Text Analysis with Voyant Tools & AntConc

Mia Partlow and Erica Hayes
NCSU Libraries

Learning Goals

- Introduce you to Text Analysis & best practices for building a corpus
- Learn the capabilities of Voyant Tools and AntConc for text reading, analysis and visualization
- Load files of various types and URLs into Voyant Tools & AntConc
- Navigate the interfaces of both tools in multiple ways
- Export your views and download images of views

What is Text Analysis?

"Text Analysis is the search for and discovery of patterns and trends in a corpus of texts. The analysis of those patterns and trends can help researchers uncover previously unseen characteristics of a specific corpus, deconstruct a text, and reveal new ideas and theories about a particular genre, topic or author." **-Meredith Dabek (2014)**

Two guiding questions for text analysis -Sinclair & Rockwell (2015):

- For texts with which I am already familiar, how can computers help me identify and study interesting things I had not noticed before, or things I had noticed but did not have reasonable means to pursue?
- How can computers help me identify and understand texts with which I am not familiar with or which I cannot reasonably read?

Concepts & Challenges

- **Corpus**: a collection of texts.
  - Consider the size of your corpus and your inclusion and exclusion criteria
- **Data cleaning**: removing information not relevant to your analysis.
  - Examples: chapter numbers and titles, copyright information, headers and footers, page numbers.
  - Languages and tools such as R, Python, and Regular Expressions are necessary with a large corpus.
- **Stopwords**: words you filter out of your analysis
  - Standard lists exist (e.g. Buckley-Salton Stoplist)
  - Voyant has a stoplist included
  - Customization of the stoplist may be necessary

Choosing a Corpus

Text Data Sources:

- Project Gutenberg ebooks
- HathiTrust Digital Library
- Datasets and Scholarly APIs at NC State
- Surveys and interview data

"Quantitative analysis tends to require context before it becomes meaningful. It doesn't mean much to say that the word "motion" is common in Wordsworth, for instance, until we know whether "motion" is more common in his works than in other nineteenth-century poets. So yes, text-mining can provide clues that lead to real insights about a single author or text. But it's likely that you'll need a collection of several hundred volumes, for comparison, before those clues become legible"

- Ted Underwood (2012)

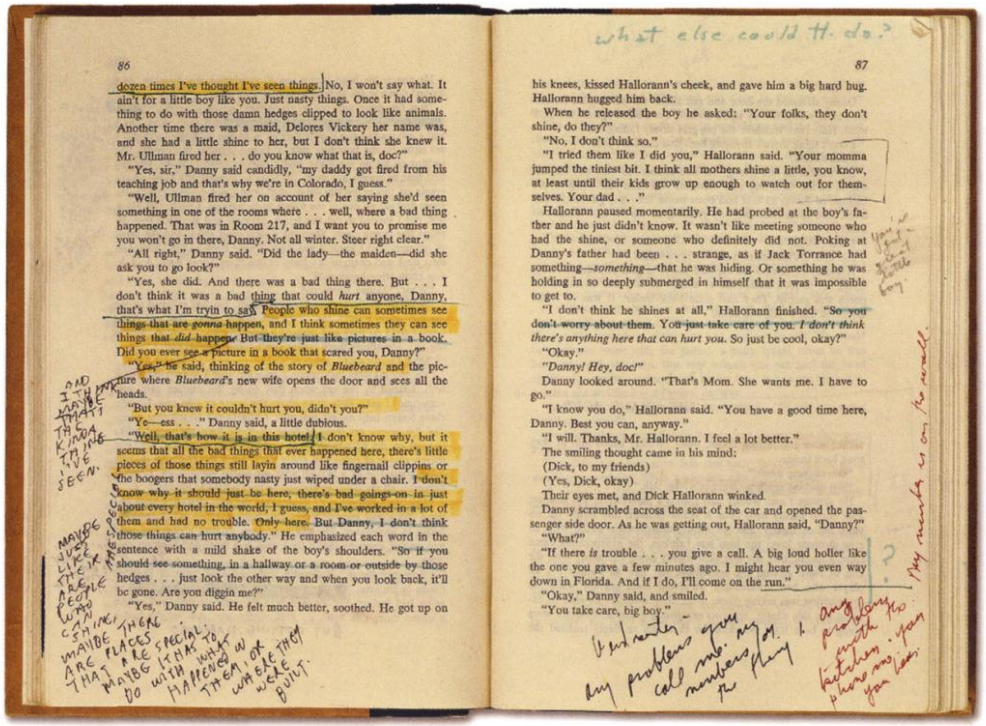
Copyright

- Mining databases or corpora to conduct scholarly analysis, without subsequently republishing the contents of those databases or corpora, can be fair use under Author's Guild v. HathiTrust, 755 F.3d 87 (2d Cir. 2014).
- A database or website you are using to create your corpus might regulate text mining, or might allow text mining but prohibit sharing excerpts from the materials that they made available for mining. Therefore, it's important to carefully read any terms of use or licenses as you prepare your project for publication.

Close Reading or Microanalysis

Close Reading/Microanalysis

"Essentially, close reading means reading to uncover layers of meaning that lead to deep comprehension." - Nancy Boyles (2013)



86

dozen times I've thought I've seen things. No, I won't say what. It ain't for a little boy like you. Just nasty things. Once it had something to do with those damn hedges clipped to look like animals. Another time there was a maid, Delores Vickery her name was, and she had a little shine to her, but I don't think she knew it. Mr. Ullman fired her . . . do you know what that is, doc?"

"Yes, sir," Danny said candidly, "my daddy got fired from his teaching job and that's why we're in Colorado, I guess."

"Well, Ullman fired her on account of her saying she'd seen something in one of the rooms where . . . well, where a bad thing happened. That was in Room 217, and I want you to promise me you won't go in there, Danny. Not all winter. Steer right clear."

"All right," Danny said. "Did the lady—the maiden—did she ask you to go look?"

"Yes, she did. And there was a bad thing there. But . . . I don't think it was a bad thing that could *hurt* anyone, Danny, that's what I'm tryin to say. People who shine can sometimes see things that are *gonna* happen, and I think sometimes they can see things that *did* happen. But they're just like pictures in a book. Did you ever see a picture in a book that scared you, Danny?"

"Yes," he said, thinking of the story of *Bluebeard* and the picture where *Bluebeard*'s new wife opens the door and sees all the heads.

"But you know it couldn't hurt you, didn't you?"

"Y—ess . . ." Danny said, a little dubious.

"Well, that's how it is in this hotel. I don't know why, but it seems that all the bad things that ever happened here, there's little pieces of those things still layin around like fingernail clippins or the boogers that somebody nasty just wiped under a chair. I don't know why it should just be here, there's bad goings-on in just about every hotel in the world, I guess, and I've worked in a lot of them and had no trouble. Only here. But Danny, I don't think those things can hurt anybody." He emphasized each word in the sentence with a mild shake of the boy's shoulders. "So if you should see something, in a hallway or a room or outside by those hedges . . . just look the other way and when you look back, it'll be gone. Are you diggin me?"

"Yes," Danny said. He felt much better, soothed. He got up on

87

what else could H. do?

his knees, kissed Hallorann's cheek, and gave him a big hard hug. Hallorann hugged him back.

When he released the boy he asked: "Your folks, they don't shine, do they?"

"No, I don't think so."

"I tried them like I did you," Hallorann said. "Your momma jumped the tiniest bit. I think all mothers shine a little, you know, at least until their kids grow up enough to watch out for themselves. Your dad . . ."

Hallorann paused momentarily. He had probed at the boy's father and he just didn't know. It wasn't like meeting someone who had the shine, or someone who definitely did not. Poking at Danny's father had been . . . strange, as if Jack Torrance had something—*something*—that he was hiding. Or something he was holding in so deeply submerged in himself that it was impossible to get to.

"I don't think he shines at all," Hallorann finished. "So you don't worry about them. You just take care of you. *I don't think there's anything here that can hurt you.* So just be cool, okay?"

"Okay."

"*Danny! Hey, doc!*"

Danny looked around. "That's Mom. She wants me. I have to go."

"I know you do," Hallorann said. "You have a good time here, Danny. Best you can, anyway."

"I will. Thanks, Mr. Hallorann. I feel a lot better."

The smiling thought came in his mind:

(Dick, to my friends)

(Yes, Dick, okay)

Their eyes met, and Dick Hallorann winked.

Danny scrambled across the seat of the car and opened the passenger side door. As he was getting out, Hallorann said, "Danny?"

"What?"

"If there *is* trouble . . . you give a call. A big loud holler like the one you gave a few minutes ago. I might hear you even way down in Florida. And if I do, I'll come on the run."

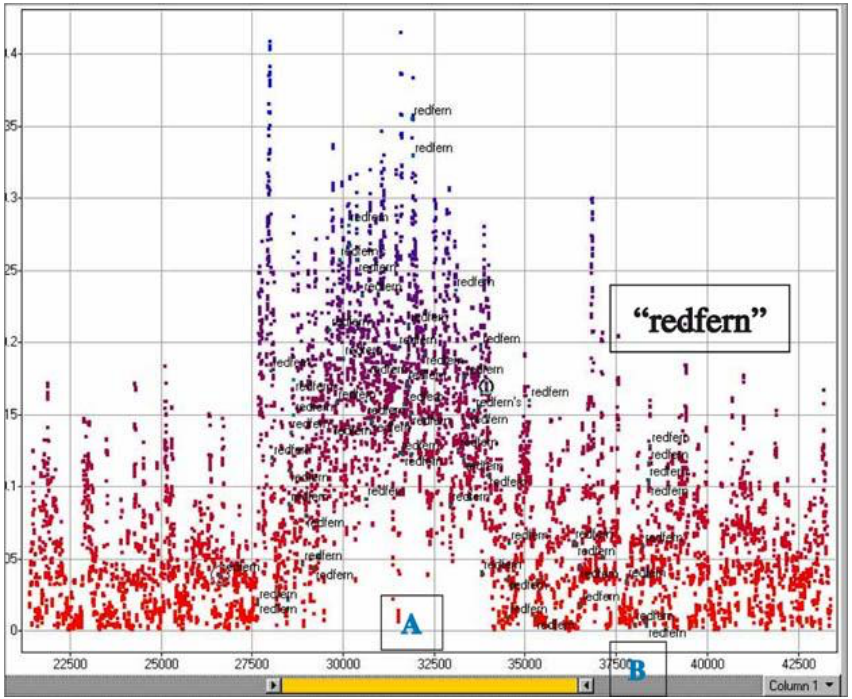
"Okay," Danny said, and smiled.

"You take care, big boy."

Distant Reading or Macroanalysis

Distant Reading/Macroanalysis

“Distant reading aims to generate an abstract view by shifting from observing textual content to visualizing global features of a single or of multiple text(s)” -Stefan Jänicke (2016)



Benefits of both forms of reading and analysis

"It is the exact interplay between the macro and micro scale that promises a new, enhanced, and perhaps even better understanding of the literary record. The two approaches work in tandem and inform each other. Human interpretation of the 'data,' whether it be mined at the macro or micro level, remains essential. While the methods of enquiry, of evidence gathering, are different, they are not antithetical, and they share the same ultimate goal of informing our understanding of the literary record, be it writ large or small"

-Matthew Jockers "On Distant Reading and Macroanalysis" (2011)

Text Analysis activity goals for today's workshop

- Explore word frequencies, occurrences of keywords, concordances and collocations in 8 fairy tales from Project Gutenberg
- The goal of this activity is to introduce you to the Voyant Tools & AntConc interfaces to enable quantitative analysis of the texts.

Voyant Tools

Voyant: The Good & The Bad

Positives:

Allows for quick reading and analysis of texts

Easier to identify themes and topics to explore further

Comparative options to analyze texts in tandem

~21 tools for visualizing the texts

HTML5 based tools (unlike Voyant 1.0)

Negatives:

Web application frequently breaks down/fails to load

Slow loading time with large documents (web)

Interface limits number of texts you can analyze

Doesn't allow you to see under the hood of the tools calculations

Voyant Server

Voyant Server is a version of Voyant that you can download and run locally on your computer. It's a Java-based application.

- Avoid problems of slow loading time and instability of web version
- Keep data private

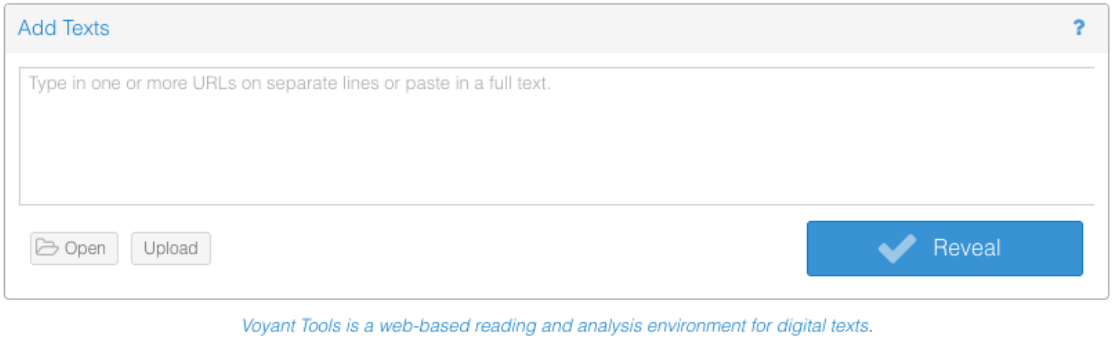
Installation instructions

Latest release download (must have Java downloaded)

VOYANT

see through your text

voyant-tools.org



Accepted file types

File types: plain text, HTML, XML, PDF, RTF, and MS Word

Zipped directories OK - (.zip, .tar, .tgz, etc.) containing documents in those formats.

Get the data

go.ncsu.edu/voyantfall2018

Download the “**fairy-tales-corpus.zip**” file

**Can’t access the link? Make sure you are logged into your NCSU gmail!

Getting started

Choose either 1 or 2:

1. Web-based version:

   Open a web browser, visit http://voyant-tools.org

2. Voyant Server (highly recommended!):

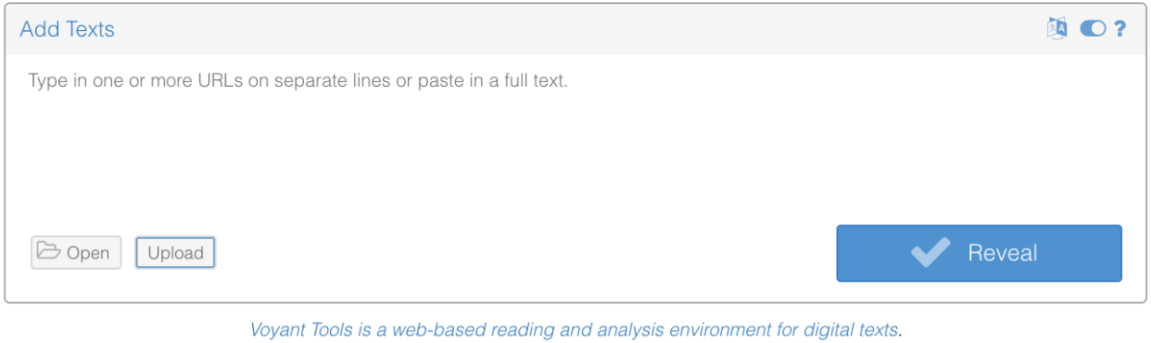
- download the VoyantServer 2.0 zip archive
- double-click on the zip archive to expand its contents
- double-click on VoyantServer.jar
    - on Mac, because of security restrictions on applications that aren't signed and approved by Apple, you may need to Ctrl-click on the VoyantServer.jar file, select open from the menu, and then click open (not the default button) in the next dialog box

Load data into Voyant

The data you load into Voyant is called a “corpus.” Three ways to load a corpus:

1. Paste text into the large box
2. Paste URLs into the large box, one URL per line
3. Click the upload button and select files to upload

Uploading corpus



Do Now:

Click **Upload** button and upload **fairy-tales-corpus.zip**

Voyant Tools

Cirrus | Terms | Links

Tab through different tools

Reader | TermsBerry

THUMBELINA

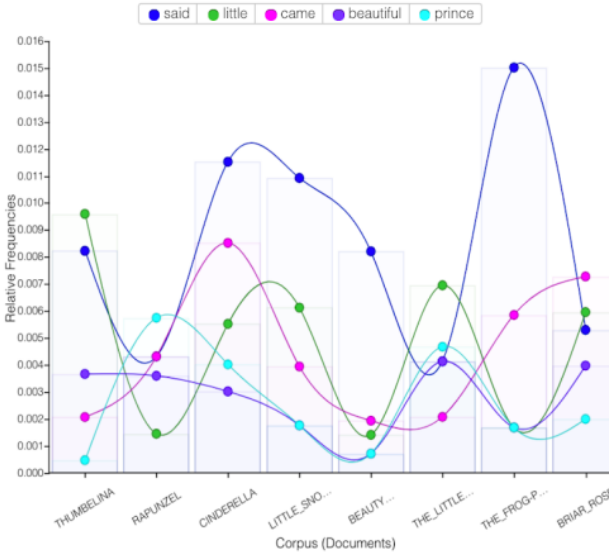
There was once a woman who wished very much to have a little child, but she could not obtain her wish. At last she went to a fairy, and said, "I should so very much like to have a little child; can you tell me where I can find one?"

"Oh, that can be easily managed," said the fairy. "Here is a barleycorn of a different kind to those which grow in the farmer's fields, and which the chickens eat; put it into a flower-pot, and see what will happen."

"Thank you," said the woman, and she gave the fairy twelve shillings, which was the price of the barleycorn. Then she went home and planted it, and immediately there grew up a large handsome flower, something like a tulip in appearance, but with its leaves tightly closed as if it were still a bud. "It is a beautiful flower," said the woman, and she kissed the red and golden-colored leaves, and while she did so the flower opened, and she could see that it was a real tulip. Within the flower, upon the green velvet stamens, sat a very delicate and graceful little maiden. She was scarcely half as long as a thumb, and they gave her the name of "Thumbelina," or Tiny, because she was so small. A walnut-shell, elegantly polished, served her for a cradle; her bed was formed of blue violet-leaves, with a rose-leaf for a counterpane. Here she slept at night, but during the day she amused herself on a table, where the woman had place[d] [a] plate were wreaths of flowers with their stems in the [wa]ter, the tulip-leaf, which served Tiny for a boat. Here th[e] [s]erself from side to side, with two oars made of whi[te] [h]orse hair, a pretty sight. Tiny could, also, sing so softly and sweetly that nothing like her singing had ever before been heard.

Select & Change Tools

Trends | Document Terms



Scale | Terms:

Reset | Display

Export

Help & Info

Summary | Documents | Phrases | Contexts | Correlations

This corpus has 8 documents with 27,732 total words and 2,909 unique word forms. Created now.

Document Length:
- Longest: THE_LITTLE_MERMAID (9223); BEAUTY AND THE BEAST (5733); THUMBELINA (4381); LITTLE_SNOW_WHITE (2290); CINDERELLA (1997)
- Shortest: THE_FROG-PRINCE (1199); RAPUNZEL (1395); BRIAR_ROSE (1514); CINDERELLA (1997); LITTLE_SNOW_WHITE (2290)

Vocabulary Density:
- Highest: RAPUNZEL (0.303); BRIAR_ROSE (0.301); THE_FROG-PRINCE (0.284); LITTLE_SNOW_WHITE (0.244); CINDERELLA (
- Lowest: THE_LITTLE_MERMAID (0.171); THUMBELINA (0.201); BEAUTY AND THE BEAST (0.202); CINDERELLA (0.223); LITTLE
(0.244)

Set Stopwords

Average Words Per Sentence:
- Highest: BRIAR_ROSE (36.0); THE_FROG-PRINCE (29.2); THE_LITTLE_MERMAID (27.8); CINDERELLA (25.9); RAPUNZEL (25.8)
- Lowest: THUMBELINA (23.2); LITTLE_SNOW_WHITE (24.4); BEAUTY AND THE BEAST (25.5); RAPUNZEL (25.8); CINDERELLA (25.9)

Most **frequent words** in the corpus: said (201); little (152); came (89); beautiful (81); prince (74)

Document	Left	Term	Right
1) THU...	went to a fairy, and	said	, "I should so very much
1) THU...	that can be easily managed,"	said	the fairy. "Here is a
1) THU...	what will happen." "Thank you,"	said	the woman, and she gave
1) THU...	It is a beautiful flower,"	said	the woman, and she kissed
1) THU...	would make for my son,"	said	the toad, and she took
1) THU...	loud, or she will wake,"	said	the toad, "and then she
1) THU...	her in the water, and	said	, "Here is my son, he
1) THU...	toad, and heard what she	said	, so they lifted their heads
1) THU...	turned up their feelers, and	said	, "She has only two legs
1) THU...	looks." "She has no feelers,"	said	another. "Her waist is quite
1) THU...	being." "Oh! she is ugly,"	said	all the lady cockchafers, although

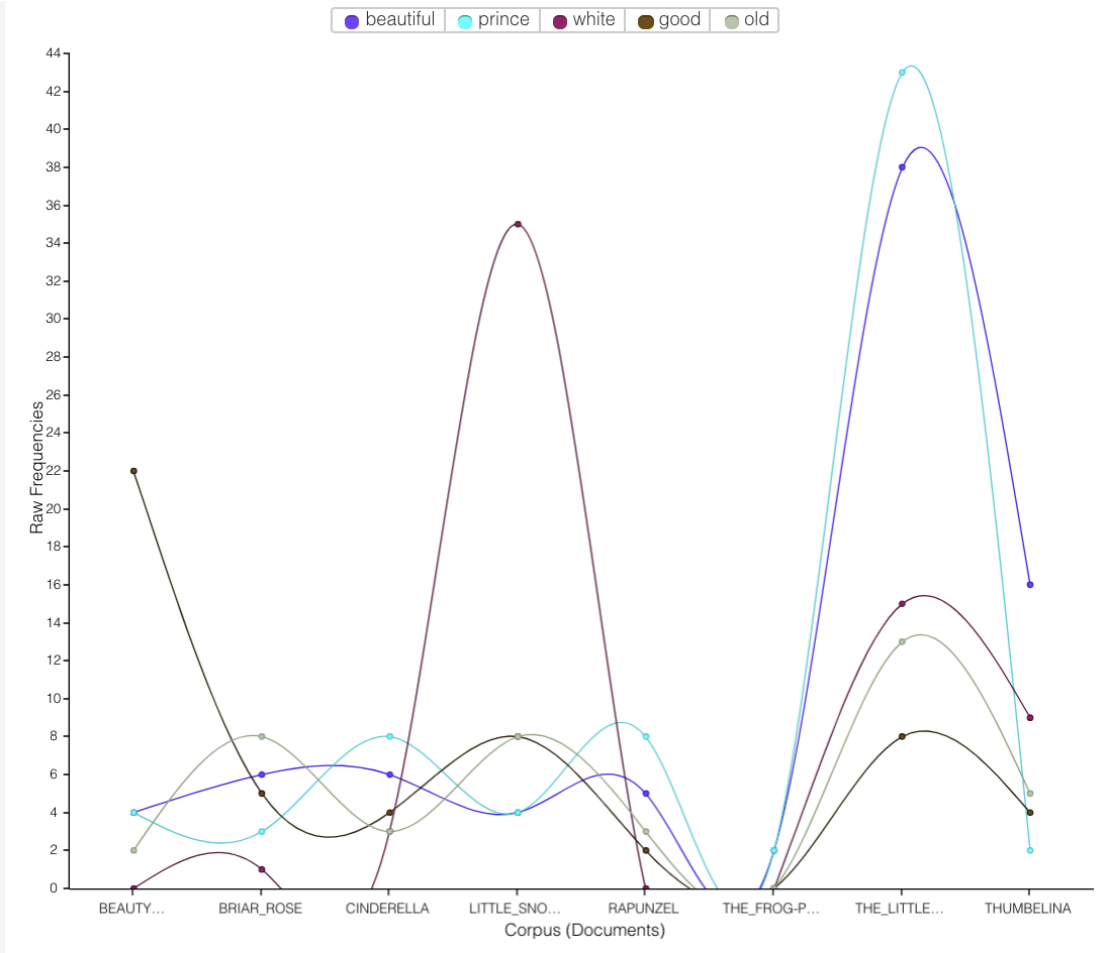
Start hands-on exploration of the corpus

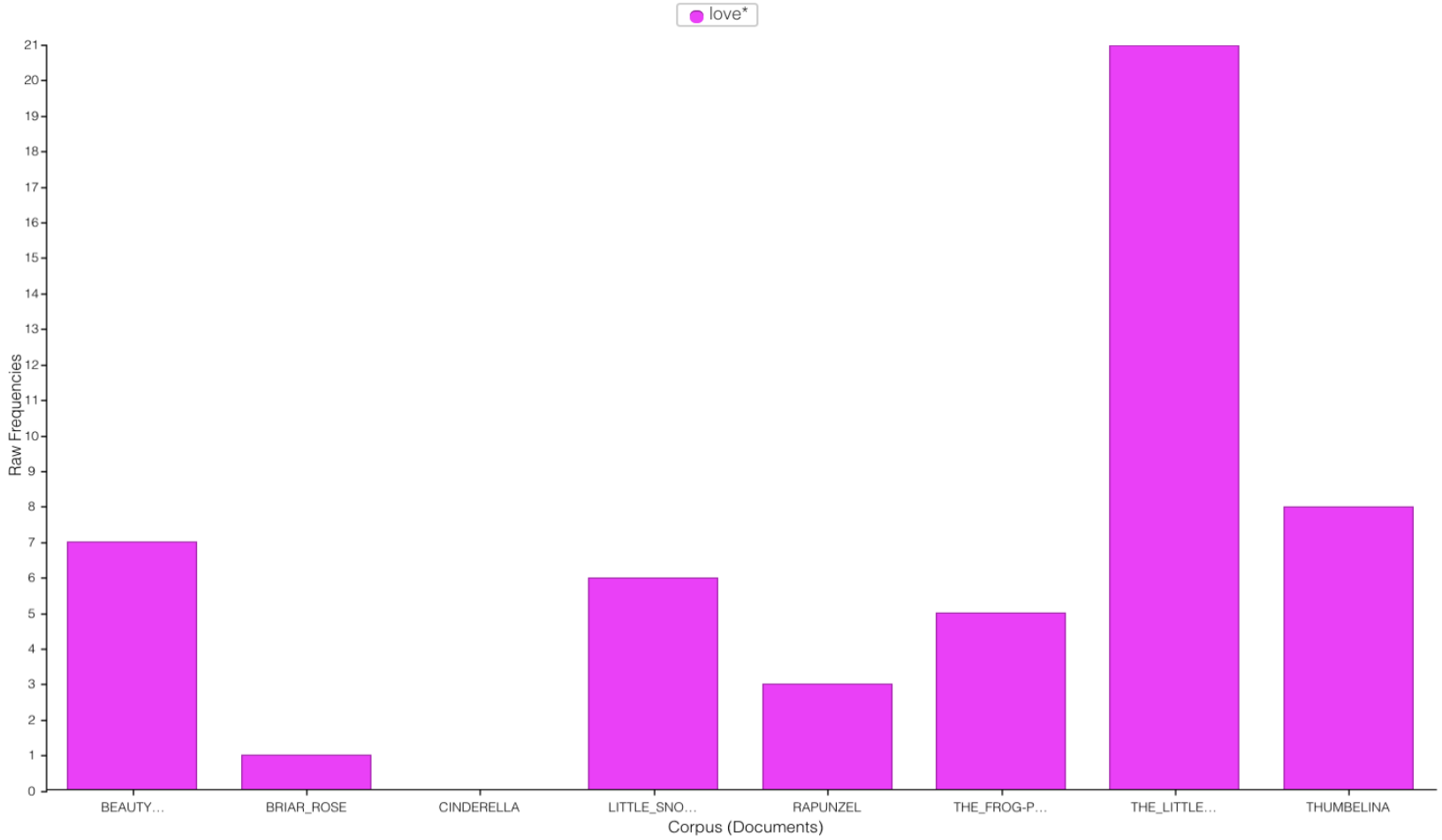
See Activity Guide handout

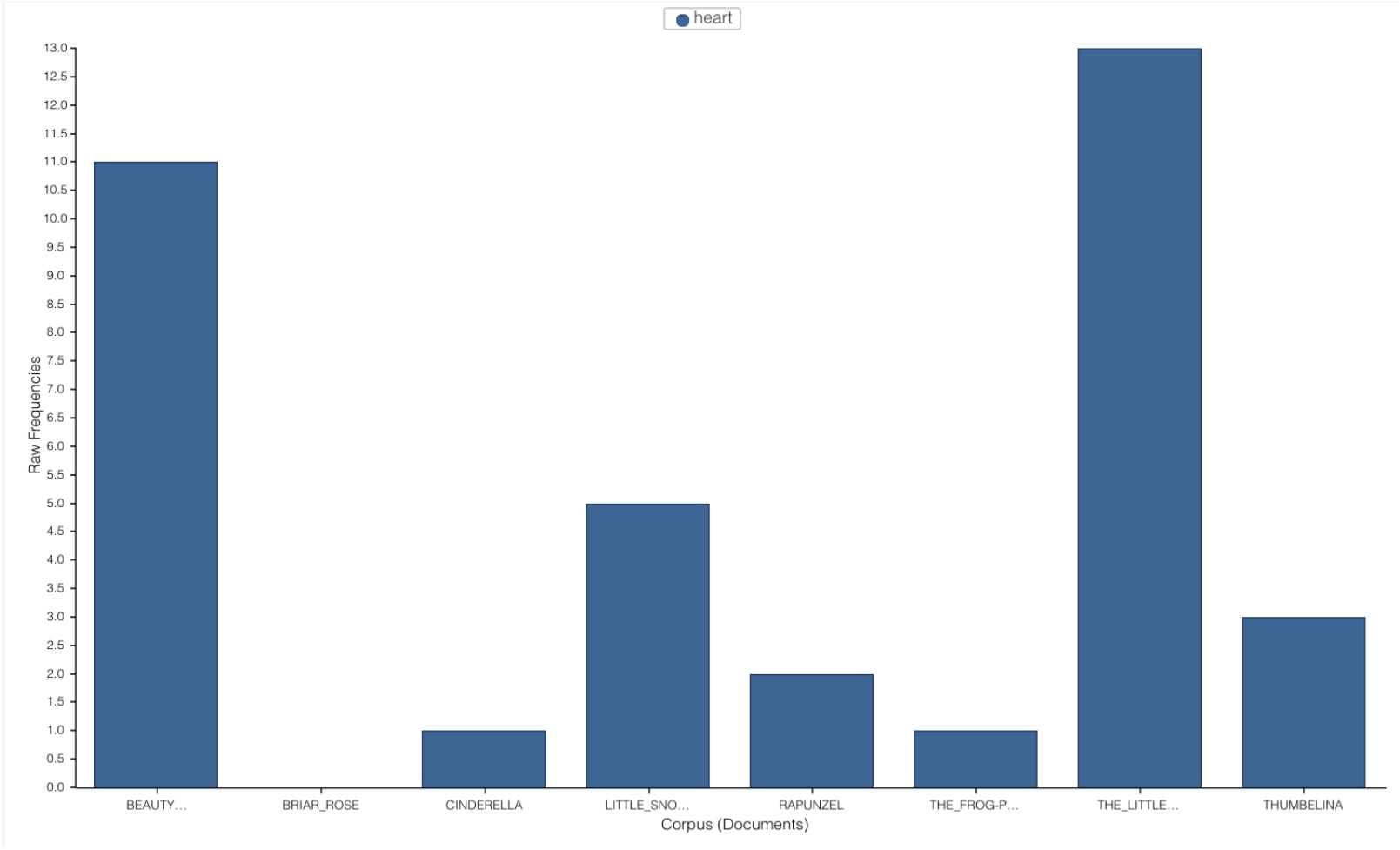
for self-directed activity

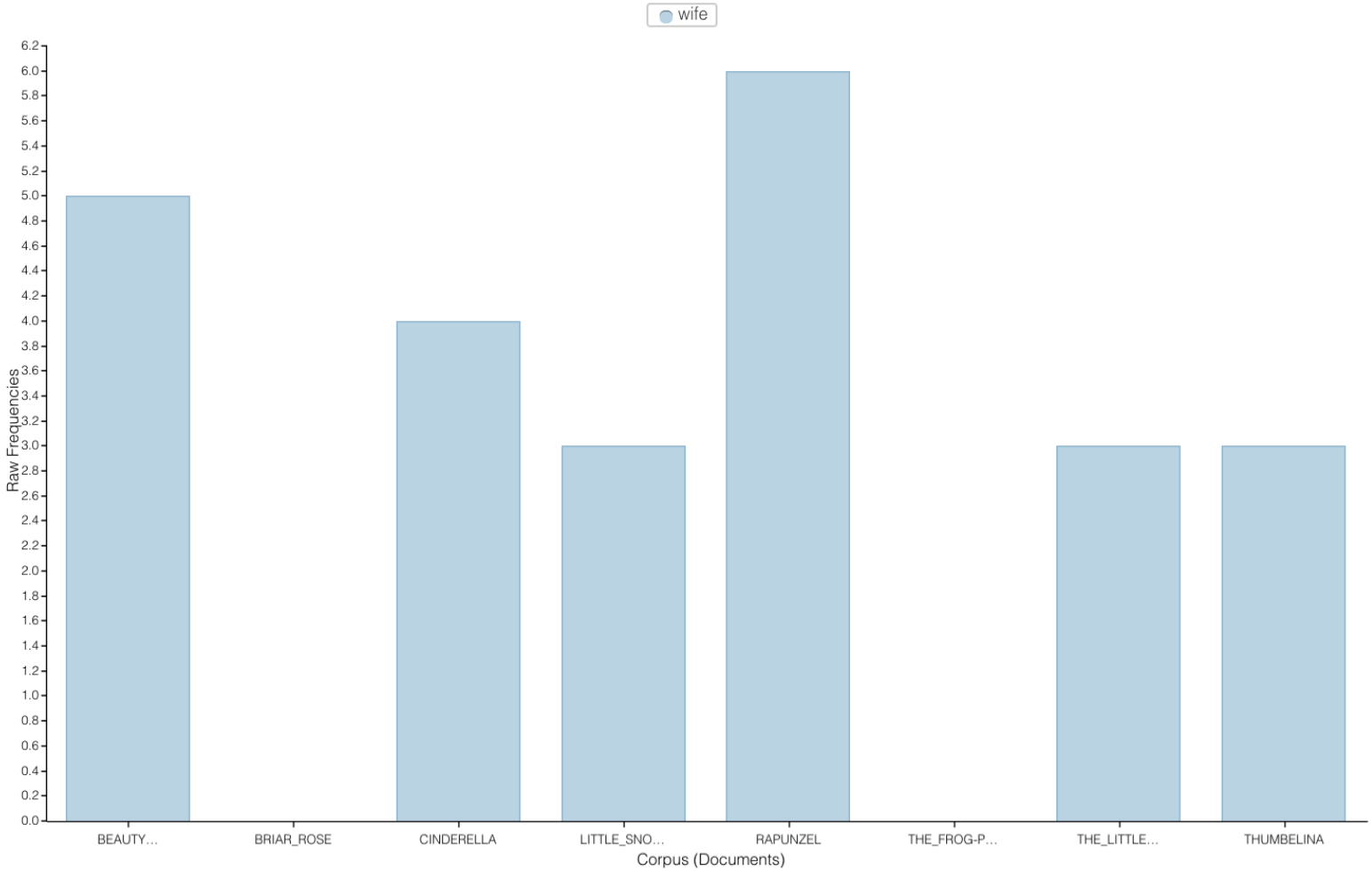
Mia and Erica are here to help!

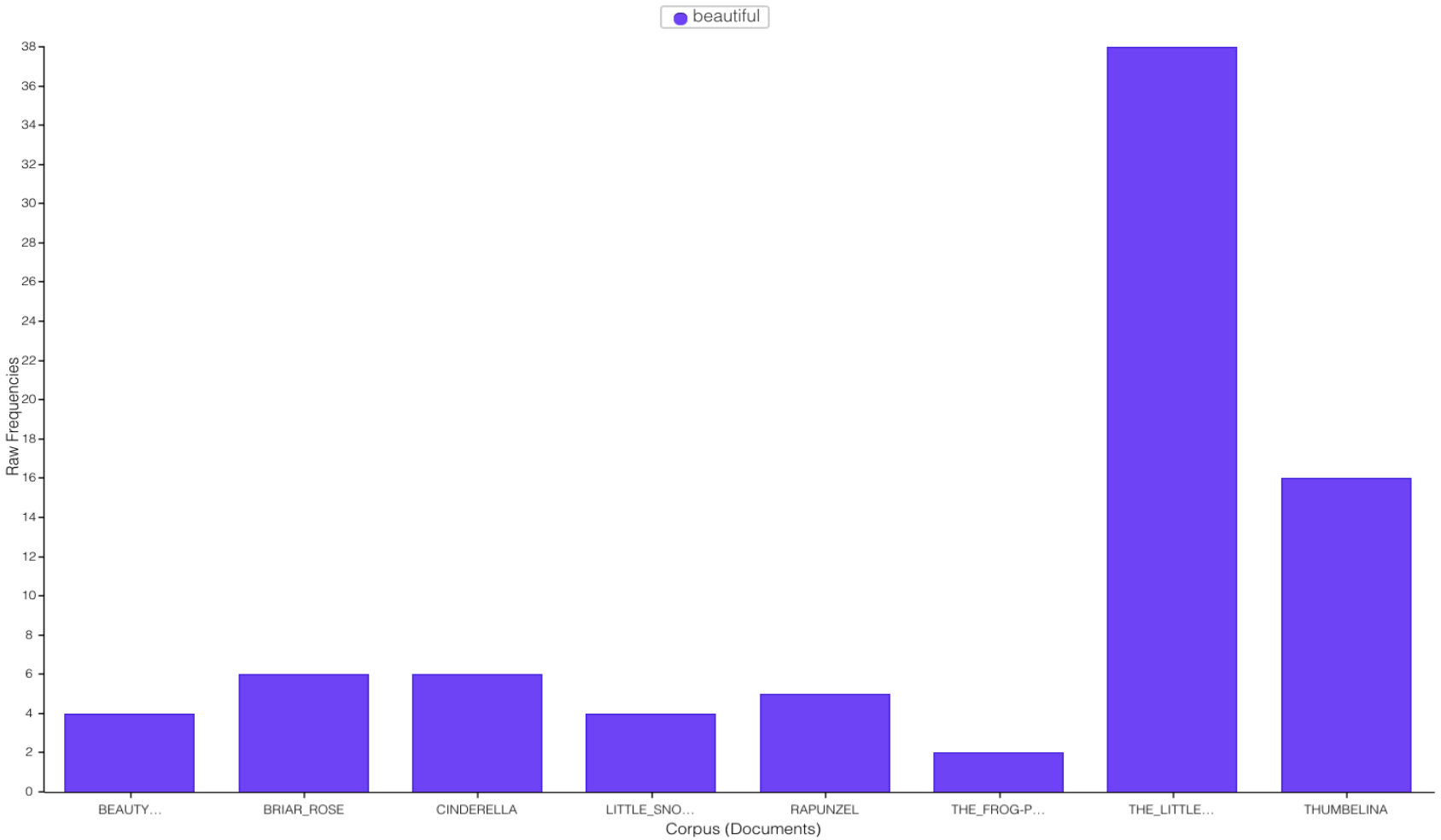
Raise a hand if you need help.


beautiful
prince
white
good
old
Raw Frequencies
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
BEAUTY...
BRIAR_ROSE
CINDERELLA
LITTLE_SNO...
RAPUNZEL
THE_FROG-P...
THE_LITTLE...
THUMBELINA
Corpus (Documents)

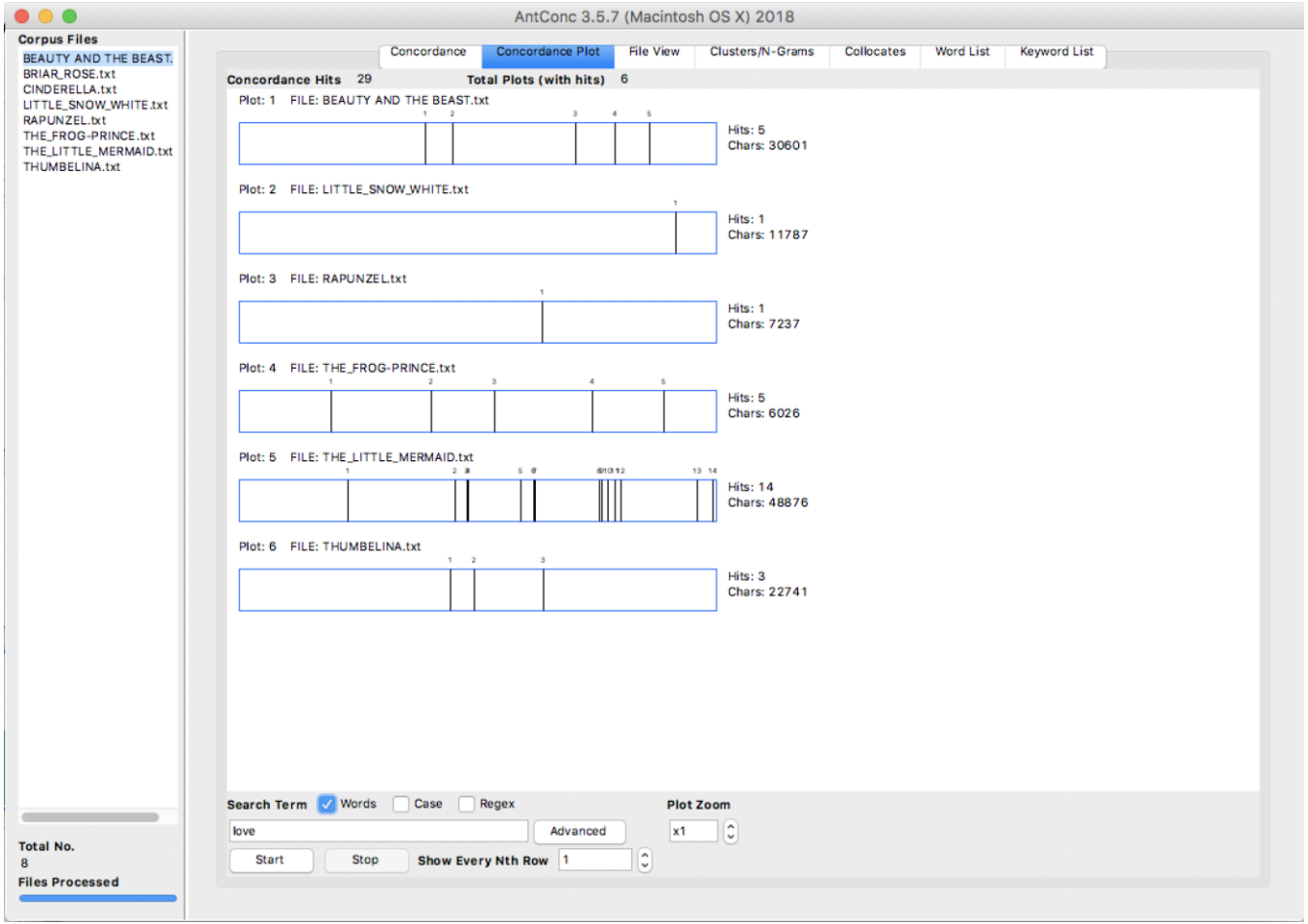

love*
Raw Frequencies
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
BEAUTY…
BRIAR_ROSE
CINDERELLA
LITTLE_SNO…
RAPUNZEL
THE_FROG-P…
THE_LITTLE…
THUMBELINA
Corpus (Documents)

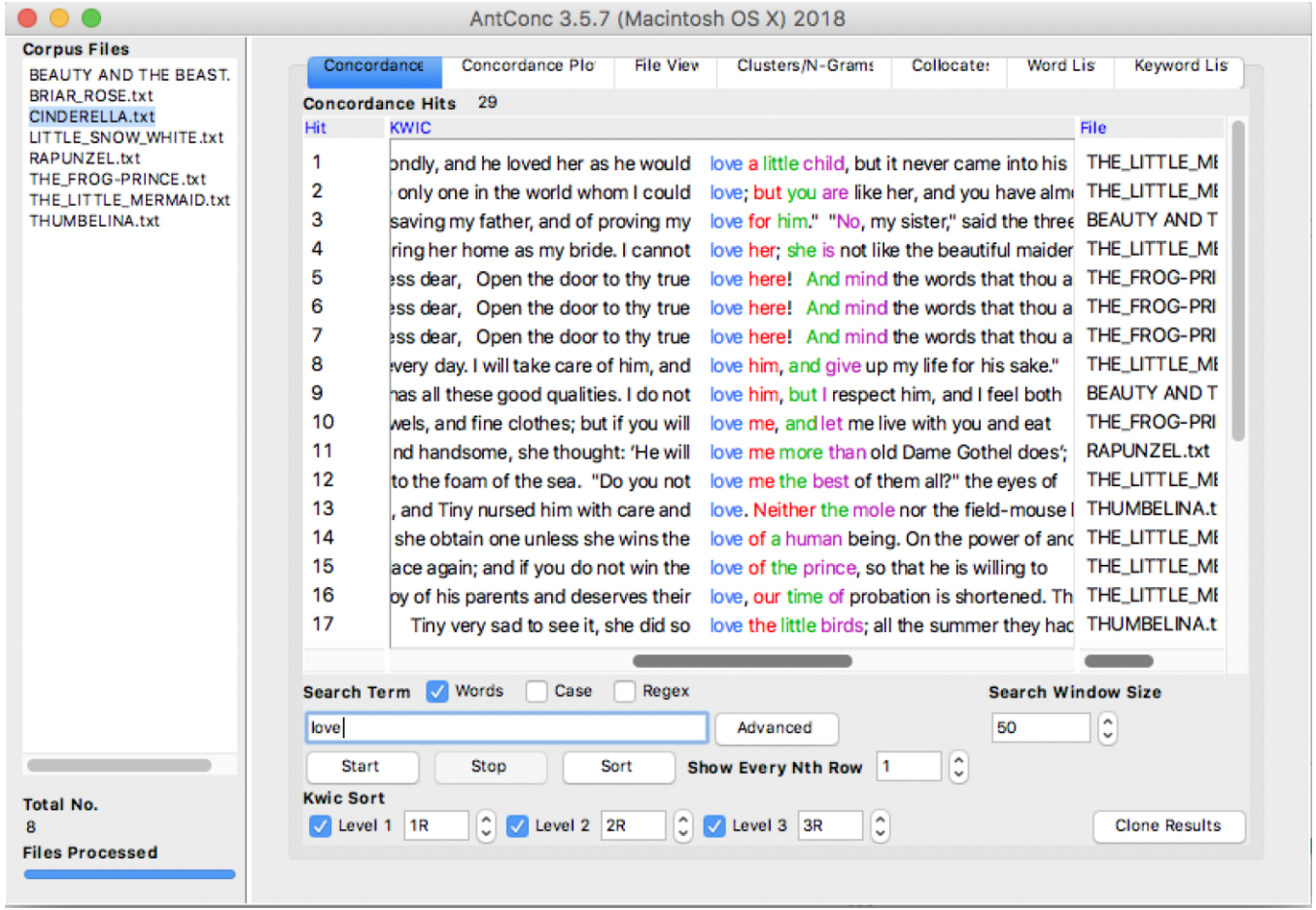

heart
Raw Frequencies
13.0
12.5
12.0
11.5
11.0
10.5
10.0
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
BEAUTY…
BRIAR_ROSE
CINDERELLA
LITTLE_SNO…
RAPUNZEL
THE_FROG-P…
THE_LITTLE…
THUMBELINA
Corpus (Documents)


wife
Raw Frequencies
6.2
6.0
5.8
5.6
5.4
5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
2.8
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
BEAUTY...
BRIAR_ROSE
CINDERELLA
LITTLE_SNO...
RAPUNZEL
THE_FROG-P...
THE_LITTLE...
THUMBELINA
Corpus (Documents)


beautiful
Raw Frequencies
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
BEAUTY…
BRIAR_ROSE
CINDERELLA
LITTLE_SNO…
RAPUNZEL
THE_FROG-P…
THE_LITTLE…
THUMBELINA
Corpus (Documents)

AntConc


AntConc 3.5.7 (Macintosh OS X) 2018
Corpus Files
BEAUTY AND THE BEAST.
BRIAR_ROSE.txt
CINDERELLA.txt
LITTLE_SNOW_WHITE.txt
RAPUNZEL.txt
THE_FROG-PRINCE.txt
THE_LITTLE_MERMAID.txt
THUMBELINA.txt
Total No.
8
Files Processed
Concordance
Concordance Plot
File View
Clusters/N-Grams
Collocates
Word List
Keyword List
Concordance Hits 29
Total Plots (with hits) 6
Plot: 1 FILE: BEAUTY AND THE BEAST.txt
Hits: 5
Chars: 30601
Plot: 2 FILE: LITTLE_SNOW_WHITE.txt
Hits: 1
Chars: 11787
Plot: 3 FILE: RAPUNZEL.txt
Hits: 1
Chars: 7237
Plot: 4 FILE: THE_FROG-PRINCE.txt
Hits: 5
Chars: 6026
Plot: 5 FILE: THE_LITTLE_MERMAID.txt
Hits: 14
Chars: 48876
Plot: 6 FILE: THUMBELINA.txt
Hits: 3
Chars: 22741
Search Term
Words
Case
Regex
love
Advanced
Start
Stop
Show Every Nth Row
1
Plot Zoom
x1


AntConc 3.5.7 (Macintosh OS X) 2018
Corpus Files
BEAUTY AND THE BEAST.
BRIAR_ROSE.txt
CINDERELLA.txt
LITTLE_SNOW_WHITE.txt
RAPUNZEL.txt
THE_FROG-PRINCE.txt
THE_LITTLE_MERMAID.txt
THUMBELINA.txt
Total No.
8
Files Processed
Concordance
Concordance Plo
File View
Clusters/N-Grams
Collocates
Word Lis
Keyword Lis
Concordance Hits 29
Hit
KWIC
File
1 ondly, and he loved her as he would love a little child, but it never came into his THE_LITTLE_ME
2 only one in the world whom I could love; but you are like her, and you have alm THE_LITTLE_ME
3 saving my father, and of proving my love for him." "No, my sister," said the three BEAUTY AND T
4 ring her home as my bride. I cannot love her; she is not like the beautiful maider THE_LITTLE_ME
5 ess dear, Open the door to thy true love here! And mind the words that thou a THE_FROG-PRI
6 ess dear, Open the door to thy true love here! And mind the words that thou a THE_FROG-PRI
7 ess dear, Open the door to thy true love here! And mind the words that thou a THE_FROG-PRI
8 very day. I will take care of him, and love him, and give up my life for his sake." THE_LITTLE_ME
9 has all these good qualities. I do not love him, but I respect him, and I feel both BEAUTY AND T
10 wels, and fine clothes; but if you will love me, and let me live with you and eat THE_FROG-PRI
11 nd handsome, she thought: 'He will love me more than old Dame Gothel does'; RAPUNZEL.txt
12 to the foam of the sea. "Do you not love me the best of them all?" the eyes of THE_LITTLE_ME
13 , and Tiny nursed him with care and love. Neither the mole nor the field-mouse l THUMBELINA.t
14 she obtain one unless she wins the love of a human being. On the power of and THE_LITTLE_ME
15 ace again; and if you do not win the love of the prince, so that he is willing to THE_LITTLE_ME
16 oy of his parents and deserves their love, our time of probation is shortened. Th THE_LITTLE_ME
17 Tiny very sad to see it, she did so love the little birds; all the summer they hac THUMBELINA.t
Search Term
Words
Case
Regex
love
Advanced
Search Window Size
50
Start
Stop
Sort
Show Every Nth Row
1
Kwic Sort
Level 1
1R
Level 2
2R
Level 3
3R
Clone Results

Online evaluation

go.ncsu.edu/libeval

Resources

- Voyant Tools Documentation
- Voyant Tools Guide
- Digging through the Hillary Clinton's Archive with Voyant Tools
- Searching for the Victorians
- Ted Underwood's Where to Start with Text Mining
- AntConc Tutorial Videos
- The Programming Historian's Corpus Analysis with AntConc