**The Voynich Manuscript**

While the world abounds with strange phenomena ripe for analysis in their raw state, there is a peculiar pleasure in scrutinising arcane information curated and obscured by the human mind.

The Voynich Manuscript is one of the most well-known and studied volumes of occult knowledge. The book’s most recent history involves its purchase in 1912 by Wilfrid Voynich, a rare book dealer, from a sale of manuscripts by the Society of Jesus at the Villa Mondragone, Frascati. Following several fruitless years of attempts to decipher the manusript and discover its origin, or to interest others in it, Wilfrid Voynich died. The book passed through a number of other hands before being donated to Yale University by the noted rare book dealer Hans P. Kraus in 1969. It now resides in Yale’s [Beinecke Rare Book and Manuscript Library](https://beinecke.library.yale.edu/) with the designation [MS 408](https://brbl-dl.library.yale.edu/vufind/Record/3519597).

Written almost entirely in an unknown script, barring a small number of words apparently in Latin and High German, the manuscript is compellingly illustrated with depictions of plants, herbs, human figures, astronomical and astrological symbols. The manuscript has resisted all attempts at interpretation by cryptographers, historians, and linguists.

Voynich Manuscript – Folio 178

From a linguistic and cryptographic perspective, this lack of success in interpretation is not surprising. The two-hundred or so folios of the manuscript, while beautifully illuminated, present a sadly limited corpus of text for the purposes of traditional analysis.

In this short series of posts we will subject the Voynich Manuscript to a range of text analysis techniques, delving into its structure, gain horrific insight into its composition, and skeptically assessing its credibility. The manuscript has been subjected to almost fifty years of furtive attempts by cryptographers, including the [US National Security Agency](https://apps.dtic.mil/dtic/tr/fulltext/u2/a070618.pdf) and a menagerie of others from the distinguished to the deranged. We will crudely mimic some earlier results, and hopefully add our own confusion to the roiling mass of current research into the Voynich Manuscript.

**Authenticity**

Since its discovery, and throughout the ongoing unsuccessful attempts to decipher its contents, many have questioned the authenticity of the Voynich Manuscript. The theory that the entire book is a hoax, either by contemporary scribes or by more modern players, has been raised repeatedly over the years.

**Data**

Due to the diligent activity of several generations of Voynich researchers, the text of the manuscript has been transcribed into a machine-readable format. As the alphabet is unknown, there are minor uncertainties in rendering the text, leading to a number of similar but competing transcriptions. The subtle details of the various transcription efforts, and their history, are available at: <http://www.voynich.nu/transcr.html>, with the raw data available at <http://www.voynich.nu/data/>.

**Crude Manipulations**

We perform the following steps to make the data usable for our analyses. For many scenarios, we would develop a generalisable set of steps to allow conversion of many documents to an appropriate form. Until and unless, however, a new cache of documents in the same language are found, it is simpler and easier to perform these one-time steps manually.

Firstly, we delete from the text all incomplete words, as marked in the IVTFF format. This includes:

* all text in angle brackets
* all words containing ?’s
* all words containing []

Secondly, we tokenize the text and remove punctuation. The transcription of the Voynich manuscript that we have chosen uses the following punctuation:

* “.” is a space
* “,” is a potential space. For simplicity, we do not treat these as a space.

Finally, we organize the document in an appropriate form to be imported into an R data frame, or tidyverse tibble.

The above steps were performed in the Vim text editor, and the commands used are reproduced in the code below:

Show Vim text manipulation commands.

# Delete all commented lines

:\%g/\^\#.\\*/d

# Remove blank lines

:\%g/\^\$/d

# Remove "," -- assume that potential spaces are /not/ spaces.

:\%s/,//g

# Replace each folio's page marker (initial for each page)

# with its contents, followed by a comma. ( -\> f1r,)

:\%s/\^\s\*\<.{-}\>\$/\rfRos,/

:\%s/\^\<(f\[0-9\]\*\[r\|v\]\[0-9\]\*)\>\s\*\<.{-}\>\$/\r\1,/

# Remove all \<\> entries (non-greedy)

:\%s/\<.{-}\>\s\*//g

# Join all paragraphs (all newlines followed by a character

# other than a newline are removed).

:\%s/\n(\[\^\\n\])/.\\1/

:\%s/\^.f/f/

:\%s/,./,/

# Replace "high ascii" rare characters from the IVTFF with their

# ASCII representation. ()

:\%s/@(\[0-9\]{-});/=nr2char(submatch(1))/g

# Replace full stops with spaces

:\%s/./ /g

The resulting raw data file is available [here](http://www.weirddatascience.net/wp-content/uploads/2019/09/GC_ivtff_s-processed.txt). This file can be read into R simply by use of the read.csv function:

voynich\_tbl <-

read\_csv( "data/voynich\_raw.txt", col\_names=FALSE ) %>%

rename( folio = X1, text = X2 )

As a first, horrifying glance into the forms of analysis that this allows, we can now use our raw data to identify the most repeated words in the manuscript, according to our transcription. The following R code extracts the entirety of the text and encodes it as a [run length encoding](https://en.wikipedia.org/wiki/Run-length_encoding). This conveniently results in a sequential list of words and the number of times that each is repeated *in sequence*. We can then simply extract the largest number of repetitions for each word in the corpus:

Count longest word repetition sequences in the Voynich Manuscript.

library( tidyverse )

library( magrittr )

# Count the number of repeated words in the Voynich Manuscript text.

# Load the raw data

voynich\_tbl <-

read\_csv( "data/voynich\_raw.txt", col\_names=FALSE ) %>%

rename( folio = X1, text = X2 )

# Extract the text as a vector of words

voynich\_vector <-

voynich\_tbl %>%

extract2( "text" ) %>%

paste( sep=" ", collapse=" " ) %>%

str\_split( " " ) %>%

unlist

# Create a run length encoding object from the vector

voynich\_rle <-

voynich\_vector %>%

rle

# Convert rle object to a data frame and report the maximum number of repeated

# cases for each word

voynich\_repetitions <-

voynich\_rle %>%

unclass %>%

as.data.frame %>%

group\_by( values ) %>%

summarise( max\_repetitions = max( lengths ) ) %>%

ungroup %>%

arrange( desc( max\_repetitions ) )

This simple analysis shows that, in the transcription we have chosen, the longest sequences of repeated words are only three words in length, occuring a total of five times in the text. While there are many other arguments against the potential validity of the Voynich Manuscript, word repetition does in itself present a compelling reason to doubt that the text is a human language.

We have now reduced the strange and beautiful elegance of the Voynich Manuscript’s centuries-old illuminations to a crude, utilitarian abstraction. With this particular act of artistic and literary desecration complete, in the next post we will examine Zipf’s Law in more detail, and interrogate the extent to which this law supports or undermines the text’s authenticity.