# Text Mining

Alun Meredith

## 1. INTRODUCTION

The corpus analysed is a collection of 24 books about antiquity, written in the 18-19th Century. Many are translations of older texts and often volumes of a larger book. The data is html documents of OCR scans from the Google Books Library Project, separated into each page.

Manually extracting some metadata about each book (title, author, translator, year written, year translated) we can see some themes worth investigating: If it is written about/by a Roman or not, written in/about antiquity, volumes within a series, written by the same author, effect of different translators.

## 2. PRE-PROCESSING

To extract the text from each html page, xml2 was used. Using html tags to isolate the second paragraph (class = 'ocr_par'), which contained the body of text without the title. The 'ocr_cinfo' contained each word separately to avoid an error where words across newlines were merged with a naive approach. Pages were concatenated to produce books.

Using the 'tm' package a raw corpus was built from these vectors. After casefolding, whitespace, non-alphabetic characters and stopwords were removed. Stopwords removed were tm's "english", numbers and common artefacts of the OCR. A regular expression was used to remove Roman numerals and words were stemmed.

Producing a term-document matrix for 1-3 ngrams, removing sparse terms and using TF-IDF[1] before producing a cosine dissimilarity matrix.

## 3. ENTROPY

The Shannon entropy **CITE** was computed on each book before the stopwords had been removed using the entropy package.

[1]Different sparsity levels used for different purposes, analysis run both with and without TF-IDF

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

We can see that within the 5 terms with the highest entropy are the dictionary/encyclopedia style books which are expected to have the highest information density. In addition volumes of the same book are generally ranked close together with the exception of Josephus IV and Tacitus History IV, the later of which is quite a remarkable outlier. Cases where the translator changes for volumes of the same book doesn't have a noticeable impact (E.g. Decline & Fall 4 vs. 1,3,5). This suggests there isn't a strong component of things being translated in totally different ways.

### 3.1 Math Equations

You may want to display math equations in three distinct styles: inline, numbered or non-numbered display. Each of the three are discussed in the next sections.

#### *3.1.1 Inline (In-text) Equations*

A formula that appears in the running text is called an inline or in-text formula. It is produced by the **math** environment, which can be invoked with the usual `\begin. . .\end` construction or with the short form `$. . .$`. You can use any of the symbols and structures, from $\alpha$ to $\omega$, available in LaTeX[**?**]; this section will simply show a few examples of in-text equations in context. Notice how this equation: $\lim_{n\to\infty} x = 0$, set here in in-line math style, looks slightly different when set in display style. (See next section).

#### *3.1.2 Display Equations*

A numbered display equation – one set off by vertical space from the text and centered horizontally – is produced by the **equation** environment. An unnumbered display equation is produced by the **displaymath** environment.

Again, in either environment, you can use any of the symbols and structures available in LaTeX; this section will just give a couple of examples of display equations in context. First, consider the equation, shown as an inline equation above:

$$\lim_{n\to\infty} x = 0 \tag{1}$$

Notice how it is formatted somewhat differently in the **displaymath** environment. Now, we'll enter an unnumbered equation:

$$\sum_{i=0}^{\infty} x + 1$$

and follow it with another numbered equation:

$$\sum_{i=0}^{\infty} x_i = \int_0^{\pi+2} f \tag{2}$$

**Table 1: Frequency of Special Characters**

| Non-English or Math | Frequency | Comments |
|---|---|---|
| Ø | 1 in 1,000 | For Swedish names |
| $\pi$ | 1 in 5 | Common in math |
| \$ | 4 in 5 | Used in business |
| $\Psi_1^2$ | 1 in 40,000 | Unexplained usage |

just to demonstrate LaTeX's able handling of numbering.

## 3.2 Citations

Citations to articles [?, ?, ?, ?], conference proceedings [?] or books [?, ?] listed in the Bibliography section of your article will occur throughout the text of your article. You should use BibTeX to automatically produce this bibliography; you simply need to insert one of several citation commands with a key of the item cited in the proper location in the `.tex` file [?]. The key is a short reference you invent to uniquely identify each work; in this sample document, the key is the first author's surname and a word from the title. This identifying key is included with each item in the `.bib` file for your article.

The details of the construction of the `.bib` file are beyond the scope of this sample document, but more information can be found in the *Author's Guide*, and exhaustive details in the *LaTeX User's Guide*[?].

This article shows only the plainest form of the citation command, using `\cite`. This is what is stipulated in the SIGS style specifications. No other citation format is endorsed or supported.

## 3.3 Tables

Because tables cannot be split across pages, the best placement for them is typically the top of the page nearest their initial cite. To ensure this proper "floating" placement of tables, use the environment **table** to enclose the table's contents and the table caption. The contents of the table itself must go in the **tabular** environment, to be aligned properly in rows and columns, with the desired horizontal and vertical rules. Again, detailed instructions on **tabular** material is found in the *LaTeX User's Guide*.

Immediately following this sentence is the point at which Table 1 is included in the input file; compare the placement of the table here with the table in the printed dvi output of this document.

To set a wider table, which takes up the whole width of the page's live area, use the environment **table*** to enclose the table's contents and the table caption. As with a single-column table, this wide table will "float" to a location deemed more desirable. Immediately following this sentence is the point at which Table 2 is included in the input file; again, it is instructive to compare the placement of the table here with the table in the printed dvi output of this document.

## 3.4 Figures

Like tables, figures cannot be split across pages; the best placement for them is typically the top or the bottom of the page nearest their initial cite. To ensure this proper "floating" placement of figures, use the environment **figure** to enclose the figure and its caption.

This sample document contains examples of **.eps** files to be displayable with LaTeX. If you work with pdfLaTeX, use files in the **.pdf** format. Note that most modern TeX system will convert **.eps** to **.pdf** for you on the fly. More details on each of these is found in the *Author's Guide*.

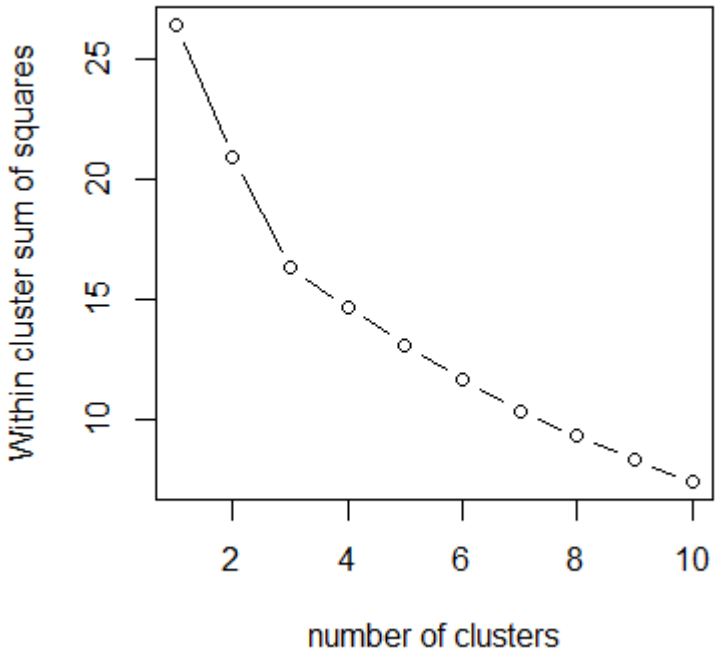


**Figure 1: A sample black and white graphic.**

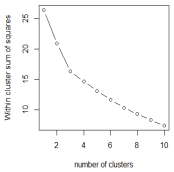

**Figure 2: A sample black and white graphic that has been resized with the `includegraphics` command.**

As was the case with tables, you may want a figure that spans two columns. To do this, and still to ensure proper "floating" placement of tables, use the environment **figure*** to enclose the figure and its caption. and don't forget to end the environment with figure*, not figure!

## 3.5 Theorem-like Constructs

Other common constructs that may occur in your article are the forms for logical constructs like theorems, axioms, corollaries and proofs. There are two forms, one produced by the command `\newtheorem` and the other by the command `\newdef`; perhaps the clearest and easiest way to distinguish them is to compare the two in the output of this sample document:

This uses the **theorem** environment, created by the `\newtheorem` command:

THEOREM 1. *Let $f$ be continuous on $[a, b]$. If $G$ is an*

**Table 2: Some Typical Commands**

| Command | A Number | Comments |
|---|---|---|
| `\alignauthor` | 100 | Author alignment |
| `\numberofauthors` | 200 | Author enumeration |
| `\table` | 300 | For tables |
| `\table*` | 400 | For wider tables |

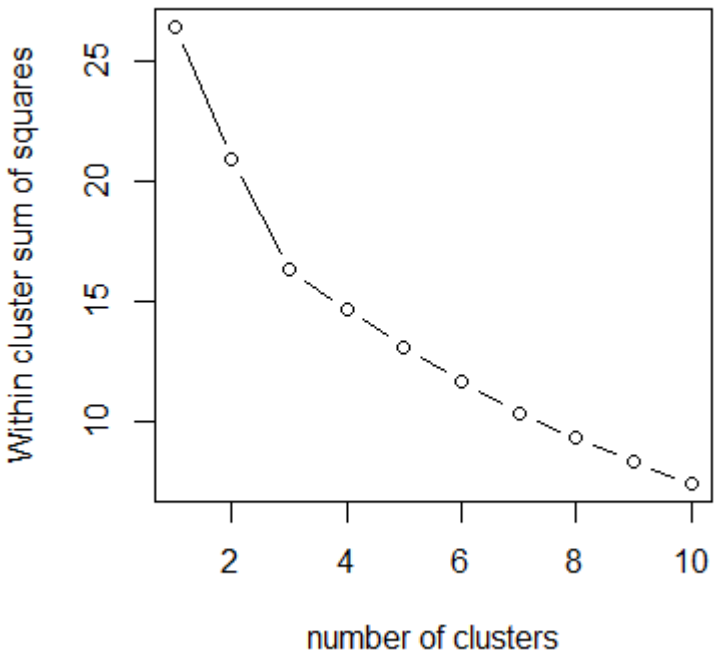


**Figure 3: A sample black and white graphic that needs to span two columns of text.**

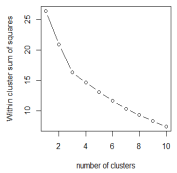


**Figure 4: A sample black and white graphic that has been resized with the `includegraphics` command.**

*antiderivative for $f$ on $[a, b]$, then*

$$\int_a^b f(t)dt = G(b) - G(a).$$

The other uses the **definition** environment, created by the `\newdef` command:

*Definition 1.* If $z$ is irrational, then by $e^z$ we mean the unique number which has logarithm $z$:

$$\log e^z = z$$

Two lists of constructs that use one of these forms is given in the *Author's Guidelines*.

There is one other similar construct environment, which is already set up for you; i.e. you must *not* use a `\newdef` command to create it: the **proof** environment. Here is a example of its use:

Proof. Suppose on the contrary there exists a real number $L$ such that

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = L.$$

Then

$$l = \lim_{x\to c} f(x) = \lim_{x\to c} \left[ gx \cdot \frac{f(x)}{g(x)} \right] = \lim_{x\to c} g(x) \cdot \lim_{x\to c} \frac{f(x)}{g(x)} = 0 \cdot L = 0,$$

which contradicts our assumption that $l \neq 0$. □

Complete rules about using these environments and using the two different creation commands are in the *Author's Guide*; please consult it for more detailed instructions. If you need to use another construct, not listed therein, which you want to have the same formatting as the Theorem or the Definition[**?**] shown above, use the `\newtheorem` or the `\newdef` command, respectively, to create it.

## A *Caveat* for the TEX Expert

Because you have just been given permission to use the `\newdef` command to create a new form, you might think you can use TEX's `\def` to create a new command: *Please refrain from doing this!* Remember that your LATEX source code is primarily intended to create camera-ready copy, but may be converted to other forms – e.g. HTML. If you inadvertently omit some or all of the `\def`s recompilation will be, to say the least, problematic.

# 4. CONCLUSIONS

This paragraph will end the body of this sample document. Remember that you might still have Acknowledgments or Appendices; brief samples of these follow. There is still the Bibliography to deal with; and we will make a disclaimer about that here: with the exception of the reference to the LATEX book, the citations in this paper are to articles which have nothing to do with the present subject and are used as examples only.

# 5. ACKNOWLEDGMENTS

This section is optional; it is a location for you to acknowledge grants, funding, editing assistance and what have you. In the present case, for example, the authors would like to thank Gerald Murray of ACM for his help in codifying this *Author's Guide* and the **.cls** and **.tex** files that it describes.

# APPENDIX

# A. HEADINGS IN APPENDICES

The rules about hierarchical headings discussed above for the body of the article are different in the appendices. In the **appendix** environment, the command **section** is used to indicate the start of each Appendix, with alphabetic order designation (i.e. the first is A, the second B, etc.) and a title (if you include one). So, if you need hierarchical structure *within* an Appendix, start with **subsection** as the highest level. Here is an outline of the body of this document in Appendix-appropriate form:

## A.1 Introduction

## A.2 The Body of the Paper

### *A.2.1 Type Changes and Special Characters*

### *A.2.2 Math Equations*

*Inline (In-text) Equations.*

*Display Equations.*

### *A.2.3 Citations*

### *A.2.4 Tables*

### *A.2.5 Figures*

### *A.2.6 Theorem-like Constructs*

*A Caveat for the TEX Expert*

## A.3 Conclusions

## A.4 Acknowledgments

## A.5 Additional Authors

This section is inserted by LATEX; you do not insert it. You just add the names and information in the `\addition-alauthors` command at the start of the document.

## A.6 References

Generated by bibtex from your .bib file. Run latex, then bibtex, then latex twice (to resolve references) to create the .bbl file. Insert that .bbl file into the .tex source file and comment out the command `\thebibliography`.

## B. MORE HELP FOR THE HARDY

The sig-alternate.cls file itself is chock-full of succinct and helpful comments. If you consider yourself a moderately experienced to expert user of LaTeX, you may find reading it useful but please remember not to change it.