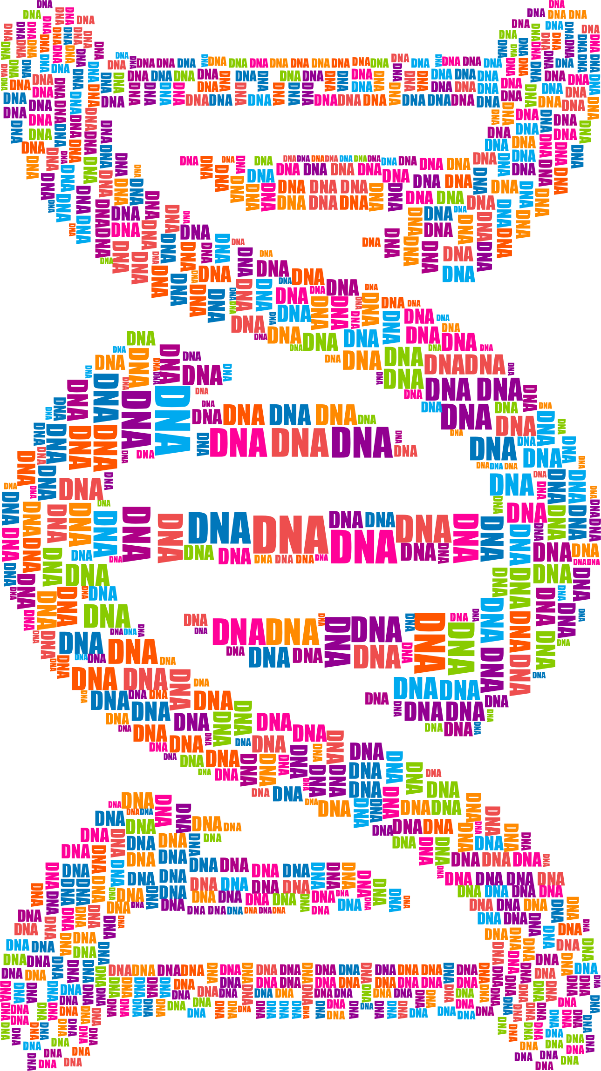
27/06/2018

UCSI software project

User guide



**University of Colorado Summer Internship (UCSI) software project user guide**

Table des matières

[Introduction 2](#_Toc521411642)

[I. Installation 3](#_Toc521411643)

[a. R version 3](#_Toc521411644)

[b. Packages to install 3](#_Toc521411645)

[II. Usage 5](#_Toc521411646)

[a. Running the app 5](#_Toc521411647)

[b. Using the app 5](#_Toc521411648)

[1. Data 5](#_Toc521411649)

[2. Pre-processing 8](#_Toc521411650)

[3. Overview 8](#_Toc521411651)

[4. Plot filtering 9](#_Toc521411652)

[5. Report 10](#_Toc521411653)

[III. Identified bugs and improvement 11](#_Toc521411654)

[a. One word word cloud 11](#_Toc521411655)

[b. Table options change 11](#_Toc521411656)

[c. Plots 12](#_Toc521411657)

[d. Wordcloud 12](#_Toc521411658)

[Contact 12](#_Toc521411659)

# Introduction

This app is a visualization tool for Natural Language Processing. It has two purposes:

* Process textual data and analyze it
* Visualize the data obtained after analysis

The app should help researchers in linguistics understanding the structure of the documents they study. It provides a set of statistical metrics, sentence structure analysis and other analysis. It then represents them in order to extract useful information from it. It is built on Shneiderman's mantra “Overview first, zoom and filter, then details-on-demand”. It should cover a wide range of the possible analysis in Natural Language Processing and present them efficiently.

# Installation

## R version

In order to use the application, you need to have the 3.5.0 (2018-04-23) R version, which is currently the latest R version. It should work with updates, so if you have a later version, it should also run.

## Packages to install

In order to use the application, you also need specific packages. A R program will be developed to facilitate the installation of the software. However, currently you need to install the following packages:

* *shiny*
* *DT*
* *plotly*
* *wordcloud2*
* *tm*
* *crosstalk*
* *webshot*
* *phantomjs*
* *tinytex*
* *shinydashboard*

The packages needed will expand as the back-end evolves. Indeed, the application doesn’t really process the data yet (only the minimum to test the app), so there will be new packages to install in order to perform this task.

Firstly, you need to install Rtools, which can be done at <https://cran.r-project.org/bin/windows/Rtools/> .

To install the packages, you need to copy/paste the following lines into the console:

install.packages("shiny")

install.packages("DT")

install.packages("plotly")

install.packages("wordcloud2")

install.packages("tm")

install.packages("crosstalk")

install.packages("webshot")

webshot::install\_phantomjs()

install.packages("tinytex")

tinytex::install\_tinytex()

install.packages("wordcloud")

install.packages("rmarkdown")

install.packages("knitr")

install.packages("httpuv")

install.packages("stringi")

install.packages("shinydashboard")

If that fails, you can try to install also the two following packages with the two lines below:

install.packages("ggplot2")

install.packages("NLP")

# Usage

## Running the app

To run the app, create a folder, and put the R file in it. It is important not to change the name of the R file. It must always be app.R. So that when you call the file, R knows exactly what to do with it. After this, you need to load the shiny library and use the runApp() command as showed just below:

library("shiny")

#Path is the path where the app.R file is.

runApp("Path", launch.browser = TRUE)

You need to be careful about one thing when you put in the path. You have to change “\” in “/”. For instance:

runApp("C:\Users\Projet\Desktop\Internship\Application", launch.browser = TRUE)

will not work, but:

runApp("C:/Users/Projet/Internship\_Visua\_Combine", launch.browser = TRUE)

will work.

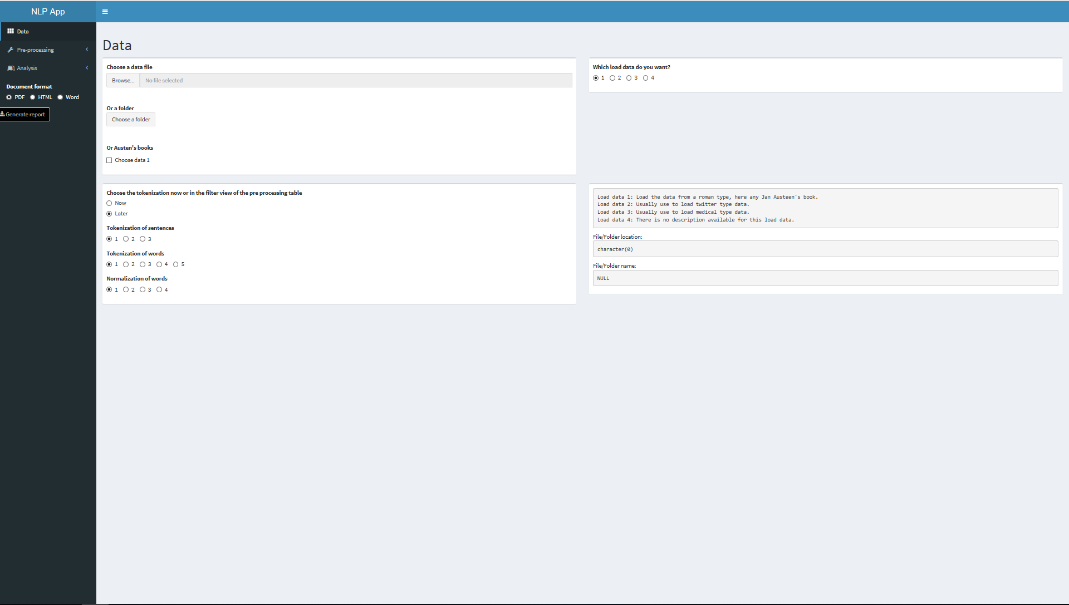
When you type this command line into the console, the application will open on your default browser. It could be open in R, but some functionalities don’t work or badly work. However, if you want to open it in R, just remove launch.browser = TRUE from the command above.

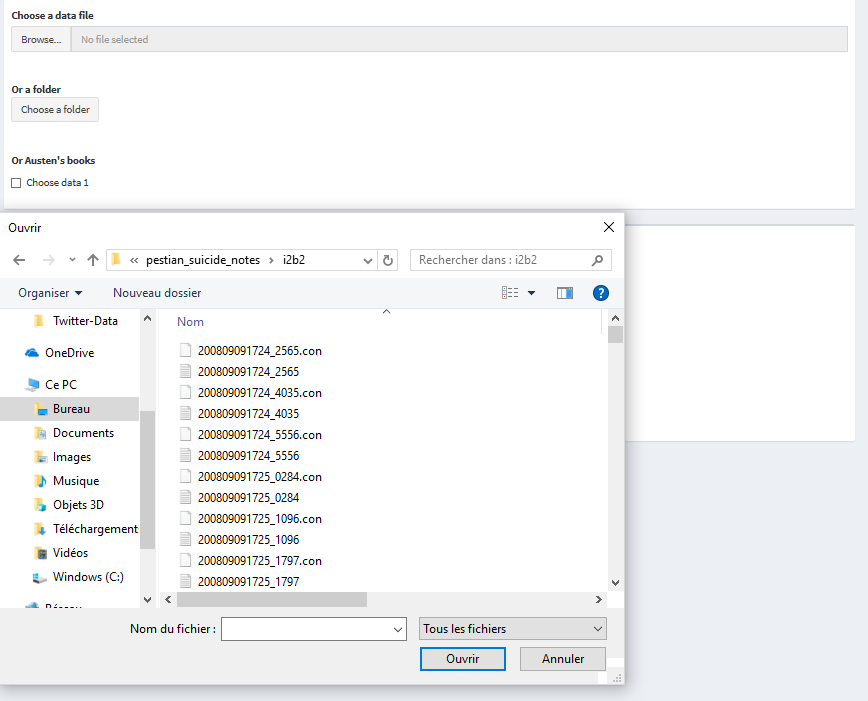
## Using the app

### Data

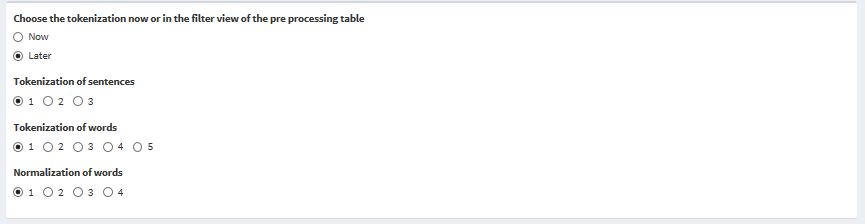
Files used for this screen:

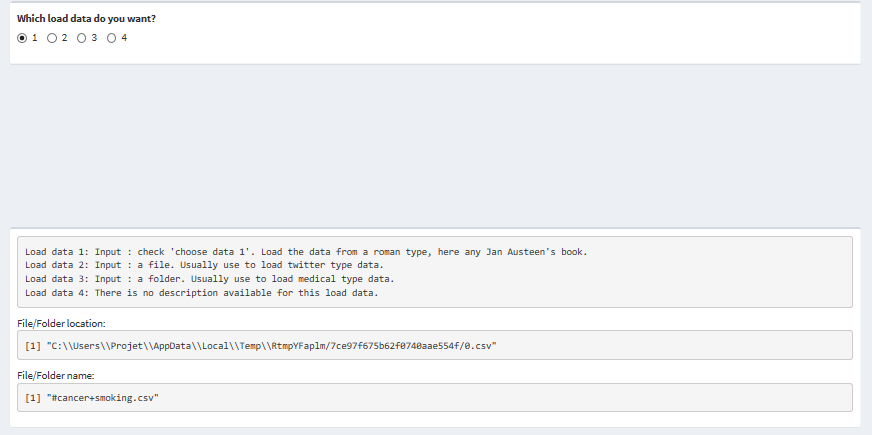
* Intership\_NLP\_CU-master\description\ description\_type\_data.R
* Intership\_NLP\_CU-master\load\_data
* Intership\_NLP\_CU-master\app
* Particularly: line 5 to 106 from Intership\_NLP\_CU-master\app\server.R

The data screen is mainly used to load your data into the app. You can choose how to read it, and if you want, which tokenization to apply. If you want to use the app without putting your data in it, choose “Choose data 1” and “1” in “Which load data do you want?”.

This first box is used to load your data. Depending on how to read it, you can either choose a folder or a file. If you want just to test the app, you can also choose an extract of Jane Austen’s books.

The folder option can be used to read different files that are in the same folder and merge them. For example, you can analyze many last words of people which are all in one file.

One of the goal of the app is to help you choose the right tokenization and see the impact of this choice. However, you can also only use the app to analyze the text you want. So if you already know which tokenization you want to use, you can choose it in this screen by checking “Now”. It will automatically process the text with the right tokenization in the analysis. In the default case, you choose the tokenization in the filter part of the pre-processing (“Later”). Each number of tokenization corresponds to a file number located in Intership\_NLP\_CU-master\preprocessing folder.

This is the box where you can choose which function will read your data. There might not be an appropriate function here for your data. In this case, read ………. to learn how to add one.

You can see below the checkboxes the description of each function, along with the input needed. You can also see the file location and name of the file you selected or the folder location if you have chosen this option.

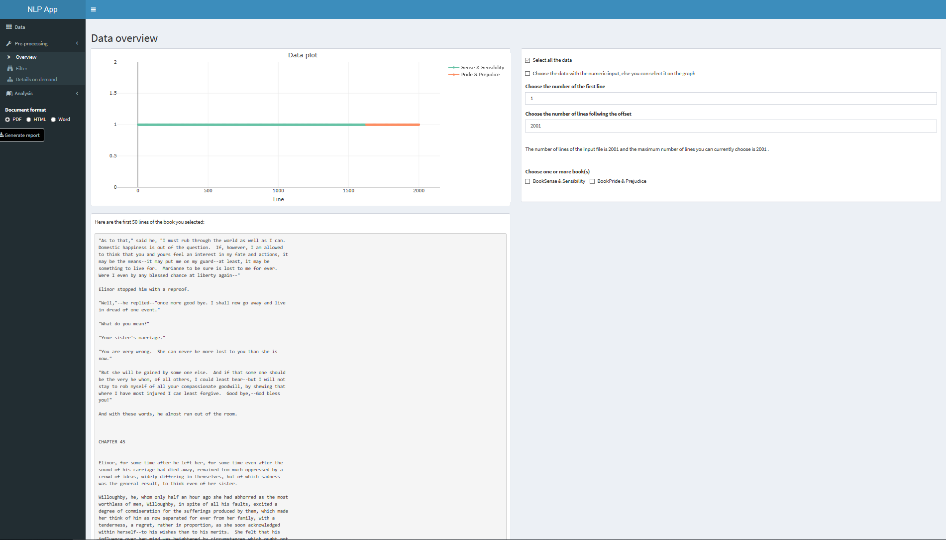
### Pre-processing

This part is to pre-process the text. It will enable you to choose more precisely which part of the text you want to study, show the importance of the choice of tokenizations and see how regular the data is. At the end of this process, you will have chosen the tokenizations you want.

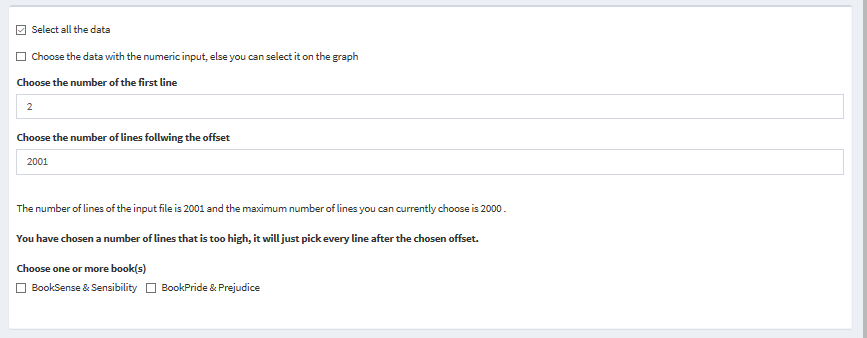
#### Overview

Files used for this screen:

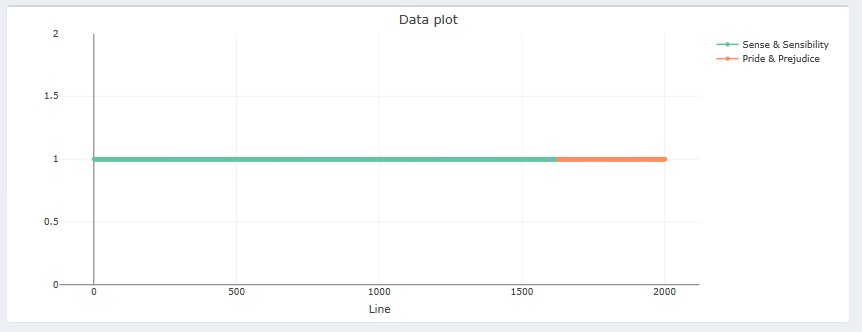
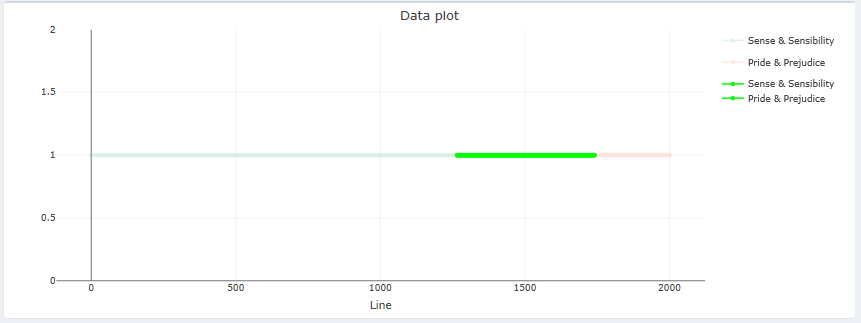
* Intership\_NLP\_CU-master\app
* Particularly: line 106 to 270 from Intership\_NLP\_CU-master\app\server.R



The overview screen is the one to use to choose precisely which part of the text to analyze. You have several ways to select it that will be explained just below. Different parts of the text can have different linguistics particularities. And this overview should enable you to highlight that.



In this box, you have three options to choose the data. There can’t be two options at the same time. If you check “Select all the data”, the text will be analyzed totally. This is the default option. You can also select which lines you want. To do that, check the second check box, and then put the number of the first line and the offset you want. There is a message telling you how many lines there are in the text and the maximum number of lines you can choose. If you have chosen a number of lines too high compared to the number of lines and the offset, a bold message will appear warning you of this problem. The third way to select your data is to select it by book. You can select as many books as you like with the checkboxes. If you didn’t put any book in your input data, there will just be a default name of book for the whole data (“BookAllTheSame”).

The fourth and last way to select data is to select it manually. You can use the mouse to select it on the plot directly. If there is a checkbox selected, you must select once to decheck the checkbox and then another time to select the data for real. The advantage of this method is that you can see which book to select and which part of that book you select (the beginning, the end…). The different colors on the plot indicate different books.

The number of lines corresponds to the size of the text. And the abscissa of the plot is the number of lines.

If you can’t manage to select on the plot, see …….. . It is probably due to the browser you’re using.

You can see just on the left what happens when you select directly on the plot.

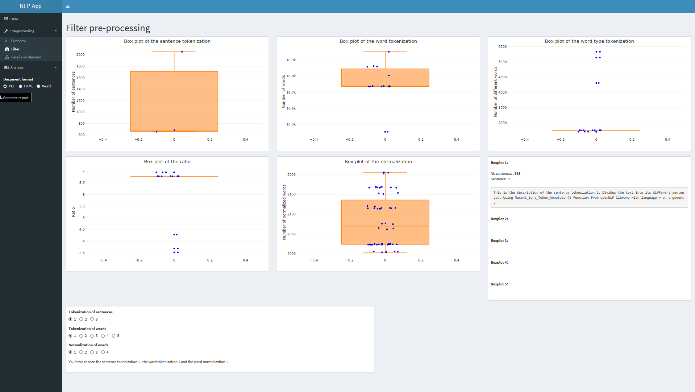


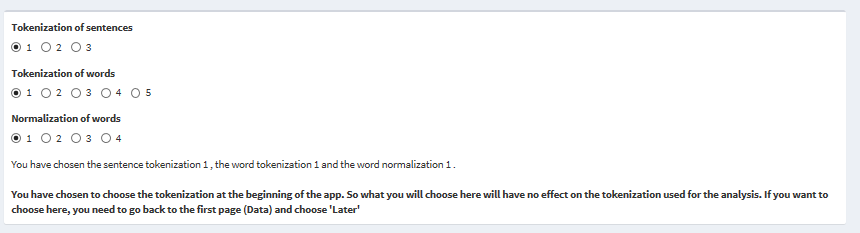
The last feature of the overview is the box on the left here. It prints the first 50 lines of the selected text. It is a good way to see where you’re in the text. It creates a link between the number, the books and the real text. Usually, when analyzing a text, there is a shifting towards the numbers (number of words, frequency, line number…). This feature enables you to not forget the text you’re analyzing. By visualizing the text, it enables you to be more precise in your selection. For example, if you want to select just the introduction, although you don’t know exactly the line it ends, you can see where it is.

#### Filter

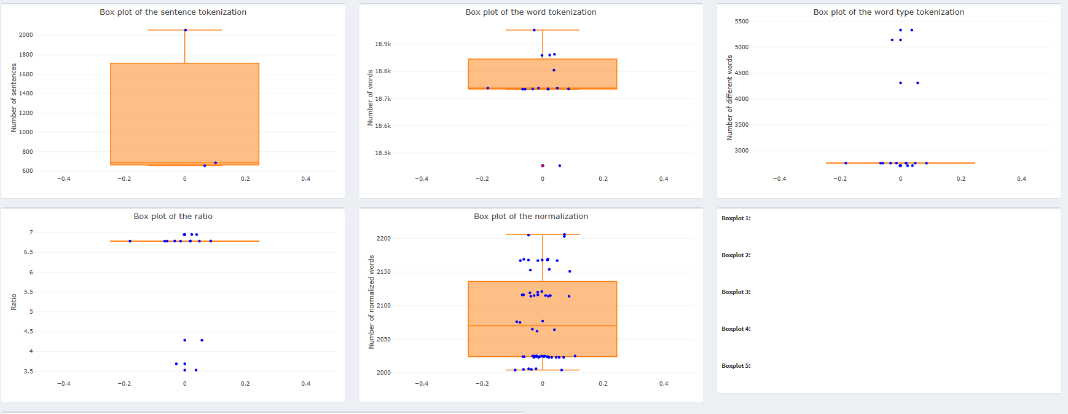
Files used for this screen:

* Internship\_NLP\_CU-master\backend\_analysis\token\_boxplot
* Internship\_NLP\_CU-master\backend\_analysis\modulo\_not\_null
* Internship\_NLP\_CU-master\preprocessing
* Intership\_NLP\_CU-master\app
* Particularly: line 270 to 501 from Intership\_NLP\_CU-master\app\server.R

The filter screen is the one to see the impact of the choice of tokenization. This is also where you can make your choices of tokenzations for the rest of the analysis considering the results of the boxplots. It will enable you to choose wisely your tokenization, which is usually a big problem to face.



This is the box to choose the tokenization. The choices are automatically taken from the files where the tokenzations are (see Internship\_NLP\_CU-master\preprocessing for more details). If you have chosen “Now” on the data screen, a message will appear saying that the choices you make here will have no impact on the rest of the analysis unless you change your choice on the data screen. It will also always show you your choices of tokenization.



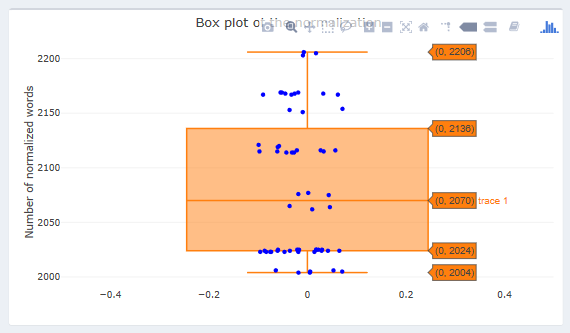
The first boxplot shows the variation of the number of sentences depending on the tokenization\_sentence function used.

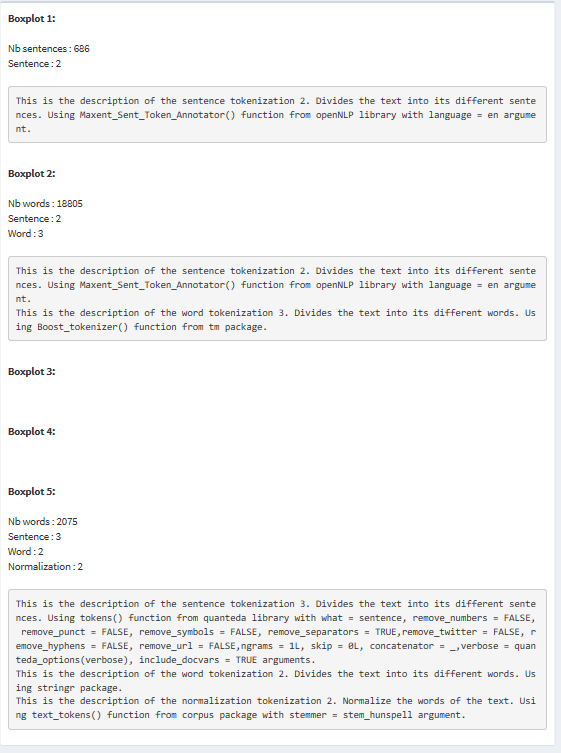
This boxplot shows the variation of the number of word occurrences depending on the tokenization\_word function used. That means the total number of words in the text selected (with repetition).

This boxplot represents the variation of the number of word type depending on the tokenization\_word function used. That means the number of distinct words in the text select (without repetition).

This boxplot shows the variation of ratio (which is the word occurrences divided by the number of word types) depending on the tokenization\_word used. The ratio is a very meaningful information about the text. A huge variation of this data regarding on the tokenization \_word could be a real problem, in a sense that a really accurate value for the ratio has a really specific meaning.

This boxplot shows the variation of the number of type words after normalization depending on the normalization function used. Normalization is a process during which each word is transform into a single canonical form those forms are regrouped so they appear just ones (e.g. ‘go’, ‘goes’, ‘going’, ‘gone’ and ‘went’ will be associated to ‘go’).

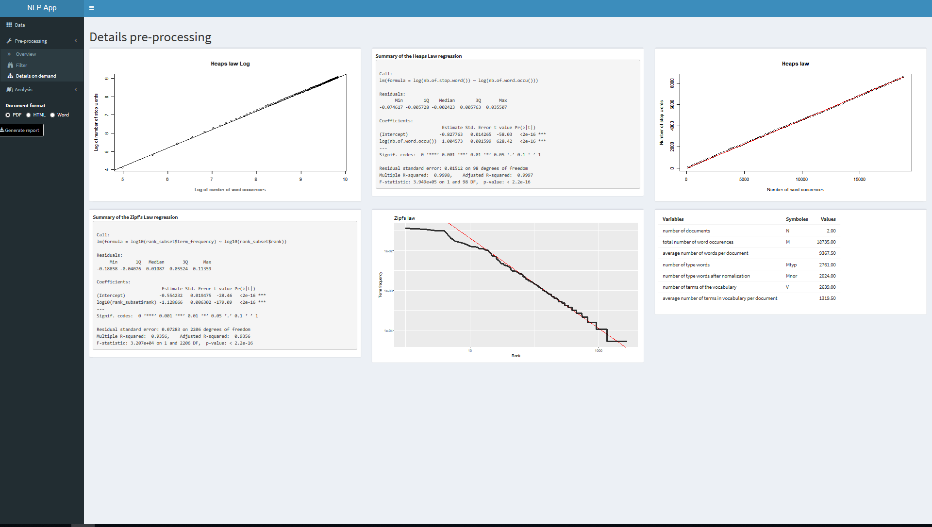
This are the five boxplots. On all of them, you have two interesting features. The first one is when you hover over the boxplots you can see the lower whisker, first quartile, median, third quartile and the upper whisker. Each point corresponds to a number of a certain tokenized identity after applying a certain tokenization (e.g. the number of words after applying tokernization\_word\_1). The abscissa of a point is determined randomly in order to be distinguishable from the others.

The second interesting feature is that if you click on one the blue points, you will have information about these one appearing on the left of the screen. You will have the ordinate of the point selected, its tokenization(s) and the description of its tokenization(s). The description should enable you to better understand which point and whether it is pertinent or not.

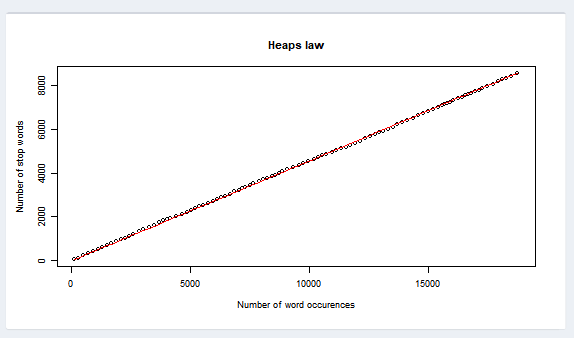
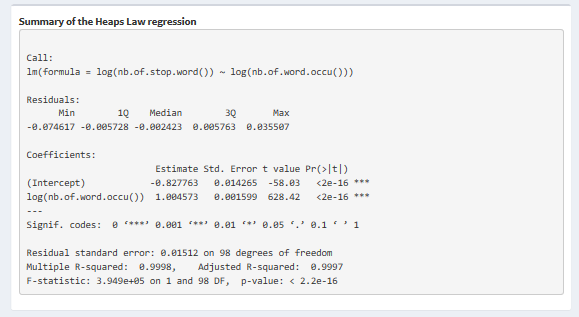
#### Details on demand

Files used for this screen:

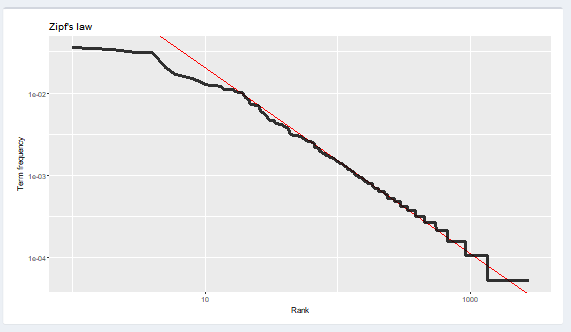
* Internship\_NLP\_CU-master\backend\_analysis\heaps\_law.R
* Internship\_NLP\_CU-master\backend\_analysis\zipfs\_law.R
* Internship\_NLP\_CU-master\ backend\_analysis\table\_info.R
* Internship\_NLP\_CU-master\ backend\_analysis\after\_choose\_token.R
* Intership\_NLP\_CU-master\app
* Particularly: line 501 to 604 from Intership\_NLP\_CU-master\app\server.R



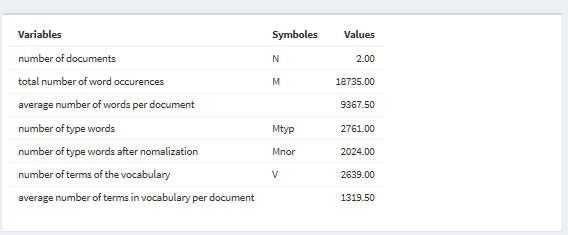
The details on demand screen is used to see how regular the data is. It tests that by plotting two important laws that verify this. It also shows summary tables with interesting numbers from the tokenization. The data analyzed here is the data with the tokenizations chosen before.

In the screen, there is three plots. The first two ones are about Heaps law. One is in log and shows how well the regression is. The other one is Heaps law itself. Heaps law is a law that verify that the bigger the number of occurrences there are the bigger the vocabulary is.

Along with this plots, there is a summary if the linear regression. It enables you to see how well the regression is and if Heaps law works well. And know the value of the empirical formula of Heaps law. If you want to know more about summaries in R, go to [this link](https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R).



The last plot is Zipf’s law. This law indicates that the frequency of occurrence for a word is inversely proportional to its rank. The rank is obtained when we sort the frequency by decreasing order. There is also a summary next to it to see the regression. It is the same idea as for Heaps law.



Finally, this table shows interesting numbers concerning the chosen tokenization. It summarizes and go further in the analysis. If you want to add other numbers, go see ......... to see how to do that.

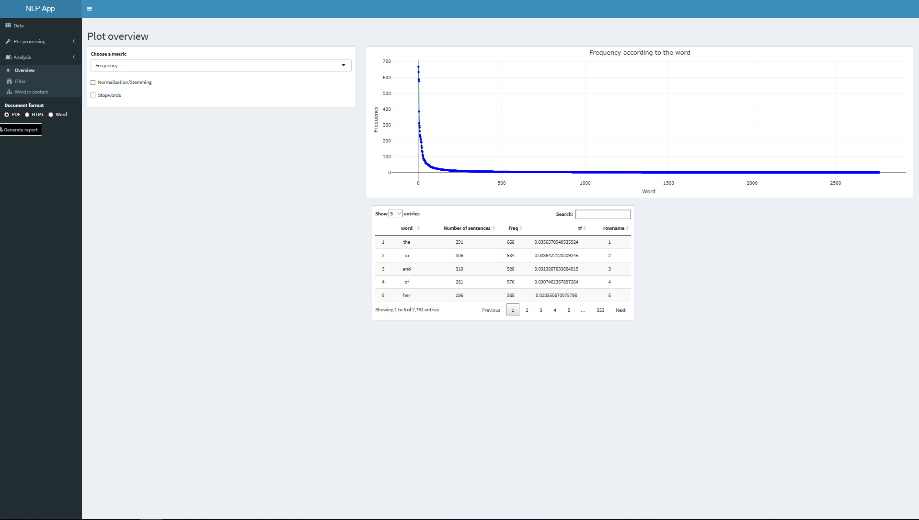
### Analysis

The goal of this part is to analyze a text. It should link two views of the language that are the symbolic one and the statistical one. The overview is a statistical analysis, and the “word in context” is a symbolic analysis while the filter is the link between the two. Here the statistical analysis is a macroscopic analysis while the symbolic one is a microscopic one. That is one of the reason Schneiderman’s mantra was chosen to implement this app.

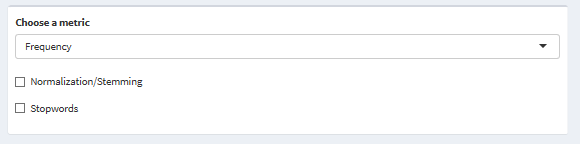
#### Overview

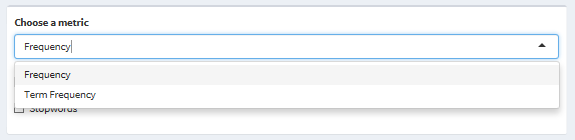
Files used for this screen:

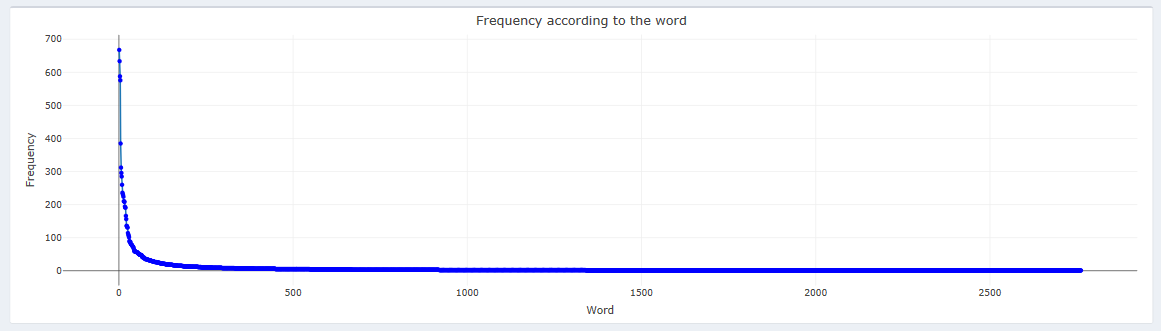
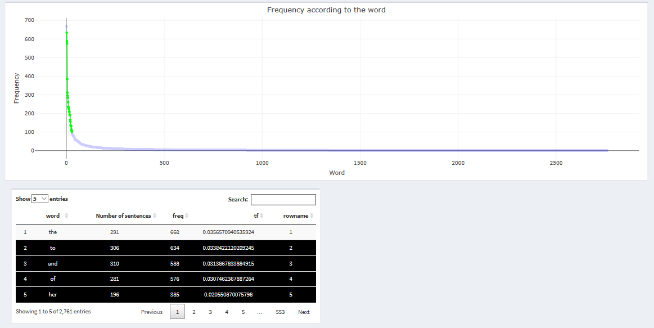
* Internship\_NLP\_CU-master\preprocessing\stop\_word
* Internship\_NLP\_CU-master\preprocessing\normalization
* Intership\_NLP\_CU-master\app
* Particularly: line 605 to 653 from Intership\_NLP\_CU-master\app\server.R

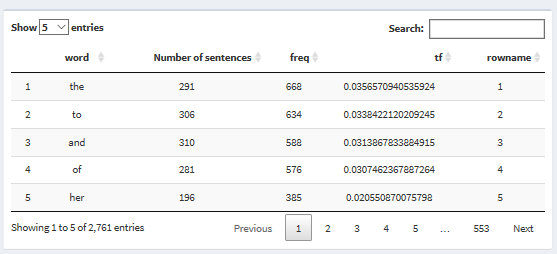


This screen enables you to do several tasks. You can choose a metric, and then according to this one, you can select the words you want to focus on for the next step. In order to do this, there is also a table in which you can search more precisely.



On this box, you can choose a metric (frequency or term frequency) on the select input button. To add another metric, see ……. You can also select if you want your text to be normalized, without stopwords or both, by checking the box below the select input button.

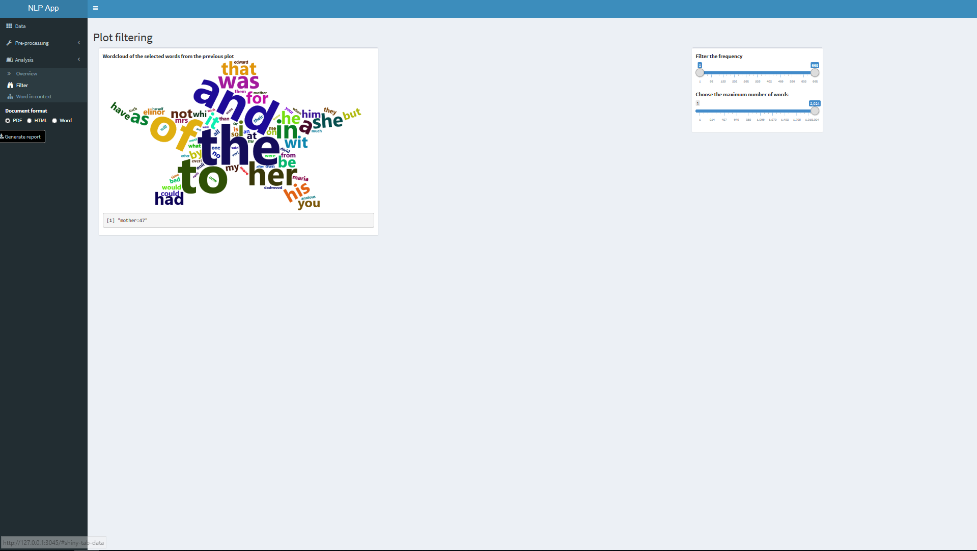
This box shows an interactive plot, that changes accordingly to the chosen metric. In the abscissa, the words are sorted by frequency. In almost every text, the frequency curve should look like the one on the right figure. You should select data points on the plot. It is a necessary step to continue the analysis afterward. Indeed, the selected words will be the one to be analyzed. In order to deselect them, you need to double-click on the plot. They will automatically appear on black in the data table, while the non-selected points will appear on white in the data table. You can see this on the right figure.

Finally, this box is a data table in which you can search very precisely. You can sort by every column (by frequency, tf…). You can search for a particular word or value. You can also choose to show more entries. It is a powerful tool to complete your statistical analysis.

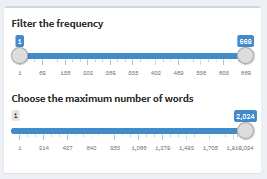
#### Filter

Files used for this screen:

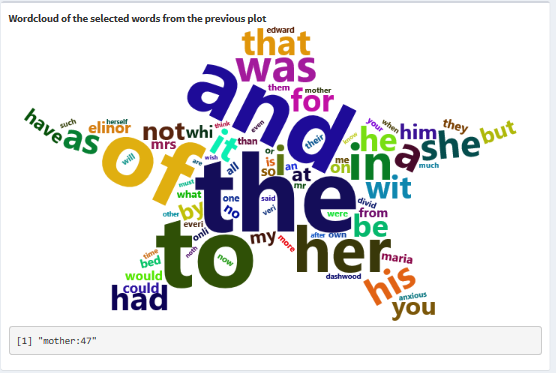
* Intership\_NLP\_CU-master\app
* Particularly: line 654 to 694 from Intership\_NLP\_CU-master\app\server.R



The filter screen enables you to choose more precisely which word you want to study. It also helps with a wordcloud of the words. This is both a filter to analyze in more depth and a tool to deduce interesting points.



This box enables you to choose a frequency range for the words in the cloud and a maximum number of words in the cloud with the slider inputs. This is a good tool to filter even more your text. There is one common bug that appears when you select only one word. It is described in …… .

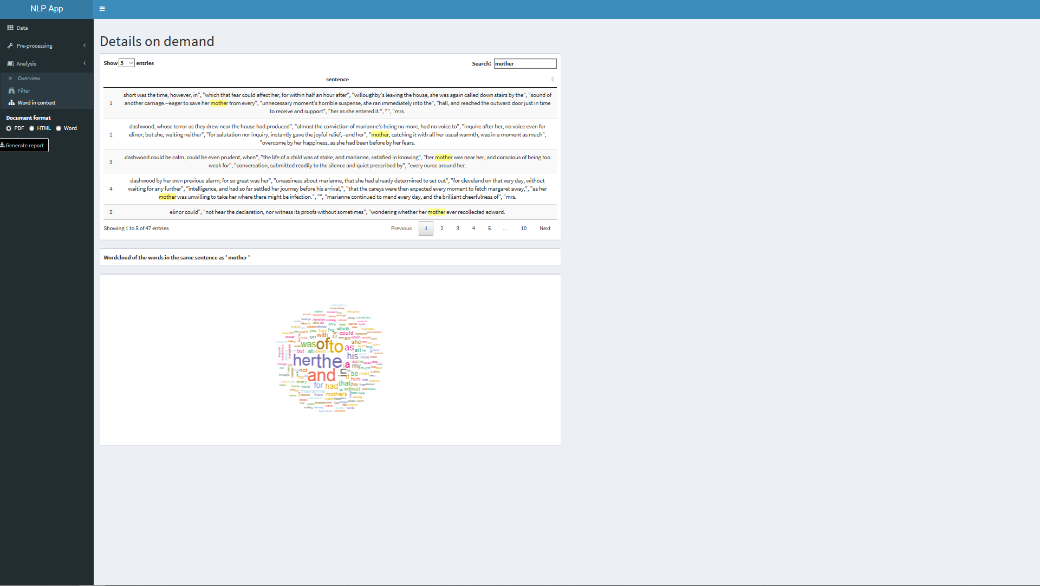


This box is a wordcloud. The sizes of the words are proportional to the frequency. The words in this wordcloud are the ones selected in the previous plot (in the “Overview”). You hover on the words and see the associated frequencies. You can also click on a word and see it appears at the bottom of the word cloud with its associated frequency. You should click on a word you want to analyze. Indeed, the last step uses the selected word here to analyze more in depth.

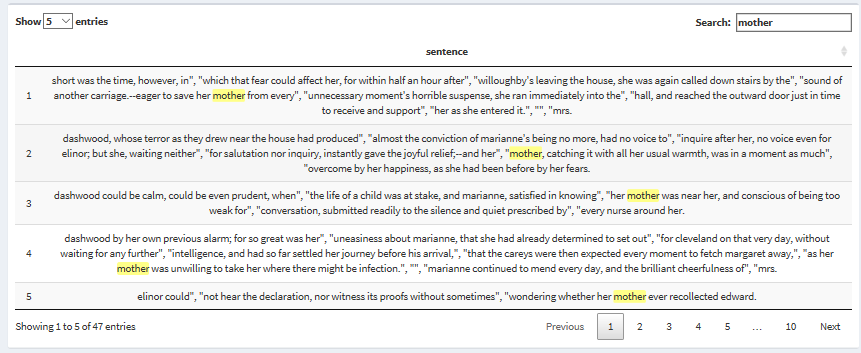
#### Word in context

Files used for this screen:

* Internship\_NLP\_CU-master\backend\_analysis\ wordcloud\_data\_func.R
* Intership\_NLP\_CU-master\app
* Particularly: line 695 to 762 from Intership\_NLP\_CU-master\app\server.R



This screen allows a more detail analysis link to a specific word. This word is the one chosen at the last step (“Filter”). You can see all the sentences it appears in, and all the words which are in the same sentence with their frequencies. This tool is the last part of the analysis and constitutes the symbolic analysis.



This box shows a table in which all the sentences which the chosen word in it appear. This selected word is highlighted in yellow. This enables a grammatical analysis and links the whole analysis to the text itself.

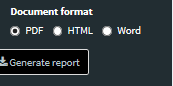


This box shows a wordcloud with all the words in the same sentence as the selected word (here ‘mother’). The sizes of the words depend on their frequencies. However, these frequencies are not the same as earlier. They are the frequencies of this words in the sentences where the selected word appears.

### Report

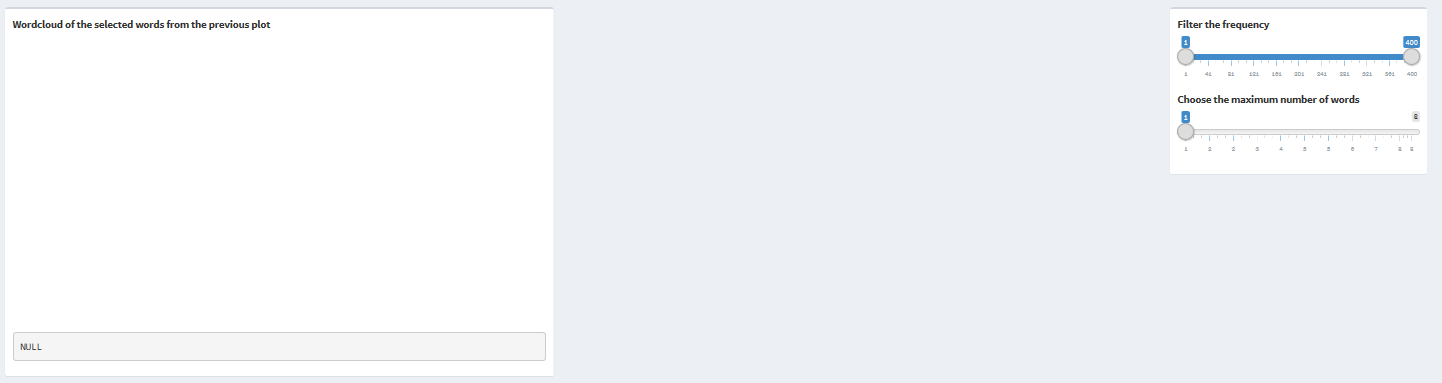
Files used for the report:

* Intership\_NLP\_CU-master\app
* Particularly: line 763 to 812 from Intership\_NLP\_CU-master\app\server.R
* And Intership\_NLP\_CU-master\app\report.Rmd

The app enables you to download a report of what you did during the session. This report includes the plots parameters that you chose during this session, the boxplots, the laws, the plots of the analysis overview, the plots, table and wordcloud of the selected points, and the sentences and wordcloud of the selected word. It is available in three output formats (PDF, HTML and WORD). It also contains a table of content. You can see an example of report in Internship\_NLP\_CU-master\guide. The power of this report is to have all the results of your analysis stored in your computer so that you can use them for your articles. It is also a good way to compare different documents. It is particularly easy with the reminder of the parameters you chose to reproduce it to another document. There is a common bug with the PDF that can occur that is described in ……. .

# Identified bugs

## One word word cloud



In the "Filter" screen of the analysis, when there is only one word selected, no word appears. This problem can appear either by selecting only one word in the plot in the "Overview" screen, either by putting the slider bar of the maximum of words to 1. There is no fix of this bug. It is not a big problem because people usually don't need to have one word wordclouds.

# Improvements

## Plots

In the “Plot overview” page, there are two possible improvements for the plots:

* When selecting points on one plot, and then changing the metric, the points are no longer selected in the new plot with the new metric. This could be nice to have the points previously selected when changing the metric.
* Ordering the points according to the metric. Currently, the points are ordered by the the words’ id, so they don’t appear nice with metrics. There is an easy way to change that and adapt it to the data but it requires either a new column for each metric either a long computational process each time a plot is drawn, so there is a trade-off to do for this feature. The best option is probably the one that uses more space.

# Contact

If you have any question about the app, troubles to install/use it or would like to report bugs, please free to contact us.  
  
Colette Voisembert [c.voisembert@gmail.com](mailto:c.voisembert@gmail.com)

Leon Migus [leonmigus@gmail.com](mailto:leonmigus@gmail.com)