*Non-Canonical Acts and their Topical Distribution**

Christian Vrangbæk*†, Eva Vrangbæk†, Márton Kardos, Kristoffer Nielbo, and Jacob Mortensen[1]

[1] *Aarhus University, Ringgaden 1, 8000 Aarhus C, Denmark*

Abstract

This paper investigates how we can use topic modelling to characterize and place four apocryphal, i.e. non-canonical, "Acts stories" in a corpus of ancient Greek texts. In the research field of New Testament Apocrypha, there remains uncertainty concerning the classification of apocryphal text The analysis serves the purpose of creating a structured ontology to be used in classifying New Testament Apocrypha. We attempt to show that topic modelling can be a viable tool in classifying and characterizing these texts. The results show that a) our four target texts of non-canonical "Acts stories" are ambiguous and multifaceted in their topical distribution compared to other texts in the corpus, and b) that topic modelling is a viable tool in this analysis.

Keywords

Apocrypha, New Testament Studies, topic modelling, classification.

1. Introduction

In the field of New Testament Studies, classification of the heterogenous group of non-canonical texts, i.e. apocrypha, is disputed. The main problem is that the taxonomy of modern scholars largely reproduces the ancient classifications which were shaped during the 4th and 5th-century debate of canon which tends to lead to a binary classification between either canonical or non-canonical, and, moreover, to utilize conceptualized labels such as "Gnosticism" and "Encratite" which does not do justice to the complexity of the topical variety in these texts [1, 2, 3].

To contribute to this debate, we want to create an ontology, i.e., a structured framework, out of (apocryphal) textual data with the overarching goal of establishing a computationally driven classification system for New Testament Apocrypha within the context of the semantic web. We will investigate the topical distribution of four non-canonical acts. Non-canonical acts are stories from roughly 2nd-3rd century CE about early

* *Short Paper for SemDH2024: First International Workshop of Semantic Digital Humanities. Extended Semantic Web Conference. Hersonissos, Greece, May 26-27, 2024.*

* Corresponding author.

† C. Vrangbæk is first and main author, E. Vrangbæk is second author with equal contribution. M. Kardos and K. Nielbo are authors of code and visualization for topic modelling, J. Mortensen hosts the text database.

chv@cas.au.dk (C. Vrangbæk)

Christian apostles' legendary deeds and speeches [11, 3, 23]. In this paper, we investigate the stories called *Acta Joannis*, *Acta Thomae*, *Acta Barnabae* and *Acta Philippi* [25]. These texts have been chosen as tests to create the basis of including more texts.

We want to contribute to define categories, textual characteristics, and relationships between texts by building a structured ontology. While the creation of this ontology extends beyond the scope of this paper, our present study of integrating topic modelling in research on New Testament Apocrypha serves a crucial step towards this endeavor. By utilizing ontological relationships and semantic annotations, in this context as topical distributions, this study explores the potential for integrating the classical theological concepts present in New Testament Apocrypha with modern computational methods, thereby bridging the gap between traditional textual analysis and advanced semantic technologies within the context of the semantic web. Our working question is *how and in what way can this model contribute to the discussion of classifying non-canonical acts in a corpus of ancient Greek texts*?

2. Data and Methods

Our starting point is that digital methods are not quick and magic tools to solve complex questions [16], rather we find that constant and critical exchange between traditional and new methods in qualitative collaboration is the way forward [8, 9, 24]. Due to the circumstance that this study originates from a research discipline where computational methods are not embedded, we find it important to provide a detailed description — tending to tedious — of the methodological process, since we believe that this can help to bridge the disciplinary divides with a low-practical how-to-style.

2.1. Text Corpus and Preprocessing

The first step is to provide our text corpus. The corpus serves in our experiments as the literary context for the target texts, for which reason a historical relation between corpus and target texts are needed. Our target texts, the four apocryphal acts, are written in the Ancient Greek language, for which reason we have gathered a database consisting of 2153 Ancient Greek Texts. These texts are retrieved from the *Perseus Corpus, First1KGreek*, *Pseudepigrapha.org* and *Deutsche Bibelgesellschaft* [27, 28]. The selection of this corpus is chosen based on its relevance for our target texts. The corpus largely sets the parameters for our topic modelling at the later stages of the experiments. The Ancient Greek texts are in a solid machine-readable state. For the corpus text to be prepared for calculations we perform a set of preprocessing steps, so that the text is cleaned and lemmatized.

We clean out for any Latin characters, digits, extra whitespaces and stop words. The textual cleaning follows the logic of standard natural language processing utilities. The models for cleaning and parsing the text were built with a transformer-based pipeline called *OdyCy* [4]. This pipeline is a single-transformer pipeline that uses the following workflow: Transformer — Parser — Morphologizer — Lemmatizer. The transformer is built on *Ancient Greek Bert* [5]. The parser, morphologizer and lemmatizer follows the

infrastructure of SpaCy [18]. Before running topic modelling, we had already preprocessed the input text, so we set max_df and min_df to 1.0, since we did not want to ignore any terms in this experiment [26].

When the textual preprocessing is done, we move on to *vectorization*. Vectorisation is a highly qualitative choice, almost a language philosophical step, in which we must decide how to represent our textual corpus as numerical vectors [7]. In this case, we employ the method of *term frequency-inverted document frequency* (TF-IDF) which represents the logarithmic scale between a term's frequency and the inverted frequency of total documents in the corpus [6, 7, 26].

2.2. Topic Modelling: Non-Negative Matrix Factorization

Topic modelling is a much-utilized tool when engaging in natural language processing classification tasks, mostly for the purpose of assigning a category based on the most probable topic to a text in a corpus [13, 17, 26]. This is also partly our aim, although we do see possibilities of detecting more complex layers in the topical distribution besides a text being only a part of one single category, rather, topic modelling gives due credit to the complex topical distribution and its significance for the position of our target texts in the corpus. Måske: moreover. Topic modelling is a way of structuring our textual data to be included in a future knowledge graph and similar semantic web technologies.

Topic modelling assumes that a corpus of texts consists of topics and that these topics are comprised by the words of the corpus. We utilize the kind of topic modelling called Non-negative Matrix Factorization (NMF) [10, 15]. NMF is an approach that decomposes a high-dimensional term-document matrix, i.e., a matrix consisting of the corpus documents in columns, the words in rows, and their occurrence-values in the cell entries. The occurrence-values are, as mentioned above, chosen to be TF-IDF. Based on optimization, the method of NMF calculates topical patterns by associating words with topics and topics with documents. The NMF model, so to say, produces latent topical patterns by grouping similar and co-occurring words in the corpus [14]. These topics are then backtracked to fit to the documents. The number of topics to choose is important for the analytical task. If we choose too many topics, there were no interpretable coherence, so based on our knowledge of the corpus and trial-and-error process the most robust output in topics was 10 topics [12].

The 10 topics and a selection of their top words are:

0. city, war, ruler, Hellene, land [πόλις, πὸλεμος, βασιλεύς, ἕλλην, χώρα]
1. god, soul, heaven, word, Christ [θεός, ψυχή, λόγος, χριστός]
2. god, lord, people, human being [θεός, κυριός, λαός, ἄνθρωπος]
3. body, matter, air, earth, water, fire [σῶμα, ὕλη, ἀήρ, γή, ὕδωρ, πῦρ].
4. Zeus, child, desire, Cypris, Apollo [ζέυς, παῖς, ἔρως, κύπρις, ἀπόλλων]
5. part, character, being, necessity, cause [μόριον, τρόπος, ούσία, ἀνάγκη, αἰτία]
6. law, justice, city, witness, possession [νόμος, δίκη, πὸλις, μὰρτυς, χρῆμα]
7. reason, city, law, fortune, favor [λόγος, πόλις, νόμος, τύχη, χάρις]
8. circle, angle, center, sign, appearance [κύκλος, γωνία, κέντρον, σημεῖον, ἐπιφάνεια]

9. reason, nature, virtue, human being [λόγος, φύσις, ἀρετή, ἄνθρωπος]

How the topics in a topic modelling are to be interpreted in a qualitative way is disputed in scholarship [19, 20, 21], so we decided to be transparent about our process, which was as follows: Our domain knowledge enables us to notice and interpret the words of the topic into a meaningful collective description, like e.g. "Historical-political topic" for Topic 0 and "Element philosophy" for Topic 9. To nuance this interpretation of topics, we investigated our corpus for those texts with the cleanest topic distribution. An example of a clean topical distribution is the *Book of Jeremiah* from the Old Testament, which has an almost 95% of Topic 2. Another example is Aristotle's *Problemata*, which is dominated by Topic 9. Finally, we have a metadata set that groups our texts in predefined genres, like, for example, Jewish Philosophy, Tragedies and New Testament Gospels. These genre categories did not influence the topic model's calculations but can be used by us as a navigating tool to interpret the topics. For example, we can see that Topic 7 and Topic 6 have overlapping words about justice and city, but we can see that the texts that are dominated by Topic 6 are rhetorical texts like Demosthenes, whereas Topic 7 dominate many of Philo of Alexandria's texts as well as Libanius' *Declamationes* which are more philosophical. These extra steps enable us to distill the words of the topic into a qualified label or notion about how to describe the topic presented.

3. Analysis 1: Topical Distribution of Four Non-Canonical Acts Stories

In our first analysis, we interpret the topical distribution of the *Acta Joannis*, *Acta Thomae*, *Acta Barnabae* and *Acta Philippi* [29]. The results can be seen in Figure 1, which shows four pie charts over the topical distribution of the four non-canonical acts.

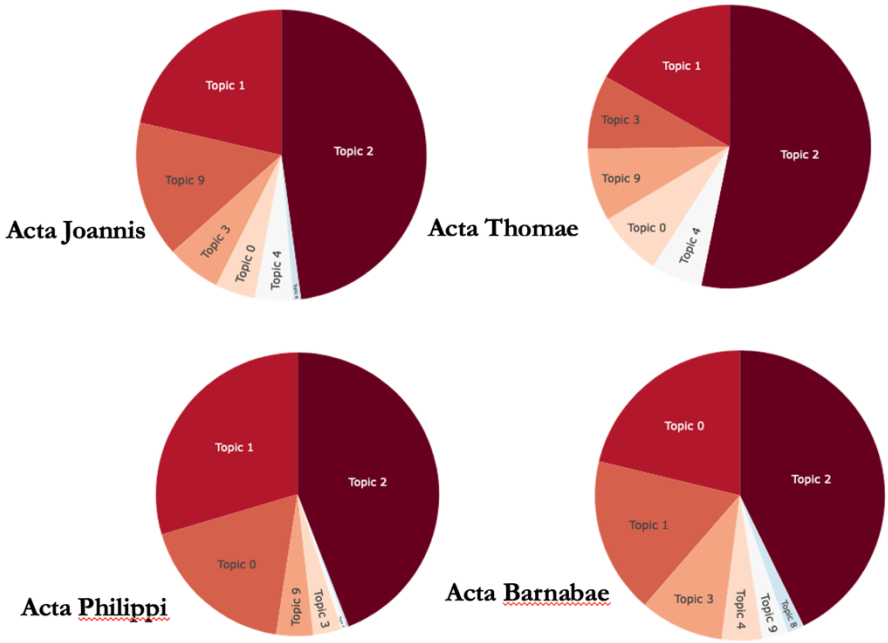

Figure 1: Pie chart visualization of topical distribution of the four target texts in the corpus of Ancient Greek texts. From top left to bottom right the four Acts stories are marked in the order *Acta Joannis, Acta Thomae, Acta Philippi, Acta Barnabae*.

Overall, all the four target texts are dominated by topic 2 and generally they share similar groups of topics, but the distribution is not equal. If we wanted to group the target texts together in the corpus on a coarse level, they would be set in the group of topic 2. However, this would not be a discovery. Where topic modelling can lead us further is in the presence and distribution of minor topics compared to each other.

The topics of ***Acta Joannis*** are distributed over mainly topics 2, 1, 9, 3, 0, 4 and 6. Topic 1 is, like topic 2, also a theological topic. The presence of topic 9, 3 and 4 is revealing of the content of the story, since these topics represent philosophically oriented words which tell about how the story engages in Hellenistic religion and philosophy. Topic 0 is the historical-political topic which is understandable since the text narrates sequences of events and speeches. We also see a small portion of Topic 6, which concerns justice and law, which resonates with a few scenes in the narrative. From the topical distribution, we can generally characterize *Acta Joannis* as being situated in a Jewish-Christian theological context where Hellenistic anthropocentric philosophy and religion is also present.

Concerning ***Acta Thomae*** *,* the topical distribution is similar to that of *Acta Joannis*, with a large topic 2 and 1, but *Acta Thomae* has a more equal distribution between its subtopics, 3, 9, 4 and 0. The *Acta Thomae*-story is set in legendary India, where Thomas is sent off as a missionary. From manual reading, *Acta Thomae* and *Acta Joannis* are comparable in the sense of the mix between narrative and preaching, where the preaching might drag in the more philosophically weighted topics, and this relation is also visible in the topical distribution.

When we inspect the topics of ***Acta Philippi****,* then the distribution is markedly different from the two previous narratives with Topic 0, the historical-political topic being the third most prominent. Although *Acta Philippi* in its content has a much more adventurous and mythical tone with, for example, the protagonists arriving to a city of snakes, the structural dynamic of this narrative is dominated by sequence narration where we follow event after event [23, 4]. This makes the *Acta Philippi* into a drier, journal-like text.

The large presence of Topic 0 is also visible in the topical distribution of ***Acta Barnabae***. This circumstance is probably linked to the fact that *Acta Barnabae* is situated in a church-political context of legitimizing the so-called autocephalous, i.e., independent ecclesiastical unity in Creta in the 4th century [22].

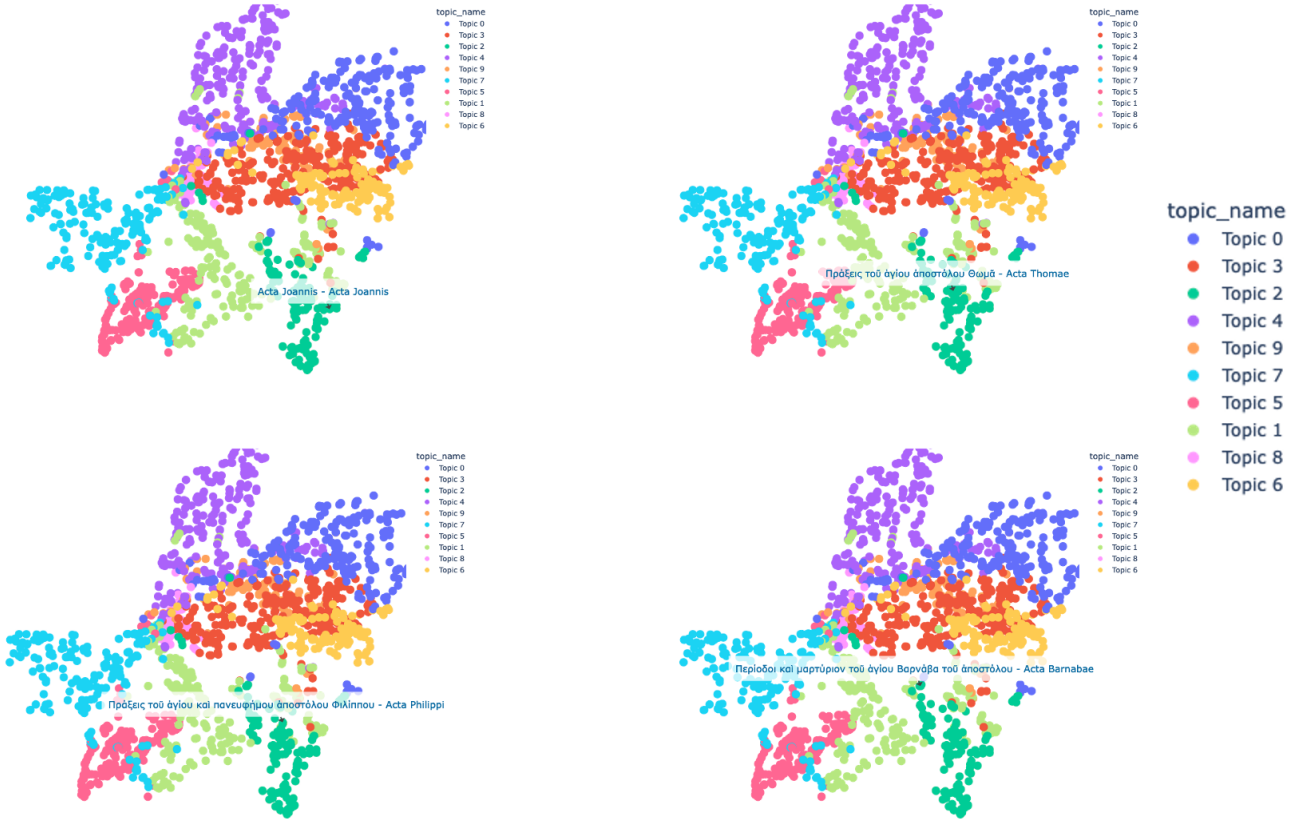

Figure 2: Four scatterplot snapshots of the same corpus with different markings of a target text. The plots' positions are based on their topical distribution. From top left to bottom right the four Acts stories are marked in the order *Acta Joannis*, *Acta Thomae, Acta Philippi, Acta Barnabae*. The figure is a frozen image from one angle of a multidimensional space. The scatterplot is created based on the topical distribution of the four texts.

4. Analysis 2: Position in Corpus

In the second part of the analysis, we want to address how the topical distribution analyzed in the previous part affects texts' position in the corpus. The position of each of the four target texts are visualized in Figure 2.

The clusters are formed based on texts with a dominant topic. Those texts which have an almost unequivocal dominance of one topic, e.g., Aristotle's *Problemata* and *The Book of Jeremiah* are placed in the outskirts of each colored cluster dragging away from the center of complexity. Conversely, those texts that are close to the center and also other topical

groups display diversity and complexity in their topical distribution. Some clusters appear like relatively demarcated cone-shapes like e.g. Topic 4 of Hellenistic religion and philosophy and Topic 2 of god, people, human, whereas topic 3 of Greek element-philosophy is more flat, which indicates that this topic is dispersed more evenly in the corpus. The topical distribution on the corpus level gives a *navigational tool* with which texts can be grouped with relative clarity. For example, it is noteworthy that classic, and almost foundational texts, of Ancient Greek language, the *Iliad* and *Odyssey*, are clearly situated in topic 4 of Greek religion and philosophy, and that later texts that are trying to imitate these, like Tryphiodorus' *Sack of Troy* (4th century CE) and Nonnus' *Dionysiaca* from the 5th century CE, almost a thousand years after Homer, have an almost identical topical distribution.

All of the *Acts* stories are placed in the dark green cluster of Topic 2, the topic of god, people, human. But it is noticeable that they all show diversity in the distribution. *Acta Joannis* and *Acta Thomae* are placed more securely in the Topic 2 cluster, whereas *Acta Philippi* and *Acta Barnabae* are drawn toward the center which might be explained by their higher percentage of the historical-political topic Topic 0.

When texts are placed in this corpus, it becomes important to iterate the basic assumption of topic modelling: that all topics are produced based on words in the corpus. This means that the corpus words are constituent of the created topics which calls for a qualitative choice of corpus texts. Our corpus consists, as mentioned, of texts that historically have shaped our four target texts, either directly or indirectly due to the circumstance that the (mostly anonymous) authors were educated people in the Greco-Roman world about whom it can be assumed that they had basic knowledge of the texts in their historical and literary context. The method of topic modelling, then, almost backtracks the literary world of the authors of our texts, of course, with the important acknowledgement that we do not have all texts which made up the author's literary context.

5. Concluding Remarks

In this analysis assisted by topic modelling, we were able to characterize and place four non-canonical acts based on their topical distribution. The topical distribution of the analyzed texts will be used to map ontological relationships and enhance semantic annotations in order to classify New Testament Apocrypha, among which the topical distribution is a major component. The results of this topic modelling analysis will thus be able to be included in a future New Testament Apocryphal Ontology.

The navigational advantages of topic modelling allowed us to inspect the target texts in a qualitatively selected corpus consisting of texts from a similar historical and literary horizon. This ensured meaningful topics. Instead of characterizing and classifying the texts based on abstractions and taxonomy from the 4th and 5th century church-political discussions on canonization, we could characterize and classify, or at least contribute to these tasks on the basis of raw content, both on a small, close scale in the topical distribution from text to text, but also on a larger scale based on the entire corpus.

Acknowledgements

The research in this article is funded by the Carlsberg Foundation in the Semper Ardens: Accelerate-project "Computing Antiquity: Computational Research in Ancient Text Corpora." We would also like to thank Deutsche Bibelgesellschaft, Pseudepigrapha.org, The Perseus-Project and First1KGreek for providing texts.

References

[1] K.L. King, No Longer Marginalized. From Orthodoxy and Heresy Discourse to Category Critique and Beyon, in: O. Lehtipuu, S. Petersen (Eds.), Ancient Christian Apocrypha, SBL Press, Atlanta, 2023, pp. 13–32, https://doi.org/10.2307/j.ctv2rh2cqj.

[2] D. Martin, New Testament Apocrypha: Introduction and Critique of a Modern Category, in: J.C. Edwards, C. Evans, C. Wassén, Ancient Christian Apocrypha, Zondervan Academic, Grand Rapids, 2022, pp. 490-511.

[3] F. Bovon, 'Die kanonische Apostelgeschichte und die apokryphen Apostelakten', in: *Die Apostelgeschichte im Kontext antiker und frühchristlicher Historiographie*, vol. 162. Berlin, New York: Walter de Gruyter, 2009. doi: 10.1515/9783110216325.

[4] J. Kostkan, M. Kardos, J. P. B. Mortensen, and K. L. Nielbo, OdyCy – A general-purpose NLP pipeline for Ancient Greek, in: Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 128–134. Accessed: Jun. 02, 2023. [Online]. Available: https://aclanthology.org/2023.latechclfl-1.14

[5] P. Singh, G. Rutten, and E. Lefever, 'A Pilot Study for BERT Language Modelling and Morphological Analysis for Ancient and Medieval Greek', in: Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, S. Degaetano-Ortlieb, A. Kazantseva, N. Reiter, and S. Szpakowicz, Eds., Punta Cana, Dominican Republic (online): Association for Computational Linguistics, Nov. 2021, pp. 128–137. doi: 10.18653/v1/2021.latechclfl-1.15.

[6] D. D. Mehare, 'Introduction to TF-IDF: To Represent Importance of Keyword within whole Dataset', IJRASET, vol. 6, no. 3, pp. 2321–2323, Mar. 2018, doi: 10.22214/ijraset.2018.3369.

[7] A. Wendland, M. Zenere, and J. Niemann, 'Introduction to Text Classification: Impact of Stemming and Comparing TF-IDF and Count Vectorization as Feature Extraction Technique', M. Yilmaz, P. Clarke, R. Messnarz, and M. Reiner, Eds., in Communications in Computer and Information Science, vol. 1442. Cham: Springer International Publishing, 2021, pp. 289–300. doi: 10.1007/978-3-030-85521-5_19.P.

[8] P. Molitor and J. Ritter, ‘Digital methods for intertextuality studies’, in: IT - Information Technology, vol. 62, no. 2, pp. 49–51, Apr. 2020, doi: 10.1515/itit-2020-0006.
[9] D. Tenen, ‘Blunt instrumentalism: On tools and methods’, in: M. K. Gold and L. F. Klein (Eds.) Debates in the Digital Humanities 2016, University of Minnesota Press, 2016
[10] D. D. Lee and H. S. Seung, ‘Learning the parts of objects by non-negative matrix factorization’, Nature, vol. 401, no. 6755, pp. 788–791, Oct. 1999, doi: 10.1038/44565.
[11] H.-J. Klauck, The Apocryphal Acts of The Apostles: An Introduction. Baylor Univ. Press, 2008.
[12] S. Vajjala, B. Majumder, A. Gupta, and H. Surana, Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems, 1st edition. Beijing Boston Farnham Sebastopol Tokyo: O’Reilly Media, 2020.
[13] T. Sherstinova, O. Mitrofanova, T. Skrebtsova, E. Zamiraylova, and M. Kirina, ‘Topic Modelling with NMF vs. Expert Topic Annotation: The Case Study of Russian Fiction’, in Advances in Computational Intelligence, L. Martínez-Villaseñor, O. Herrera-Alcántara, H. Ponce, and F. A. Castro-Espinoza, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 134–151. doi: 10.1007/978-3-030-60887-3_13.
[14] N. Gillis, ‘The Why and How of Nonnegative Matrix Factorization’. arXiv, Mar. 07, 2014. Accessed: Mar. 12, 2024. [Online]. Available: http://arxiv.org/abs/1401.5226
[15] J. Albrecht, S. Ramachandran, and C. Winkler, Blueprints for Text Analytics Using Python: Machine Learning-Based Solutions for Common Real World (NLP) Applications, 1st edition. Sebastopol, CA: O’Reilly Media, 2021.
[16] I. Uglanova and E. Gius, ‘The Order of Things. A Study on Topic Modelling of Literary Texts’, in: Proceedings of the Workshop on Computational Humanities Research (CHR 2020), Amsterdam, the Netherlands, November 18-20, 2020, pp. 57-76.
[17] [1] P. Kherwa and P. Bansal, ‘Topic Modeling: A Comprehensive Review’, EAI Endorsed Transactions on Scalable Information Systems, vol. 7, no. 24, Jul. 2019, Accessed: Mar. 12, 2024. [Online]. Available: https://eudl.eu/doi/10.4108/eai.13-7-2018.159623
[18] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python," IEEE, 2020. [Online]. Available: https://doi.org/10.5281/zenodo.1212303
[19] [1] M. Gillings and A. Hardie, ‘The interpretation of topic models for scholarly analysis: An evaluation and critique of current practice’, Digital Scholarship in the Humanities, vol. 38, no. 2, pp. 530–543, Jun. 2023, doi: 10.1093/llc/fqac075.
[20] M. Gillings and A. Hardie, ‘The interpretation of topic models for scholarly analysis: An evaluation and critique of current practice’, Digital Scholarship in the Humanities, vol. 38, no. 2, pp. 530–543, Jun. 2023, doi: 10.1093/llc/fqac075.
[21] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, ‘Reading Tea Leaves: How Humans Interpret Topic Models’, in: Y. Bengio and D. Schuurmans and J. Lafferty and C. Williams and A. Culotta, Advances in Neural Information Processing Systems 22 (NIPS 2009), 288-296.

[22] F. Cairns, ‘An early Byzantine Pseudepigraphon: the Apocryphal Acta Barnabae’, Byzantinische Zeitschrift, vol. 112, no. 1, pp. 47–66, Feb. 2019, doi: 10.1515/bz-2019-0004.
[23] F. Bovon, ‘Canonical and Apocryphal Acts of Apostles’, Journal of Early Christian Studies, vol. 11, no. 2, pp. 165–194, 2003.
[24] E. E. H. Vrangbæk and K. L. Nielbo, ‘Composition and Change in De Ciuitate Dei: A Case Study of Computationally Assisted Methods’, in: Papers presented at the Eighteenth International Conference on Patristic Studies held in Oxford 2019, Peeters, 2021, pp. 149–164.
[25] M. Bonnet, R. A. Lipsius, Acta Apostolorum Apocrypha. Vols. 1-3 Leipzig: Mendelssohn, 1903.
[26] L. Buitinck et al., *API* design for machine learning software: experiences from the scikit-learn project, ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 2013, 108—122.
[27] Code for **corpus** https://github.com/centre-for-humanities-computing/computing-antiquity.
[28] Code for **parser** https://github.com/centre-for-humanities-computing/odyCy.
[29] Code for **topic modelling** https://github.com/ankyloHryax/classic-topic.