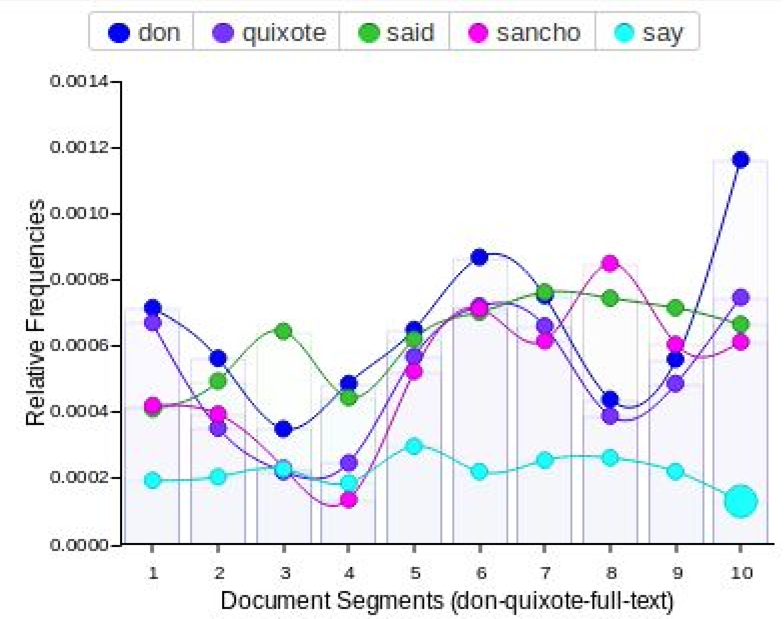
Quickstart Guide to Voyant

The McGraw Center for Teaching and Learning

Voyant (https://voyant-tools.org) is a free, online tool for performing text analysis using a wide variety of included tools. *Text analysis* is a very general term, but in this context refers to the use of software tools to generate and visualize statistics about how and how frequently words are used in a text or in a series of texts (a *corpus*). There are many other tools for performing analyses of text in different ways, such as Juxta for the comparison of two differing versions of a text, or MALLET for the classification of texts into semantic categories. Voyant belongs to a type of software that might be called *concordancers*, software that count the occurrence of words, and provides information about the co-location of words within a text. Text analysis tools are not meant to do the job of interpreting text, but rather to uncover trends and patterns of language use in text.

Preparing your corpus

The way in which you structure your corpus has a great impact on the types of information you can derive from text analysis. Texts that you submit to Voyant should preferably be in plain text (.txt) format and segmented in a way that will allow you to interpret the results in the most effective way. For example, books are often segmented into chapters. Chapters are often segmented into paragraphs, and so forth. But, in terms of analysing the text, chapters might not be the most useful segmentation. Chapters might, for example, vary widely in their length. This variation could overemphasize the longer paragraphs and underemphasize the shorter ones. One can imagine a line graphs showing the frequency of the occurrence of a word throughout a text. This sort of graph is actually very common in text analysis software. The Y axis of such a graph shows the number of times the word appears. But what about the X axis? By default, commonly, the X axis in such a graph is an even segmentation of text. In other words, the software splits the text evenly into 10 segments, each with roughly the same number of words. But a graph such as this is meant to be a *diachronic* graph. The graph is in essence a timeline through the text. What if these segments actually had chronological significance? Works such as diaries, newspaper articles, and journals, are very often subjects for automated text analysis and most definitely have a chronological component that might be useful in that analysis. In these cases, a diachronic graph could show you changes in word usage across the publication history of a journal or across the author's entire lifetime. Similarly if all the books written by a particular author were entered into Voyant, it might be useful to not only segment each book in the corpus, but also to sequence those books by publication date, so that one gets a proper sense of how words were used over time. Another

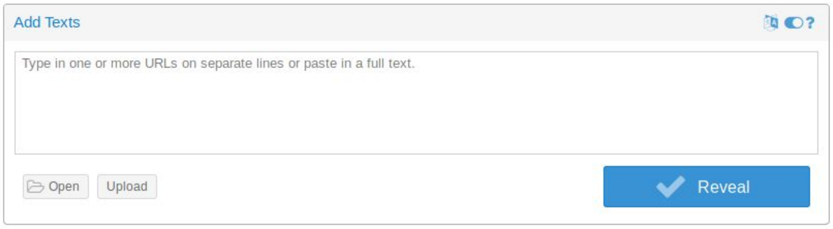


way to segment a text is by chapter. Chapters are, after all, almost never simple arbitrary divisions of the text. The author is showing some intent through the way in which he or she segments the text or tells the story in chapters. It is worth paying attention to what you think the author's reasoning was in structuring the text and perhaps to use this to make an analysis of the text more valuable and informed.

NOTE: if you are getting your text from gutenberg.org, which is of course an excellent source of text, look closely at the texts you copy. The texts always contain a lengthy legal statement at the end of the book which you most likely to do not want to include in your analysis, and often will include a foreword, which may not written by the author or even contemporaneous with the book.

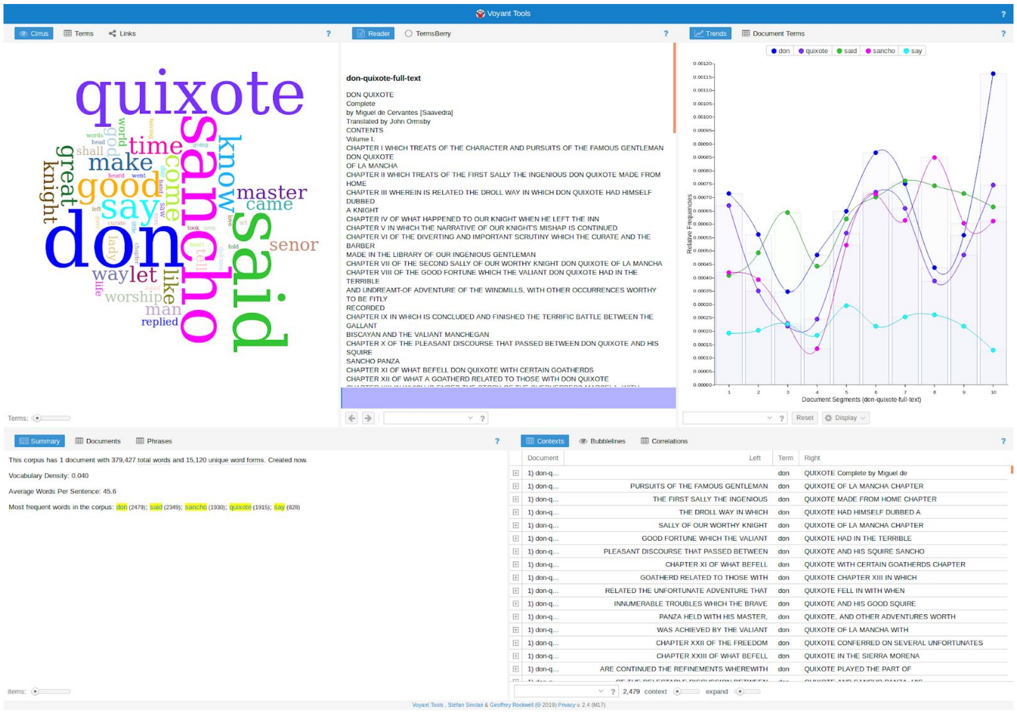
Uploading your corpus

The homepage of Voyant (https://voyant-tools.org/), allows you to upload your corpus files. (Note that there are also several predefined corpora available by clicking the **Open** button.) The Upload button allows you to select one or more files from your computer for analysis. After you have selected your files, Click the **Reveal** button to enter to main Voyant interface.


Add Texts
Type in one or more URLs on separate lines or paste in a full text.
Open
Upload
Reveal

The interface

Voyant contains many tools. Many tools do roughly the same thing but display the results in different ways. This can be very useful and is one of the real reason for using text analysis software in the first place - to look at texts in new ways. The main interface, by default, consists of five windows, each displaying a different tool. Hovering your cursor over the top section of any of these windows displays some tools for changing the options associated with the current tool and for displaying information about the tool. The top section of the entire screen also includes some tools, including, importantly, a **Home** button for returning to the homepage and beginning a new analysis.



don-quixote-full-text

DON QUIXOTE
Complete
by Miguel de Cervantes [Saavedra]
Translated by John Ormsby
CONTENTS
Volume I.
CHAPTER I WHICH TREATS OF THE CHARACTER AND PURSUITS OF THE FAMOUS GENTLEMAN
DON QUIXOTE
OF LA MANCHA
CHAPTER II WHICH TREATS OF THE FIRST SALLY THE INGENIOUS DON QUIXOTE MADE FROM
HOME
CHAPTER III WHEREIN IS RELATED THE DROLL WAY IN WHICH DON QUIXOTE HAD HIMSELF
DUBBED
A KNIGHT
CHAPTER IV OF WHAT HAPPENED TO OUR KNIGHT WHEN HE LEFT THE INN
CHAPTER V IN WHICH THE NARRATIVE OF OUR KNIGHT'S MISHAP IS CONTINUED
CHAPTER VI OF THE DIVERTING AND IMPORTANT SCRUTINY WHICH THE CURATE AND THE
BARBER
MADE IN THE LIBRARY OF OUR INGENIOUS GENTLEMAN
CHAPTER VII OF THE SECOND SALLY OF OUR WORTHY KNIGHT DON QUIXOTE OF LA MANCHA
CHAPTER VIII OF THE GOOD FORTUNE WHICH THE VALIANT DON QUIXOTE HAD IN THE
TERRIBLE
AND UNDREAMT-OF ADVENTURE OF THE WINDMILLS, WITH OTHER OCCURRENCES WORTHY
TO BE FITLY
RECORDED
CHAPTER IX IN WHICH IS CONCLUDED AND FINISHED THE TERRIFIC BATTLE BETWEEN THE
GALLANT
BISCAYAN AND THE VALIANT MANCHEGAN
CHAPTER X OF THE PLEASANT DISCOURSE THAT PASSED BETWEEN DON QUIXOTE AND HIS
SQUIRE
SANCHO PANZA
CHAPTER XI OF WHAT BEFELL DON QUIXOTE WITH CERTAIN GOATHERDS
CHAPTER XII OF WHAT A GOATHERD RELATED TO THOSE WITH DON QUIXOTE

This corpus has 1 document with 379,427 total words and 15,120 unique word forms. Created now.

Vocabulary Density: 0.040

Average Words Per Sentence: 45.6

Most frequent words in the corpus: don (2479); said (2349); sancho (1930); quixote (1915); say (828)

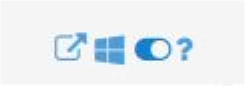
Document	Left	Term	Right
1) don-q...		don	QUIXOTE Complete by Miguel de
1) don-q...	PURSUITS OF THE FAMOUS GENTLEMAN	don	QUIXOTE OF LA MANCHA CHAPTER
1) don-q...	THE FIRST SALLY THE INGENIOUS	don	QUIXOTE MADE FROM HOME CHAPTER
1) don-q...	THE DROLL WAY IN WHICH	don	QUIXOTE HAD HIMSELF DUBBED A
1) don-q...	SALLY OF OUR WORTHY KNIGHT	don	QUIXOTE OF LA MANCHA CHAPTER
1) don-q...	GOOD FORTUNE WHICH THE VALIANT	don	QUIXOTE HAD IN THE TERRIBLE
1) don-q...	PLEASANT DISCOURSE THAT PASSED BETWEEN	don	QUIXOTE AND HIS SQUIRE SANCHO
1) don-q...	CHAPTER XI OF WHAT BEFELL	don	QUIXOTE WITH CERTAIN GOATHERDS CHAPTER
1) don-q...	GOATHERD RELATED TO THOSE WITH	don	QUIXOTE CHAPTER XIII IN WHICH
1) don-q...	RELATED THE UNFORTUNATE ADVENTURE THAT	don	QUIXOTE FELL IN WITH WHEN
1) don-q...	INNUMERABLE TROUBLES WHICH THE BRAVE	don	QUIXOTE AND HIS GOOD SQUIRE
1) don-q...	PANZA HELD WITH HIS MASTER,	don	QUIXOTE, AND OTHER ADVENTURES WORTH
1) don-q...	WAS ACHIEVED BY THE VALIANT	don	QUIXOTE OF LA MANCHA WITH
1) don-q...	CHAPTER XXII OF THE FREEDOM	don	QUIXOTE CONFERRED ON SEVERAL UNFORTUNATES
1) don-q...	CHAPTER XXIII OF WHAT BEFELL	don	QUIXOTE IN THE SIERRA MORENA
1) don-q...	ARE CONTINUED THE REFINEMENTS WHEREWITH	don	QUIXOTE PLAYED THE PART OF

2,479

Summary and Cirrus Text Cloud

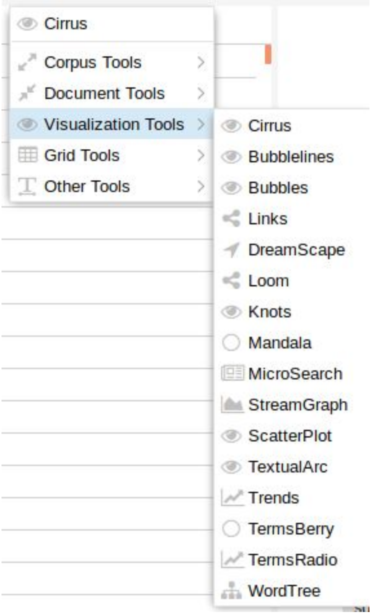
Starting in the bottom left, the Summary window displays statistical information about your corpus, including the total number of words, unique forms of words, and vocabulary density, a ratio of the total number of unique words to the total number of words. Notice that they use the phrase ‘unique word forms’. Voyant performs process called tokenization. A token is the equivalent of a string of characters separated from other strings by spaces. So, in essence, words. After the software has done this splitting, it counts how many words it found and then proceeds to remove duplicates so that it can count the number of unique words. It does not however have any ‘knowledge’ of different forms of the same word. The word ‘house’ is completely different that the word ‘houses’. There is a process called ‘stemming’ that attempts to convert all variant forms of words to a common, base forms, but Voyant does not include these tools. Keep in mind that as you use Voyant, you may want to consider related variant forms of the words you are search for.

Stopwords



During this process of counting words, Voyant ignores several very common words, such as ‘the’, ‘a’, and ‘you’. Because these words tend to be som commons, if they were not ignored, they would in almost every case constitute the most frequent words, obscuring those words that might have more impact. Voaynt uses what is called a **Stopwords** list. This list can be examined and edited by selecting the option icon in the header region of the Summary tool.

Cirrus and Terms Tools



Continuing in a counter-clockwise direction, the **Cirrus** tool, just above the Summary window, displays a wordcloud of the most frequent words in the text(s). Hovering your cursor over any of the words in this cloud tells you how many times this particular word appeared in the text(s). The neighboring **Terms** tool does the same thing, but in list view, sorting the words from most frequent at the top of the list.

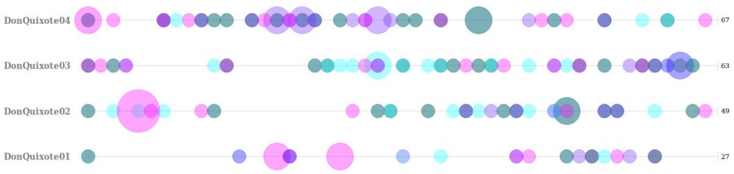
Changing tools

Voyant includes dozens of tools. Any window can be switched to another tool at any time. Hover your mouse over the top section of the top-left window and select the second, window icon. The resulting menu allows you to select a new tool to be added to this window. The *TermsBerry* tool, for example, found under Visualization Tools, displays a similar term frequency count but in a different way.

Trends

The top-right window displays, by default, one of the most used tools in Voyant, the **Trends** tool. This tool displays the diachronic graph mentioned previously. Each line on the graph represents a word. The y axis represents the relative frequency of that word and the x-axis, the segment of text. The word displayed with this tool can be changed either by clicking on a word in one of the other tools on the screen, or by searching for words in the search box provided just below the tool. Pay particular attention to the document sections along the x-axis. If you had uploaded a single text file, these sections would simply be the entire text split into ten segments. If you had uploaded several text files, each columns should correspond to one of the text files. Be careful however, because the ordering of these columns mirrors the sequence in which text files were uploaded. If the files were not uploaded correctly, the information in this tool can be misleading.

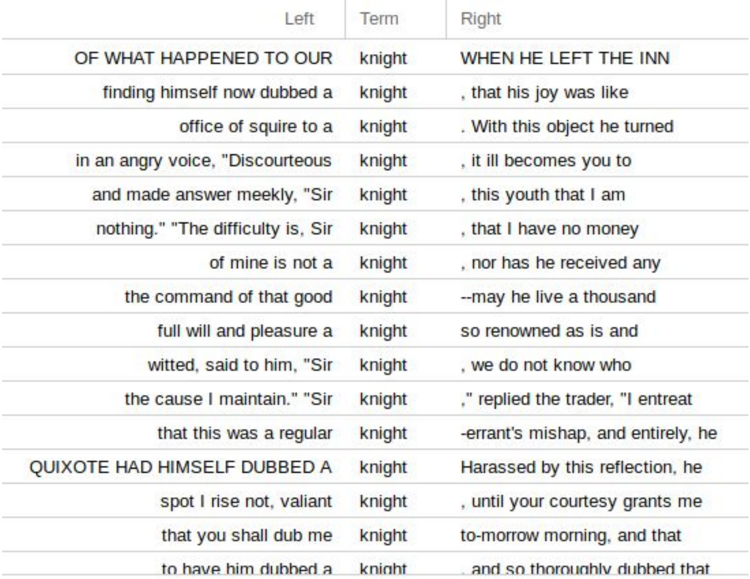
Bubblelines



A tool similar to the Trends tool is **Bubblelines**. Use the tool selection option in one of the windows to bring up this tool. Bubblelines shows each text in the corpus as a separate line and displays colored circles representing the relative occurrence of words.

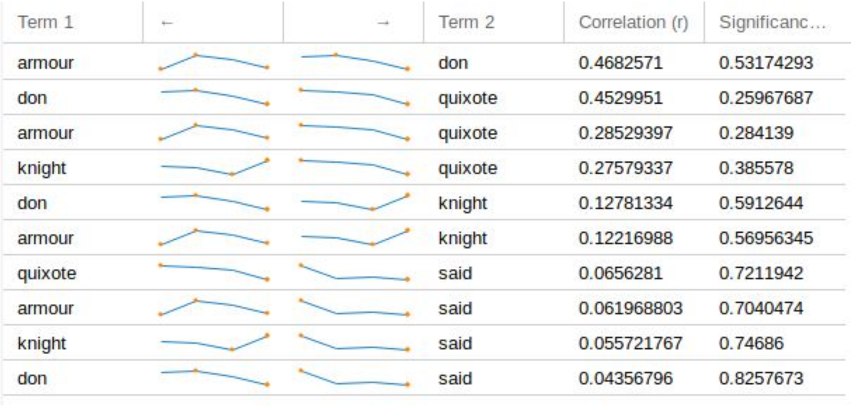
Getting to some meaning: Contexts and Correlations

So far we have been looking at different of analyzing the frequency of the use of words in the text. We have even seen how the frequency of changes over time through the course of a text. This sort of analysis is useful, but Voyant also allows us to compare, or correlate, the use of words through the text. What are the words that tend to be used within the same contexts? What words occur frequently together? The **Contexts** and **Correlations** tools are useful in answering these questions. By default the bottom-right window will show the Contexts tool. This tool, often called a Keyword in context (KWIC) tool is perhaps the oldest method for analyzing text. The tool displays each occurence of a single word, surrounded to the left and right by its context. The column to the left or the right of the target word can be sorted alphabetically to quickly get a sense for what words tend to precede or follow the target word.



Left	Term	Right
OF WHAT HAPPENED TO OUR	knight	WHEN HE LEFT THE INN
finding himself now dubbed a	knight	, that his joy was like
office of squire to a	knight	. With this object he turned
in an angry voice, "Discourteous	knight	, it ill becomes you to
and made answer meekly, "Sir	knight	, this youth that I am
nothing." "The difficulty is, Sir	knight	, that I have no money
of mine is not a	knight	, nor has he received any
the command of that good	knight	--may he live a thousand
full will and pleasure a	knight	so renowned as is and
witted, said to him, "Sir	knight	, we do not know who
the cause I maintain." "Sir	knight	," replied the trader, "I entreat
that this was a regular	knight	-errant's mishap, and entirely, he
QUIXOTE HAD HIMSELF DUBBED A	knight	Harassed by this reflection, he
spot I rise not, valiant	knight	, until your courtesy grants me
that you shall dub me	knight	to-morrow morning, and that
to have him dubbed a	knight	, and so thoroughly dubbed that

The **Correlations** tool is a statistical equivalent of this, displaying the pairs of words that display some statistical proximity. Notice that this, just as with all other tools in Voyant, is very much dependent upon the stopwords defined previously. In the example to the right, the word ‘said’ is displayed prominently. If this word is not important to you, you might want to redefine your stopwords list and try the tool again.



Term 1	←	→	Term 2	Correlation (r)	Significanc…
armour			don	0.4682571	0.53174293
don			quixote	0.4529951	0.25967687
armour			quixote	0.28529397	0.284139
knight			quixote	0.27579337	0.385578
don			knight	0.12781334	0.5912644
armour			knight	0.12216988	0.56956345
quixote			said	0.0656281	0.7211942
armour			said	0.061968803	0.7040474
knight			said	0.055721767	0.74686
don			said	0.04356796	0.8257673

In conclusion

As you have seen, Voyant includes many tools and this quickstart guide has only touched on a few of them. Try them out and find out what they do. Text analysis tools such as Voyant are not so much intended to give you answers as they are to facilitate the asking of new questions. Viewing a text in a multitude of ways, including seemingly dry statistical ways, might spark some ideas about ways in which you might re-read or reconsider the text.