Computational Methods for Determining the Similarity between Ancient Greek Manuscripts

Eddie Dunn[1], **Curry Guinn**[1], **and George Zervos**[2]
[1]Department of Computer Science, University of North Carolina Wilmington, United States
[2]Department of Philosophy and Religion, University of North Carolina Wilmington, United States

Abstract - *This paper describes research applying computational document classification methods to the domain of biblical paleography. Ancient manuscripts were preserved by the laborious process of hand-copying from prior versions. The scribes often made changes in spelling, word usage and syntax. Further, scribes might delete or alter passages that reflect a theology or understanding different from the scribe's contemporary view. Additional material not found in the prior version(s) might be inserted, perhaps combining information from other sources. In this paper, we examine over a hundred different versions of a single apocryphal gospel, the Protoevangelium of James, in order to group the documents into families of related documents in order to better understand the history of the document and how it evolved over time. This research uses the computational techniques of k-means analysis, hierarchical clustering, and correspondence analysis to find similarities and differences between documents. These results are then compared to the work of Daniels and Zervos, scholars in the field of biblical paleography who have studied this gospel.*

Keywords: Natural language processing, machine learning, clustering.

1 Introduction

Paleography is the study of ancient writing including deciphering, understanding, and dating manuscripts. Prior to the printing press, documents were preserved by hand-copying existing sources. In the ancient world, because spelling, vocabulary, and syntax were not standardized and changed over time, scribes often made modifications in order to make the text more accessible to current readers. If multiple prior versions existed, a scribe might decide to merge content from those documents or choose content from one source over others. Further, because of theological issues, a scribe might decide to delete or amend passages that seemed to contradict the scribe's current theological understanding. Passages might also be inserted from other (perhaps unrelated) sources in order to introduce or reinforce a particular theological point. As a result of this process, an ancient biblical manuscript like Genesis or The Gospel of Mark exists today in hundreds of different versions. Which version is the "correct" version? Which version most closely resembles the earliest (perhaps no longer extant) version? Can we tell which documents were the source documents for later versions? Can we trace the evolution of a document, and thus see the influences of evolving linguistics and theologies?

In this paper, we will explore techniques that have been used in authorship attribution, document classification, and data visualization in order to explore these questions. We will analyze a collection of documents which are all versions of the *Protoevangelium of James*, an apocryphal manuscript whose original dates may be from 200 CE or earlier. These documents have been studied by biblical paleographers, B. Daniels [1] and George Zervos [9], and we will compare the results using these computational techniques with these researchers' prior analyses. The goal of this study is to 1) corroborate past results, 2) discover new connections between documents, and 3) suggest to paleographers particular features or passages that deserve more exploration.

2 Textual criticism and the *Protoevangelium of James*

Textual criticism is the area that aspires to remove errors (whether intentional or unintentional) in an attempt at coming as close to creating the "original" or source documents(s) as possible. The hallmark of this type of work is a critical apparatus to show variant readings alongside a primary text (also called a base text).

The *Protoevangelium of James* (PJ) most likely dates back to the middle to later part of the 2nd Century CE [10]. This document has been known by several names. In the earlier years of its life it was likely called Book of James as it is referred to in the writings of Origen who died in the middle of the 3rd Century CE [2]. As its most commonly known name in the literature today implies, proto-gospel means just that -- it is a story before the gospels or life of Jesus. It seems to have been composed largely in reaction to the accusations by contemporary critics that were assaulting the burgeoning religion on the grounds that the parents of its messiah were commoners. The writer of this document portrays Joseph as a rich building contractor and Mary as herself being immaculately conceived and brought up (with her chastity protected) as a revered temple virgin perhaps in direct response to these accusations.

Figure 1: A scanned page of the Protoevangelium of James from Bodmer V, a, dated c. 200.

This document is also intriguing as the oldest existing manuscript is complete and it is significantly different from its closet contemporary versions as well as the majority of the later surviving examples (Figure 1). PJ also enjoys the luxury of, while not being part of the canon, being widely copied and distributed throughout the ancient world, especially in those eastern traditions with highly developed Mariological themes. Mariology is, as its name implies, the study of Mary the mother of Jesus. This term has a much more profound meaning to the traditions that evolved in the eastern world. In fact, there are Eastern Orthodox feast days established based on information in this document. While all of the documents we will be performing computations on are in Greek, there are surviving copies of this document in many languages including Coptic, Syriac, Ethiopic, Armenian, Georgian, and Slavonic. There is also a scholarly notion that an Arabic copy might have "influenced Qur'anic and later Islamic understandings of the place of Mary in the Christian tradition" [5], yet another way this text has impacted western religion.

For this study, the focus is on two separate collections of this gospel. These collections that are the basis for the dataset come from the dissertations from Duke University of Daniels (BD) [1] and Zervos (GZ) [9]. These collections are both presented in their own critical apparatus. A critical apparatus in this context is an accepted way of showing how different copies of the same documents vary (called variant readings). There are over 167 extant Greek versions of the PJ. Scholars have found that the earliest copy (from the Bodmer V collection [8]) is decidedly different from the base text used by Daniels and the base text used by GZ.

3 Authorship attribution and document classification techniques

3.1 Authorship attribution techniques

While this research does not attempt to identify particular authors, the techniques employed in authorship attribution are relevant as an introduction. While there were previous attempts dating back to the 19th century at using statistical measures in attributing authorship, it was not until the publication of *Inference and Disputed Authorship: The Federalist* by Mosteller and Wallace in 1964 that this area of "non-traditional" authorship attribution study gained widespread attention [5]. Previous work had attempted to use features such as average sentence length and rate of use of articles and pronouns. They found that while the rates of use in the case of some words such as "the" did not vary in a statistically significant manner from author to author, the use of what they refer to as connector words, such as "upon", can vary by as much as 3 standard deviations. Mosteller and Wallace used such features as word counts and rate of use of specific, non-article or pronoun words. By examining the distributions of individual words it was discovered that some word rates were best described by a Poisson distribution and others were better approximated with a negative binomial distribution. Bayesian inference was then applied using the probabilities calculated using the appropriate distribution. Their analysis was ultimately to come down on the side of supporting the historical notion that Madison was likely the author of the 12 then-disputed Federalist papers. Their study also outlines a basic work flow of technique application that is still followed.

3.2 Document classification techniques

More recently, with the profusion of massive amounts of textual data via the internet, document classification techniques have been used to compute the similarity between documents. The core of these techniques relies on using term frequency (TF) and inverse document frequency (IDF). Term frequency can be computed as simply taken a count of a term within a document. To prevent a bias towards longer documents, this value may be normalized by a variety of techniques. In this study, we normalize term frequency by dividing by the maximum frequency of any word in the set of documents. f(t, d) is the frequency of a word in a document"

$$\mathrm{tf}(t,d) = \frac{\mathrm{f}(t,d)}{\max\{\mathrm{f}(w,d) : w \in d\}}$$

Inverse document frequency weights each feature in inverse proportion to its relative occurrence, thus giving infrequently used words higher importance. We employ the standard IDF formula:

$$\mathrm{idf}(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

where *t* is the term or word, *D* is the set of documents, |*D*| is the number of documents, and the denominator is the number of documents where the term *t* appears.
These two measures can them be combined to compute the relative importance a word is to a document in a measure known as the term frequency-inverse document frequency (tf-idf):

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \times \mathrm{idf}(t,D)$$

For each feature selected (and as described below in our Research Experiment we will use features other than complete words), we compute a log normalized vector in multi-dimensional space of TF-IDF values.

3.3 Machine learning techniques

We then employ a number of well-established techniques in machine learning/document classification such as k-means analysis, hierarchical clustering, as well as DCA correspondence analysis, a technique often used in ecologists in the study of populations [6]. For all of these techniques, we use the statistical programming package, R, which contains libraries to support all of these analyses plus corresponding visualization tools.

3.4 Similar research applied to ancient texts

Research has been done by Finney [3] in the study of ancient manuscripts where he employs similar techniques in his analysis of ancient documents. His work focuses on a difficult (and different) problem in comparing different versions of the same document: namely, automatic alignment of the texts. The alignment problem is also commonly encountered in machine translation.

4 Research Experiment

4.1 Data set

All of the versions of the Protoevangelium of James (PJ) used in this study are in Greek. Our analyses examine 135 documents: 89 manuscripts analyzed by Daniels [1], 45 by Zervos [9], and Bodmer V (the oldest known version of PJ) [8]. Using OCR-software (Read Iris Pro), the original documents are scanned and converted to UTF-8 character codes. Then each document was put into an HTML-like document format for review by human readers where proofreading was done in multiple passes by multiple persons to better ensure data accuracy and integrity.

Our experiments examined the full set of documents (135) as well as a subset of the oldest 32 manuscripts (those definitively dated before 1100 CE).

4.2 Feature selection

This research uses a composite feature space consisting of unigrams (single Greek words), bigrams (neighboring pairs of words), and character sequence n-grams of length 2, 3, 4, and 5. The reason for the inclusion of character n-grams was to capture variations in letter patterns and spelling that occurred in the millennia from the time of the first document to the most recent documents in the collection. Obviously, there will be overlap in the feature set (as some words are 5 characters or less); however, an analysis conducted by removing some of those character n-gram features produced slightly worse results (not included in this article). The feature space is very large, over 85,000 unique tokens. Fortunately, the techniques employed all typically work very well with large feature spaces.

4.3 Analysis techniques

Using the "vegan" library within R [6], we computed several analyses:

- Sorensen (Bray/Curtis) similarity index cluster analysis
- Detrended correspondence analysis (DCA)
- Nonmetric Multidimensional Scaling (NMDS)
- Canonical correspondence analysis (CCA)
- K-means clustering

for each document and for each chapter within the document.

5 Results

Using the ordination plotting function (ordiplot in R), we can visualize the results of DCA, CCA and NMDS

analyses. For brevity, we present only the ordiplot from the DCA in this paper. For the full set, the DCA ordination plot is presented in Figure 2. For the subset including only the oldest 32 manuscripts, refer to Figure 3. Note that the document numbers used are the ones given by BD and GZ.

Cluster analysis visualization using dendrogram (tree) is also an intuitive appealing technique for examining differences and simulations. Using Bray-Curtis and the hclust function within R we obtained the results for the complete set (Figure 4) and the oldest 32 documents (Figure 5).

5.1 Comparison with Daniels and Zervos

In their respective PhD dissertations, Daniels [1] and Zervos [9] analyze the documents based on their own observations focusing on the inclusion or absence of various passages. Further, they made use of where the documents were located or found. Several subsets of documents were found at particular monasteries. Often, multiple versions of the document found at one monastery were found to be extremely similar, presumably because these documents were copies of each other and earlier documents. The computational analysis we performed does not take into account any of this information.

One group of documents is described as the largest by both BD and GZ consisting of 003, 005, 103, 115, 118, 201, 204, 206, 214, 502, and 609. The hierarchical plot (Figure 4) shows all of these manuscripts as being in the same cluster. This corresponds to groups E and G from the DCA groupings (Figure 2). GZ suggests that 612 and 409 might also be close, and we do place them in the same group in the hierarchical clustering as well as being in Group C on the DCA groupings (Figure 2). However, it is not near the rest of the mentioned manuscripts. This is intriguing and should be examined with the knowledge of the Greek language.

Another family widely agreed upon is the one made up from 112, 208, 212, 402, 407, 511, 616, 702, 705, 709, and 901. All with the exception of 702 and 709 came from the St. Panteleimonus monastery in Athens (so we identify this group by the same name). It is interesting to note that GZ specifies that there are two sub-groups in this family consisting of 511, 702, and 709 that follow 212 and 616, 705, 901 that follow 208,402. The hierarchical plot confirms all of this information (Figure 4). Also notice this group corresponds with Group D in our DCA groupings (Figure 2).

The group 601, 606 is also highlighted. Document 601 has in its sub-group 512, 615, 619 and 606 with sub-group 617, 703, 707, 803, 805, and 902. Our plot also confirms these observances (see DCA Group I, Figure 2). In looking at the tree it seems proper to place the 601, 606 group and its associated documents into the Panteleimon family. It is also worth noting that 621, 631, and 704 (Group H) are mentioned by GZ as being in this group as well. It was found that while these three documents were indeed found to be similar, they were placed a good distance from the rest of its other neighbors and should be examined by experts for further analysis.

Next we examine the group from the monastery of Vatopedi on Mt. Athos. This group consists of 111, 218, 501, 513, 801, and 802. Again this is all confirmed in both the hierarchical (Figure 4) and DCA (Figure 2) plots where it shows up as group E. The group that is now being called the Jerusalem group consists of 202, 508, 603, 622, and 708. These are lumped in with our DCA Group A (Figure 2). As an aside, GZ notes the similarity of 509 and 604 and also 210 and 220. This information is also confirmed in our plots.

5.2 Chapter by chapter results

The results of the full document analysis while providing a great deal of information and striking results also shows some confusion with respect to some documents. This is especially clear with the DCA Only Old plot. There are two dynamics that contribute to this effect: The first is that the letter groupings were established from the full set which includes the very tightly grouped but also very different traditions that do not seem to be present in the earliest documents in our set. The second reason is that a more detailed examination of the chapters shows that there is a great deal of variation contained within each manuscript in certain sections versus others. This is where visualization of the correlation matrices (corrplot in R) truly shines. It provides a way to see how each chapter breaks down, and that indeed we have situations where in one chapter the scribe is using content from one tradition and then another in different sections. Figure 6 presents the correlation plots for each chapter of the oldest 32 documents. Clearly, there are families that emerge per chapter that are not as apparent when examining the entire document as a whole.

6 Conclusion and future work

The identification of "families" of related copies of the same base document has traditionally required laborious and detailed study of the documents including some knowledge of the physical history of the documents. The computational techniques described in this paper produced results that were remarkably similar to scholars' analyses. What makes this result particularly exciting to the paleographers studying this document is that there are scores and perhaps hundreds

more copies of the *Protoevangelium of James* that have not been carefully analyzed yet. These algorithms can automatically suggest which documents are related and which are dissimilar. Further, these algorithms can actually point to which features are most relevant for distinguishing the documents. Such tools will be invaluable to researchers as they incorporate new versions of the documents into their study.

While this study has focused on the PJ document, there are no limitations to language or document set. This methodology can be used to explore any collection of ancient texts to suggest document family histories.

7 References

[1] Daniels, B. (1956) *The Greek Manuscript Tradition of Protoevangelium Jacobi*, Unpublished PhD Dissertation, Duke University Durham, NC.

[2] Ehrman, B and Plese, Z. (2011) *The Apocryphal Gospels Texts and Translation*s, Oxford University Press, New York, NY.

[3] Finney, T. (2012) How to Discover Text Groups, http://www.tfinney.net/Groups/index.xhtml.

[4] Foster, P (2009) The *Apocryphal Gospels: A very short introductio*n, Oxford University Press, New York, NY.

[5] Mosteller, F & Wallace, D. (1964). *Inference and disputed authorship: The Federalist*. Addison Wesley, Boston, MA.

[6] Oksanen, J. (2013) *Multivariate Analysis of Ecological Communities in R: vegan tutorial*, http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf.

[7] Schaps, D. (2011) *Handbook for Classical Research*, Routledge, New York, NY.

[8] Testuz, M. (1958) *Papyrus Bodmer V: Nativite de Marie*, Geneva: Bibliotheca Bodmeriana.

[9] Zervos, G. T. (1986) *Prolegomena to a Critical Edition of the Genesis Maria (Protoevangelium Jacobi)*, Unpublished PhD Dissertation, Duke University Durham, NC.

[10] Zervos, G. (1994) – Dating the Protoevangelium of James: The Justin Martyr Connection – *SBLSP*, pp. 415-34.

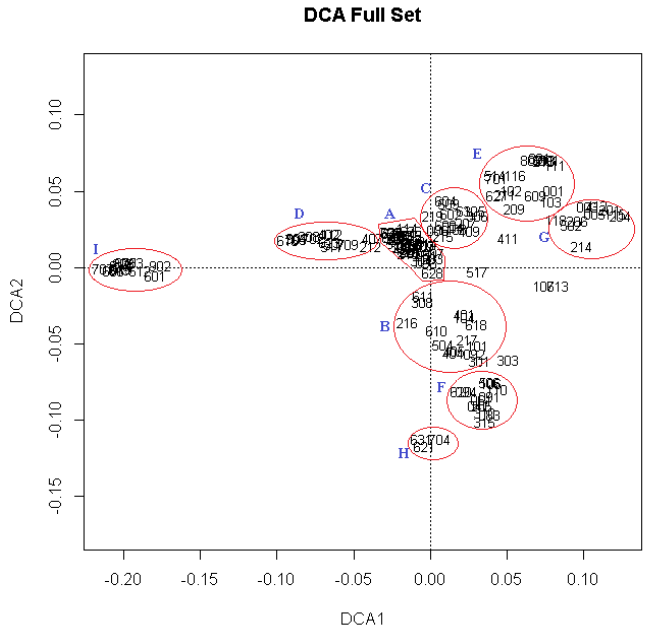

Figure 2 Detrended Correspondence Analysis (DCA) plot of the full data set

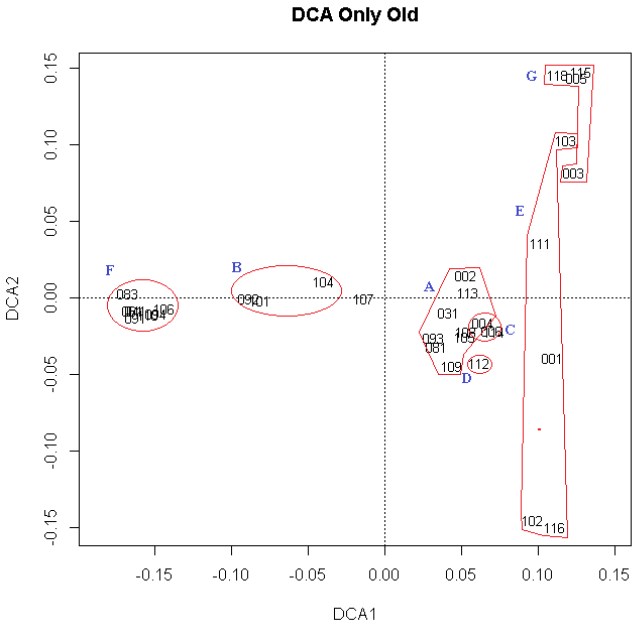

Figure 3 DCA plot of the subset of 32 oldest manuscripts

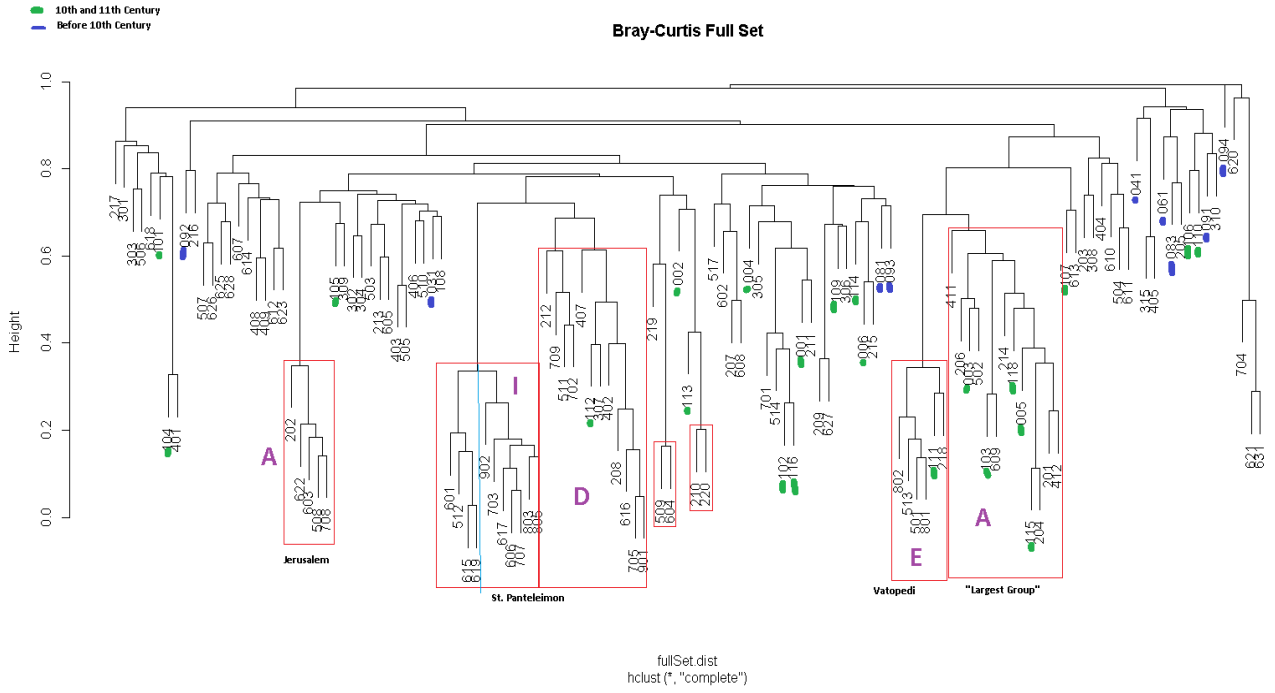

Figure 4 Hierarchical Clustering with Bray-Curtis on full data set

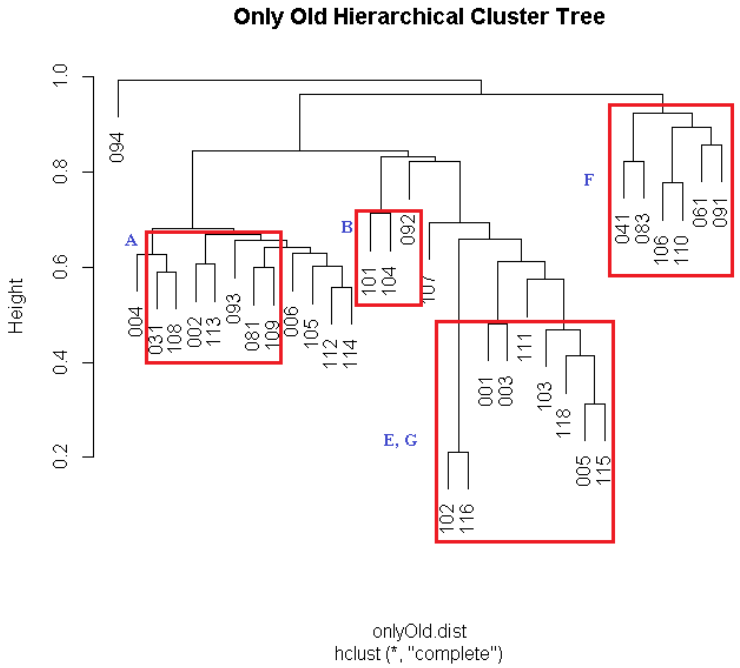

Figure 5 Hierarchical clustering using Bray-Curtis on the 32 oldest documents

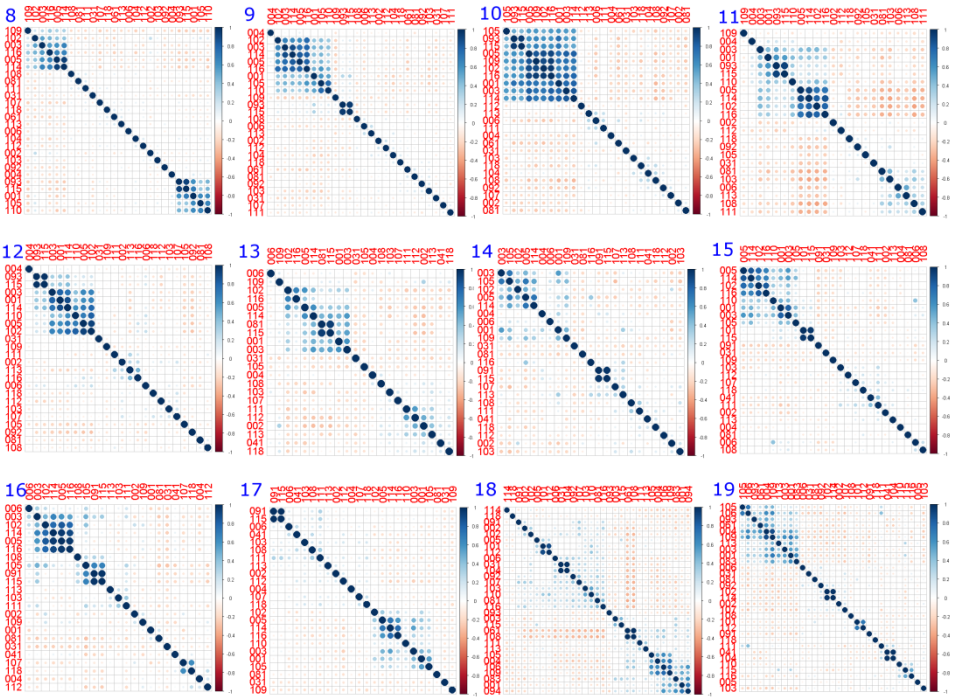

Figure 6 Correlation plot chapter-by-chapter of the 32 oldest documents