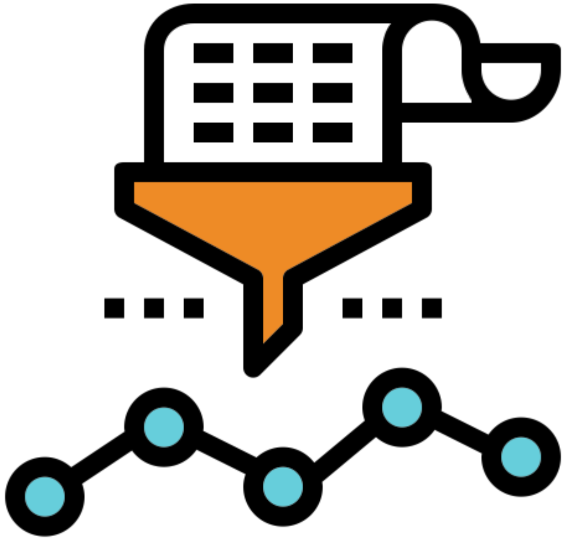
Text Data Mining
For Business Decisions

Module 4

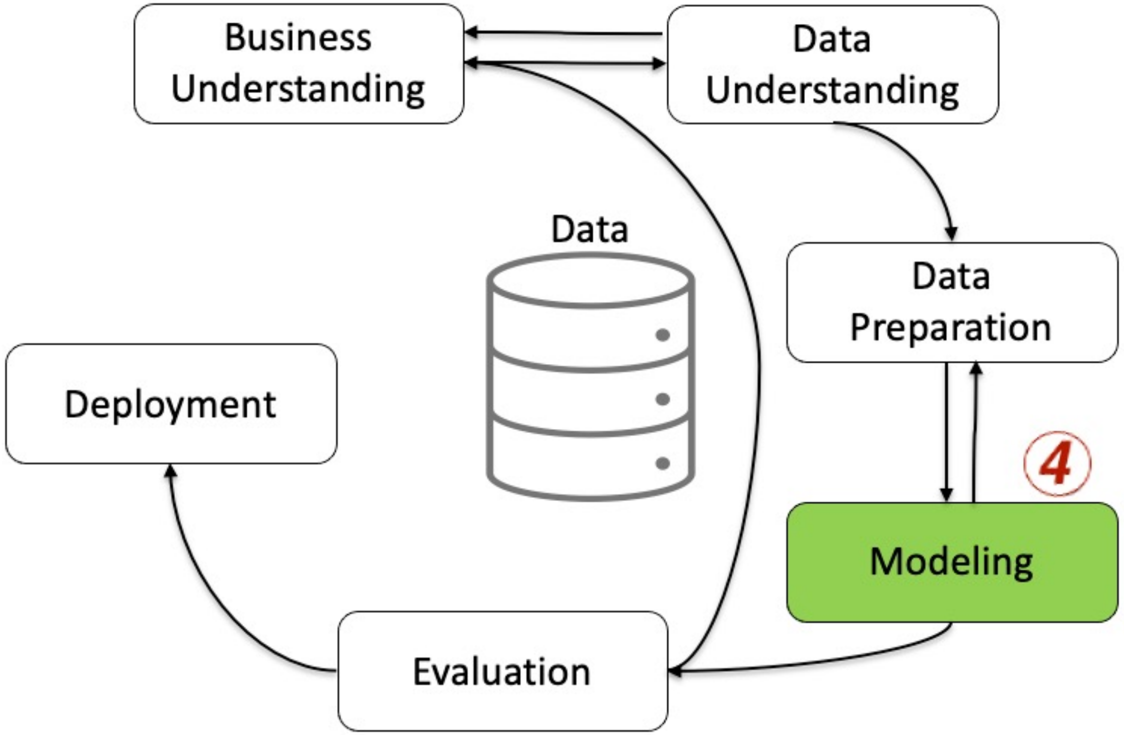
Word Frequency Analysis

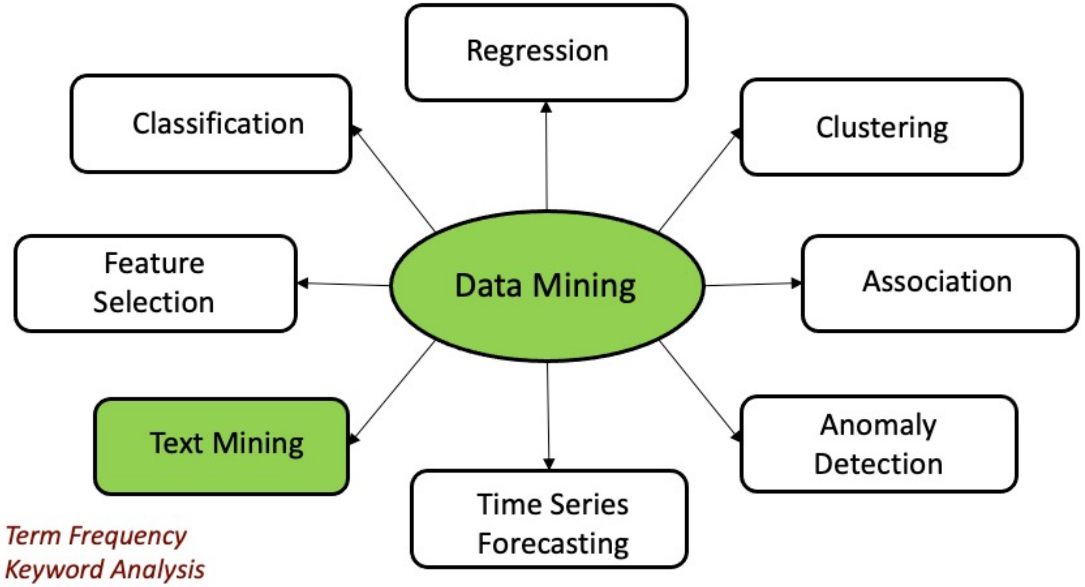
How do we discover the most frequent words in text data?



FORTINO
Global Education

Data Mining- Continuing Model-Making

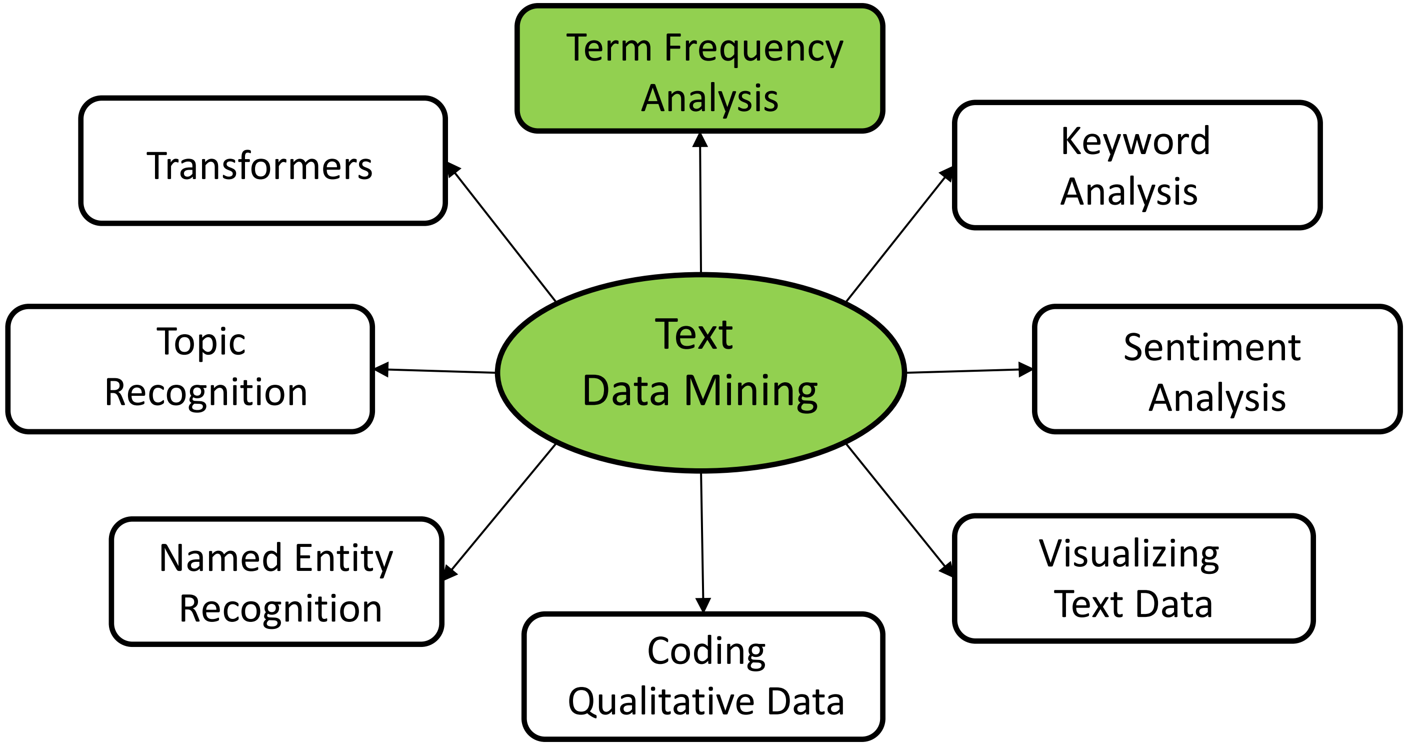




Term Frequency
Keyword Analysis
Sentiment Analysis
Similarity Scoring

FORTINO
Global Education

Term Frequency Analysis is one of the Most Basic Text Mining Algorithms



How do we count words?

- Counting words is the most basic technique in text analysis.
  - The starting point for most investigations. It is called *lexical analysis*
- Assume that the most frequently-appearing words hold some meaning: they are somewhat more important that other words.
  - We tabulate their frequency because they likely have some greater significance in the context of our text.
  - We ignore nuances, such as grammatical structure (for example, is the word a noun or a verb?) and how a term is used (Sarcasm? Irony? Fact?). – that's called *semantic analysis*
  - We also know that not all words carry significant meaning, such as propositions or conjunctions (for example, "and," "beside," "at," or "in.")
  - We assume we can safely ignore these and remove them from the text using a list of words with less meaning (stopwords).
  - We strive to create a Bag-of-Words text data file (document) or text data field (data point in a cell of a spreadsheet).
- We count the most meaningful words to determine which are more important and which are less important.
- N-grams – how many words in a pattern of words are we associating and counting as a pattern
- "No" and "good" counted separately is a 1-word ngram, "no good" is a 2-word engram

Why do we count words for business purposes?

- We survey what customers wanted in a product.
  - Besides asking quantitative questions (such as “on a scale of 1 to 5…”), we also ask open-ended questions (such as “Is there anything else you want us to know…”).
  - Doing a word frequency analysis of the text responses to the open-ended question and comparing it to the word frequency analysis of the product description tells us if we are meeting expectations.
  - The more often the word frequency tables match, the more we are matching the customer’s expectations.
- Consider a company that launched a campaign to make employees more aware of the new company mission.
  - After some time and using an employee survey, we ask the open-ended question: “Can you tell us what you think the company mission statement is?”
  - By doing a word frequency analysis of the mission statement and comparing it to the frequency analysis of the employee’s responses, we can gauge how successful our awareness campaign has been.

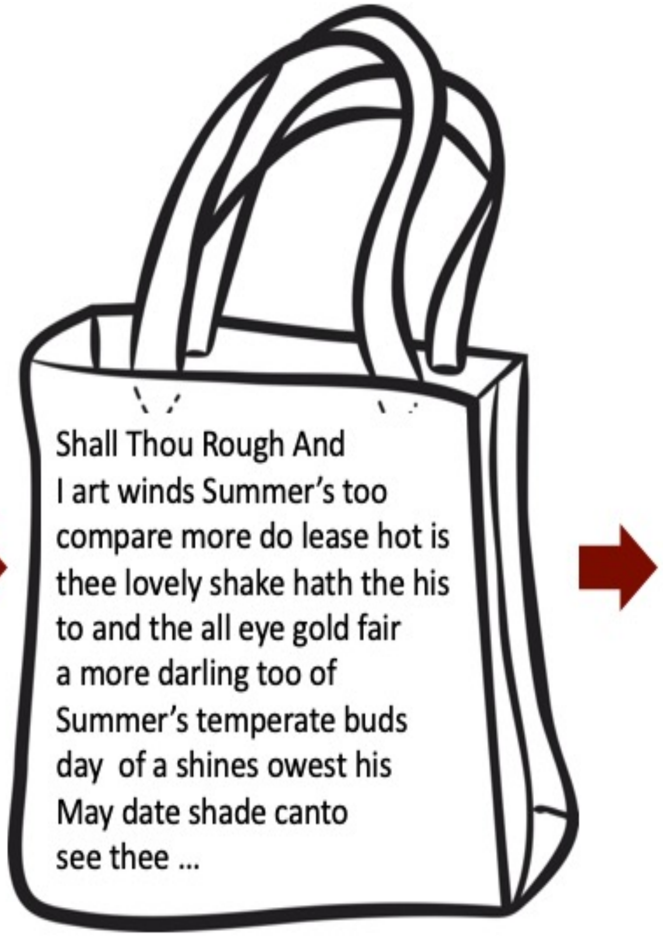
The Bag-of-Words Textual Data Model

- The Bag-of-Words model is a representation of textual data used in natural language processing (NLP) and information retrieval (IR).
- In this model, text (such as a sentence or a document)
  - represented as the bag (multiset) of its words,
  - without regard to grammar and word order,
  - with the multiplicity of the words maintained (for later counting, for example).
- One of the main problems that can limit this technique's efficacy is the existence of prepositions, pronouns, and articles in our text.
  - These words likely appear frequently in our text, but they lack information about the main characteristics and topics in our document.
  - These are removed in the process of creating a Bag-of-Words text data file or data element. Chapter 5 demonstrates techniques to remove these unwanted terms before analysis.

Converting a Text into a Bag-of-Words

Shall I compare thee to a Summer's day?
Thou art more lovely and more temperate:
Rough winds do shake the darling buds of May,
And Summer's lease hath all too short a date:
Sometimes too hot the eye of heaven shines,
And often is his gold complexion dimmed,
And every fair from fair some-time declines,
By chance, or natures changing course untrimmed:
But thy eternal Summer shall not fade,
Nor lose possession of that fare thou owest,
Nor shall death brag thou wanderest in his shade,
When in eternal lines to time thou growest,
So long as men can breathe or eyes can see,
So long lives this, and this gives life to thee.

Text



Shall Thou Rough And
I art winds Summer's too
compare more do lease hot is
thee lovely shake hath the his
to and the all eye gold fair
a more darling too of
Summer's temperate buds
day of a shines owest his
May date shade canto
see thee ...

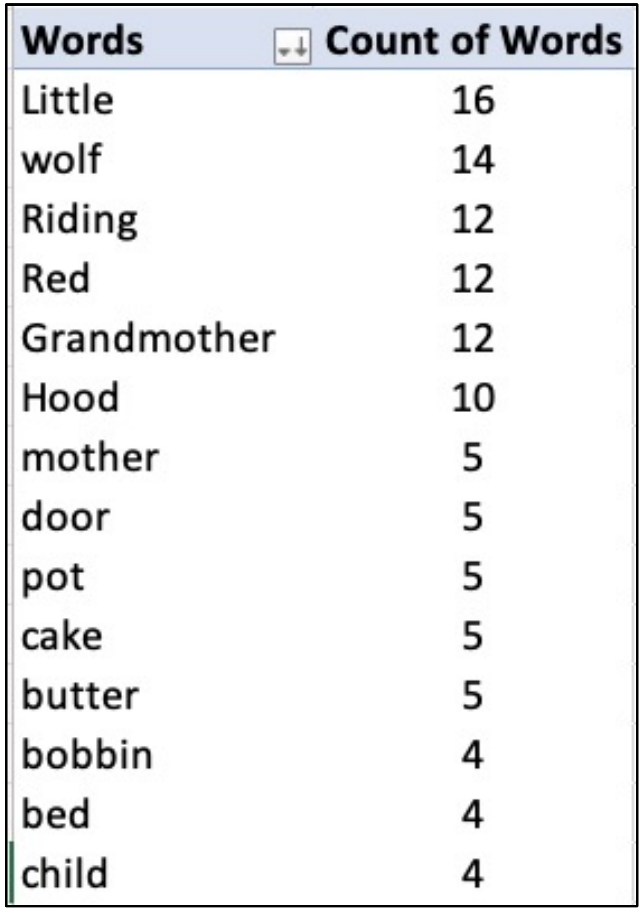
Bag of Words

Term	Count
shall	3
eternal	2
fair	2
long	2
summer's	2
time	2
art	1
brag	1
breathe	1
buds	1
chance	1
changing	1
compare	1
complexion	1

Term Frequency

Use of a Stop Word List

- Not all words are of interest when looking for meaning, so we often remove the most common words in a *stopword list*. Words such as “an,” “it,” “me,” and “the” are all words that are usually found on such a list.
- Once removed, looking at the frequency of occurrence of the remaining words could be very informative.
- If, for example, we removed the stop words and analyzed the word frequency of a *Little Red Riding Hood* fable, we would have a very informative table (Figure 5.1). It gives us an excellent picture of the story in the fable.



Words	Count of Words
Little	16
wolf	14
Riding	12
Red	12
Grandmother	12
Hood	10
mother	5
door	5
pot	5
cake	5
butter	5
bobbin	4
bed	4
child	4

COUNTIF – A Powerful Tool

- Excel has several text processing tools you should have in your toolbox:
  - COUNTIF – counts number of cells in a range that contain a word
  - LEN – counts total characters in a cell
  - =IF(LEN(TRIM(B2))=0,0,LEN(TRIM(B2))-LEN(SUBSTITUTE(B2," ",""))+1)
    - Counts the number of words in a cell
  - Excel has a limit of 32,700 characters in any one cell

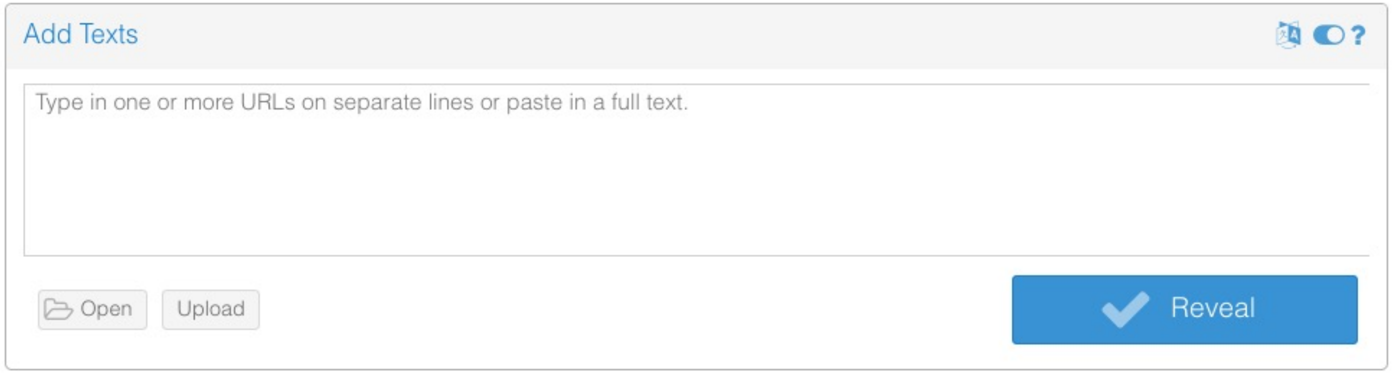
Exercises

- Open the *Graduate Course Description* file
- STEP 1 – Aggregate the course descriptions
- STEP 2 – Split cell into words
- STEP 3 – Transpose and Clean
- STEP 4 – Set up table for use of Pivot Table
- STEP 5 – Count unique words using a Pivot Table
- STEP 6 – Remove stop words
- STEP 7 – Final Term Frequency Table
- Solution file have the word SOLUTION in the file title

Voyant on the Web

VOYANT

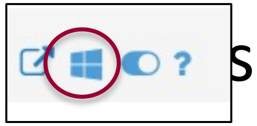
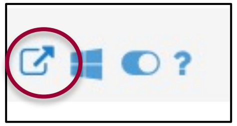
see through your text

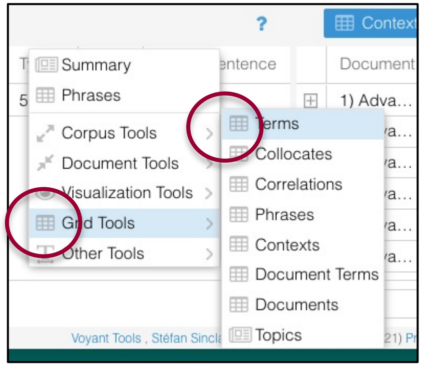


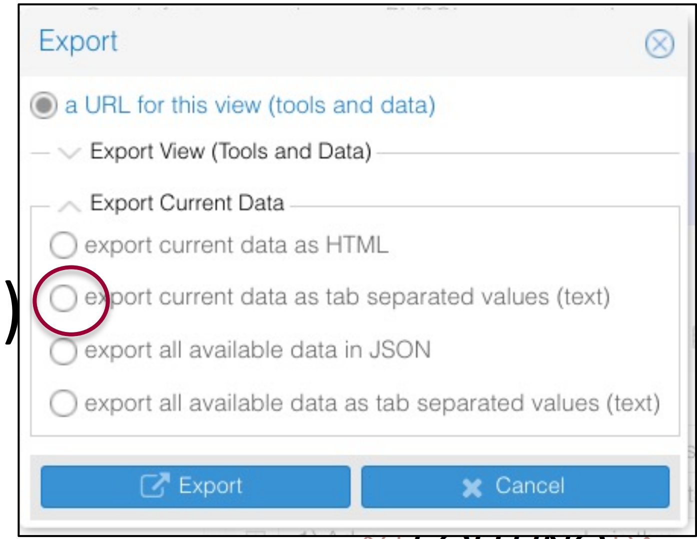
Type in one or more URLs on separate lines or paste in a full text.

Voyant Tools is a web-based reading and analysis environment for digital texts.

Term Frequency using Voyant

- In the *TF – Count Words* Excel file select the aggregated top 7 course descriptions (10,000 characters)
- Copy the aggregated course descriptions into the computer buffer
- Open Voyant on the web: https://voyant-tools.org/
- Paste the aggregated descriptions as text into Voyant and click *Reveal*
- In the tools  select Grid Tools and Terms
- Back in the selections choose Export  and (text)
- Scape table and paste into Excel – compare




Export

● a URL for this view (tools and data)

Export View (Tools and Data)

Export Current Data

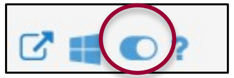
○ export current data as HTML

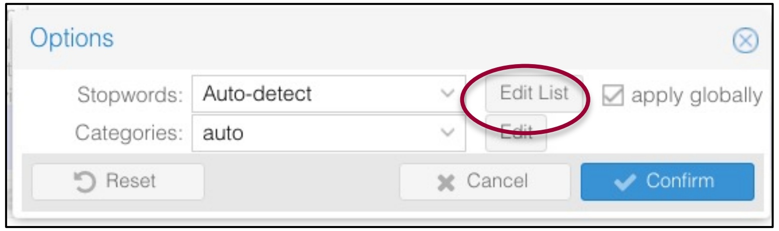
○ export current data as tab separated values (text)

○ export all available data in JSON

○ export all available data as tab separated values (text)

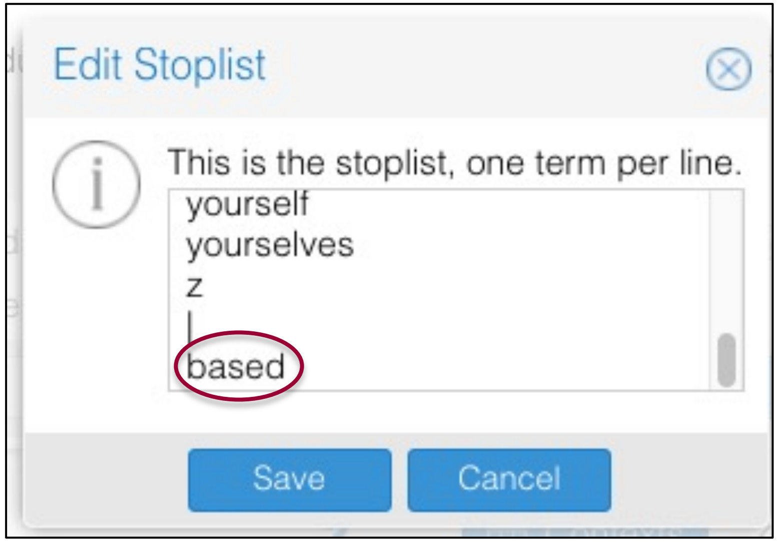
Additional Exercises

- Select a term to be removed via the stopword list,
  - edit the list  remove “based”
  - run the analysis again and assure yourself that the term is no longer in the analysis
- Run the analysis again by uploading the text file of the 7 courses into Voyant
  - Compare to the table from pasting the text directly into Voyant
- Run the analysis again by creating a corpus of each of the 7 courses as individual text files
  - Use Document Terms Grid tool in Voyant
  - Download the table of term frequency by text and using a pivot table get a crosstabulation of term frequency by course



Stopwords: Auto-detect

Categories: auto



This is the stoplist, one term per line.

yourself
yourselves
z
based

FORTINO
Global Education