During this time of year there is obviously a lot of talk about the Bible. As most people know the New Testament comprises four different Gospels written by anonymous authors 40 to 70 years after Jesus’ supposed crucifiction. Unfortunately we have lost all of the originals but only retained copies of copies of copies (and so on) which date back hundreds of years after they were written in all kinds of different versions (renowned Biblical scholar Professor Bart Ehrmann states that there are more versions of the New Testament than there are words in the New Testament). Just as a fun fact: there are many more Gospels but only those four were included in the official Bible.

So in general it is interesting to learn a little bit more about this stuff, especially because, for better or worse, it is at the core of Western civilization. One interesting question is how do the four Gospels relate to each other. To find that out you could either study ancient Greek and Christian theology for many years and decades – or you could use the stylo package and do a cluster analysis within seconds. Obviously we go for the latter…

For that matter we download one version of the four Gospels from [Project Gutenberg](https://www.gutenberg.org/) in the original ancient Greek and put it into a separate directory called ‘corpus’ (which is the name of the default directory of the package). The famous beginning of the Gospel of John (“In the beginning was the Word, and the Word was with God, and the Word was God…”) reads as follows in Ancient Greek:

Στην αρχή ‘ταν ο λόγος κι’ ο λόγος είτανε με το  
Θεό και Θεός είταν ο λόγος. Είταν εκείνος στην αρχή  
με το Θεό. Όλα τα πάντα μέσο του έγιναν, και χωρίς  
του τίποτα δεν έγινε που γίνηκε. Μέσα του είτανε ζωή  
κι’ η ζωή ‘τανε το φως των ανθρώπων, και το φως  
μέσα στο σκοτάδι φέγγει και το σκοτάδι δεν το κυρίεψε.

For your convenience I provide you with the curated texts here: [John](http://blog.ephorie.de/wp-content/uploads/2018/12/John.txt), [Luke](http://blog.ephorie.de/wp-content/uploads/2018/12/Luke.txt), [Mark](http://blog.ephorie.de/wp-content/uploads/2018/12/Mark.txt), [Matthew](http://blog.ephorie.de/wp-content/uploads/2018/12/Matthew.txt).

After that we can already call the main function of the package:

library(stylo)

## ### stylo version: 0.6.8 ###

##

## If you plan to cite this software (please do!), use the following reference:

## Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R:

## a package for computational text analysis. R Journal 8(1): 107-121.

##

##

## To get full BibTeX entry, type: citation("stylo")

stylo(corpus.lang = "Other", encoding = "UTF-8", mfw.min = 200, mfw.max = 200, custom.graph.title = "Ancient Greek Gospels", gui = FALSE)

## using current directory...

## Performing no sampling (using entire text as sample)

## loading John.txt ...

## loading Luke.txt ...

## loading Mark.txt ...

## loading Matthew.txt ...

## slicing input text into tokens...

##

## turning words into features, e.g. char n-grams (if applicable)...

##

## Total nr. of samples in the corpus: 4

## ....

## The corpus consists of 70943 tokens

##

## processing 4 text samples

##

## combining frequencies into a table...

##

##

## culling @ 0 available features (words) 5000

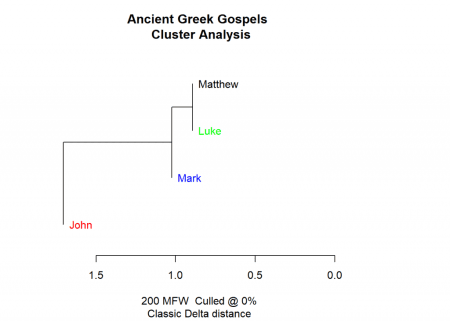
## Calculating z-scores...

##

## Calculating classic Delta distances...

## MFW used:

## 200



##

## Function call:

## stylo(gui = FALSE, corpus.lang = "Other", encoding = "UTF-8", mfw.min = 200, mfw.max = 200, custom.graph.title = "Ancient Greek Gospels")

##

## Depending on your chosen options, some results should have been written

## into a few files; you should be able to find them in your current

## (working) directory. Usually, these include a list of words/features

## used to build a table of frequencies, the table itself, a file containing

## recent configuration, etc.

##

## Advanced users: you can pipe the results to a variable, e.g.:

## I.love.this.stuff = stylo()

## this will create a class "I.love.this.stuff" containing some presumably

## interesting stuff. The class created, you can type, e.g.:

## summary(I.love.this.stuff)

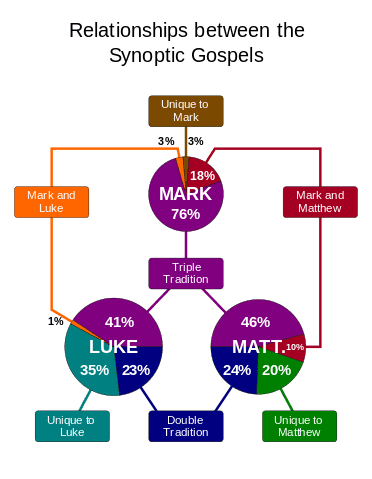
## to see which variables are stored there and how to use them.

##

##

## for suggestions how to cite this software, type: citation("stylo")

As you can see the Gospels of Matthew and Luke are more similar than Mark (the oldest Gospel), John (the last Gospel that made it into the Bible) is farthest away. This is indeed what Biblical scholars have found out by diligently studying the New Testament: Mark, Luke and Matthew are called “Synoptic Gospels” because they are so similar, John kind of stands on its own. The shared pieces of the Synoptic Gospels can be seen here (picture from Wikipedia, [Synoptic Gospels](https://en.wikipedia.org/wiki/Synoptic_Gospels)):



As you can see, Biblical scholars have found out that Luke and Matthew share about 60 to 70 percent of the text (through the so called double tradition), whereas Mark shares about 40 percent of the text (through the triple tradition). For comparison, John only shares about 10 percent with the synoptic Gospels. Now, this is exactly what our cluster analysis form above shows – what an impressive feat!

By the way, there is an even simpler way to interact with the stylo package: through the GUI. Just call stylo() and the following window will open where you can play around with different arguments:



It has never been easier to do sophisticated textual analyses, even in ancient languages one cannot understand! You might ask: how did R do that?

Well, behind the scenes R basically creates frequency tables for the words used in the different documents. It then tries to cluster the documents based on a similarity (or distance) measure. Intuitively that means that the more often authors use the same words the closer their texts are. As you can see R doesn’t really have to understand the words or the language used. It just maps statistical measures which seems to be quite effective (not only) in this case. The statistical measures often give some kind of characteristic signature of an author or a source. In a way this is the quantified version of fuzzy concepts like ‘writing style’ yet R can do it much more quickly even for long texts than any human ever could!

Happy New Year… stay tuned for much more to come in  
