences words occur in. Words are arranged in alphabetical order across these levels.

Figure 7 shows sequence co-occurrence for the test dataset when “Order By” is set to sequence co-occurrence and “Aggregation Type” is averaging. Words with the highest sequence co-occurrence are on the left of the visualization and the words with the least co-occurrence are on the right. You will notice that each word has its own column. Hovering over the first column you can see that it is devoted to representing the instances of “the” across the sequences. The next is “example,” followed by “count.” The smallest blocks at the right represent “test.” These blocks are smaller because there are fewer occurrences of “test” in any given sequence compared to the number of the “example,” the number of “count,” and the number of “the.”

Figure 7 provides a tidy example of what sequence co-occurrence looks like. Figure 8 shows demonstrates the co-occurring language in the serial novel *She: A History of Adventure* (1887), a raw-text corpus. It is what you might see in your own data, and it reads the same way. The smaller the block, the less the words it represents occur compared to the instances of other words in any given sequence. In Figure 8 you will also notice more white space, and the blocks appear to be grouped. The grouping of tiered blocks farthest right contains words that are unique—i.e., they occur only within that sequence of the corpus (they aren’t co-occuring within the corpus).



Aggregation Type: Averaging

Color By: Word Rank

Order By: Sequence Co-Occurence

Match On: Word

Color Scheme: Blue-Yellow Sequential

Secondary Ramp: Blue-Yellow Sequential

Preset Configurations:

Use Full Color Ramps □
Apply Inverted Coloring □
Apply Reference Order □

x-axis (Sequence Frequency)
2701 - 1
color ramp #1 (rank)
0 - 4996

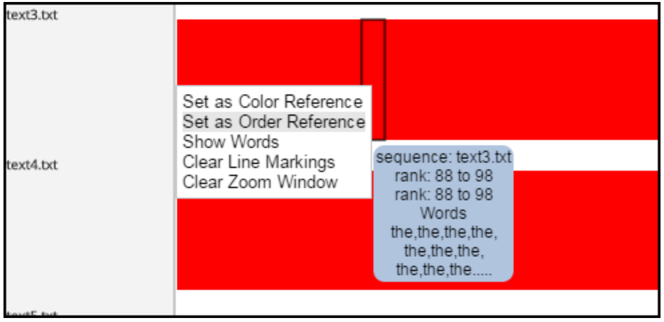
sequence: /She/She08.txt
rank: 112 to 4391
Sequence Frequency: 5 to 2
Words
accent,accent,accent,pestilence,
pestilence,pestilence,inward,
sculpture,standest,swathed.....

Figure 8: Sequence Co-Occurrence Ordering in the She: A History of Adventure dataset

In both Figures 7 and 8, color differences in the blocks, since "Color By" is set to word rank and "Aggregation Type" to averaging, demonstrate where the majority of the instances of the respective words occur. Yellow maps to words near the beginning of the sequences, while blue maps to words at the end of sequences, and mixed hues respectively in between. The first column of blocks in Figure 7, representing "the," are a blue-green color. This average color reflects that "the" appears evenly throughout all the sequences. You will also note a lot of color variance down the column for the word "count." The dark blue in "text5.txt" reflects that "count" appears mainly near the end of that sequence. (The summary CSV shows that it appears at rank 1,044.) The next darkest blue belongs to text3.txt, and the summary sheet indicates "count" appears at ranks 773 and 1,000. You can look at other blocks and see where the words appear in specific sequences to understand their colors.

Rank in Reference

Figure 9: Making a Sequence an Order Reference



sequence: text3.txt
rank: 88 to 98
rank: 88 to 98
Words
the,the,the,the,
the,the,the,
the,the,the.....

Rank In Reference allows you to order or color all data in a dataset according to a specific sequence. Since text3.txt doesn't contain all the words within the dataset, it proves a good use case. Right-click on the sequence of text3.txt and select "Set as Order Reference" (see Figure 9). TextDNA will automatically order the content in other sequences according to the selected one. TextDNA sorts other sequences according to the first instance of a word within the selected sequence.

Text3.txt is a document populated with the words "the," "example" and "count." "Example" occurs at ranks 125, 700, and 1,030. "Count" occurs at ranks 773 and 1,000. "The" appears everywhere else. Text3.txt does not contain "test."

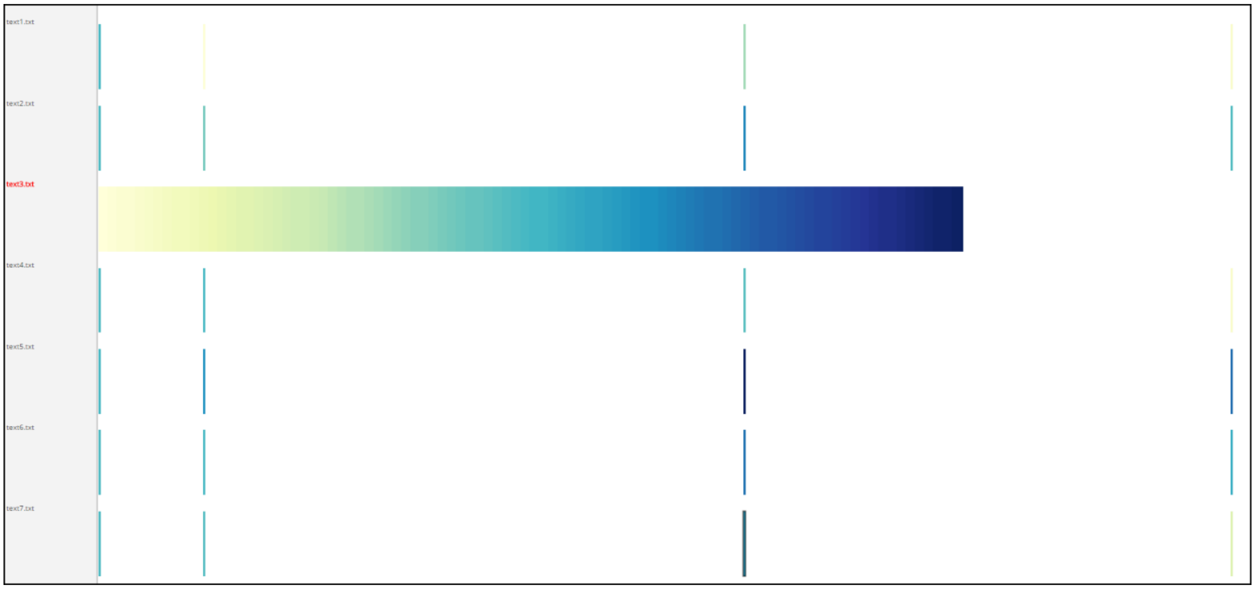


Figure 10: Test Dataset with text3.txt as Order Reference

Selecting text3.txt as the reference order transforms the visualization spatially (Figure 10). Selecting a sequence as an order reference (or color reference) maps the specific data property to the entire dataset. The selected sequence can be used as a key to read other input. Note how sequence three is a continuous horizontal bar while the others have been segmented according to the ordering of text3.txt. Hovering over the second column of block slices, you will see they represent the word "example." Besides "the," "example" comes earliest in the text. All instances of "example" in the other text sequences are mapped to blocks in this column. Since "Match On" is set to word, you can see the other instances of blocks where "example" occurs in text3.txt highlighted in red. You can see the same if you move to the column where "count" is. Because "test" does not appear in text3.txt, the blocks representing this word in other sequences are pushed spatially beyond the reference, located at the right end of the visualization. Content to the right of the reference sequence is arranged according to sequence co-occurrence. Words common to most sequences are arranged on the left, and unique words to the right.

See Figure 11 for how a raw-text corpus looks with an order reference set.

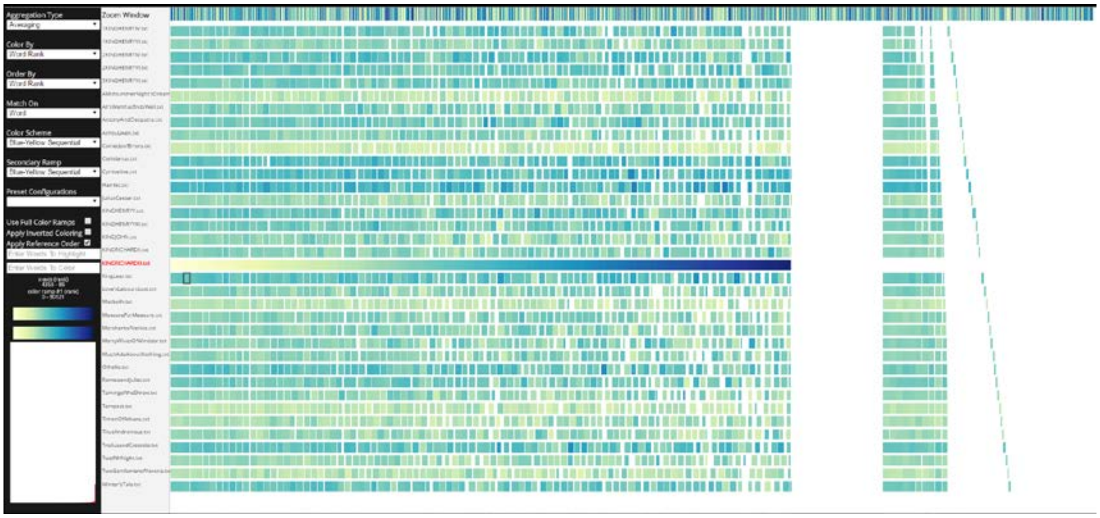
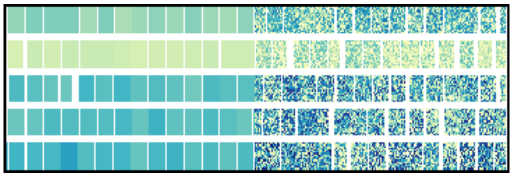

Aggregation Type: Averaging
Color By: Word Rank
Order By: Word Rank
Match On: Word
Color Scheme: Blue-Yellow Sequential
Secondary Ramp: Blue-Yellow Sequential
Preset Configurations:
Use Full Color Ramps □
Apply Inverted Coloring □
Apply Reference Order ☑
Enter Words To Highlight
Enter Words To Color

Figure 11: Shakespeare Raw Text with Richard III set as Order Reference

Aggregation Type Color Weaving

Though the examples so far do not showcase "Aggregation Type" color weaving, it is an important setting to explore data granularity. Unlike averaging, color weaving accounts for the variety of color encodings mapped to the words that constitute a block (Figure 12).

Figure 12: Averaging vs Color Weaving Aggregation

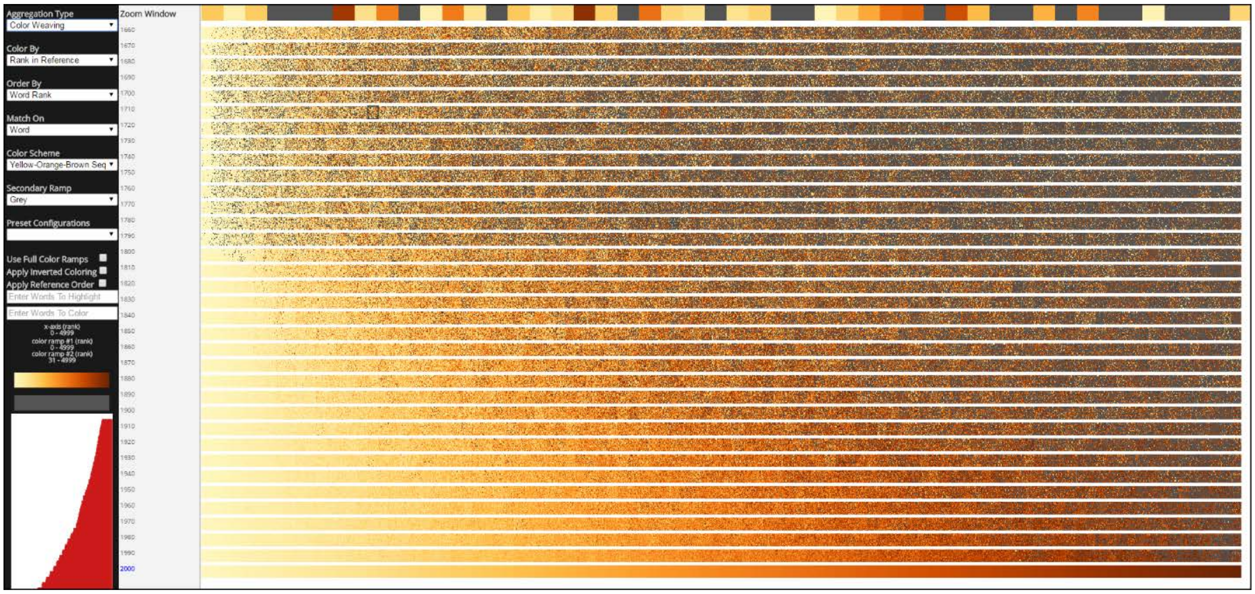


Color weaving offers an approximate distribution of values of the "Color By" parameters across elements in the block. The color parameters for each word are mapped to individual pixels in the block. After each word color encoding is mapped to a pixel, the elements are randomized and mapped to the remaining pixels. The resulting randomization repeats until the block is filled.

Color weaving highlights small-scale variations within blocks and is a powerful color encoding for locating interesting words to guide analysis of the corpus.

Split Color Encodings

Of the five default color encodings, two are split color properties: a portion of the elements are mapped to one ramp while the remainder are mapped to another. They require you to choose the primary "Color Scheme" and "Secondary Ramp." These color encodings are sequence co-occurrence and rank in reference.



Aggregation Type: Color Weaving
Color By: Rank in Reference
Order By: Word Rank
Match On: Word
Color Scheme: Yellow-Orange-Brown Seq
Secondary Ramp: Grey
Preset Configurations:
Use Full Color Ramps □
Apply Inverted Coloring □
Apply Reference Order □
Enter Words To Highlight
Enter Words To Color
x-axis (rank) 0 - 4999
color ramp #1 (rank) 0 - 4999
color ramp #2 (rank) 31 - 4999

Figure 13: Split Color Encoding with Grey Secondary Ramp in Top 5000 Google N-Grams and 2000 as Reference Decade

Figure 13 is an example of split color encoding, using the Top 5000 Google N-Grams (1660-2009) dataset. Sequence 2000 is set as the Rank in Reference for "Color By," with the "Color Scheme" set to Yellow-Orange-Brown Sequential and the "Secondary Ramp" as Grey. Sequence 2000 has been set as the "Color By" property rank in reference, so its content uses the "Color Scheme." These settings filter the words in the sequences according to how it appears in the 2000 sequence. Content that matches words from the 2000 sequence are colored according to their position within the 2000 sequence, and words that do not appear in the decade of reference map to the "Secondary Ramp," the grey color.

Figure 13 utilizes the split color encoding to compare words from the first decade of the 21st century with the words in other sequences. Color here is a way to understand what words throughout the corpus are unique to the reference decade and co-occurring with other decades. Grey maps to words not found within the reference decade.

Last Updated April 28, 2016 by Deidre Stuffer