Vassar College

Computer Science

Temporal Exploration of the Proceedings of Old Bailey

Author

Charlotte Lambert

Advisor

Jonathan Gordon

May 10, 2020

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract

The Proceedings of the Old Bailey, 1674–1913 (Hitchcock et al., 2012b) is a published record of criminal proceedings at London's central criminal court. The Proceedings primarily depict the lives of the "non-elite" population of London. This project explores these proceedings to study this specific population over the approximately 250-year time period of the publication. Because the corpus spans a significant period of history, it can be examined to identify evolving patterns related to different social groups represented in the text. This project aims to identify which computational methods can reveal interesting sociolinguistic information about this corpus. More specifically, this paper will explore unsupervised techniques like latent Dirichlet allocation (LDA) (Blei et al., 2003), Word2Vec (Mikolov et al., 2013), and Embedded Topic Modeling (ETM) (Dieng et al., 2019b) when applied to the Proceedings of Old Bailey. Additionally, temporal variants of these methods, such as Dynamic Topic Modeling (DTM) (Blei and Lafferty, 2006), Dynamic Embedded Topic Modeling (DETM) Dieng et al. (2019a), and LDA and Word2Vec manually run across different time slices, are applied to analyze the corpus over time.

Acknowledgements

First, I want to thank my thesis advisor, Professor Jonathan Gordon, for his invaluable guidance on this project and the fact that he always has an idea when I get stuck. I am also extremely grateful that he selected me to be his research assistant two years ago, introduced me to the area of natural language processing, and taught one of my favorite classes at Vassar.

I would also like to thank the second reader of this paper, Professor Nancy Ide, who provided very helpful suggestions. Additionally, I'd like to thank my unofficial third reader, Kevin Ros, for proofreading my paper and providing moral support. Furthermore, I want to express my appreciation for the entire computer science department at Vassar, I feel very lucky to be a part of it.

I wish to also thank my parents for tolerating me in my stressed-out writing state these past few months, despite not being able to leave this apartment, and for taking an interest in this project which is so outside of their areas of interest.

Finally, I want to express my gratitude to my housemates at Vassar and to my Beacon family for their support and for withstanding my thesis-related stress for the past eight months, I love you all.

Chapter 1

Introduction

Many researchers in the area of Natural Language Processing (NLP) examine historical corpora to understand change through time, both linguistically and socially. Techniques like unsupervised topic modeling and vector space modeling provide different interpretations of the nature of change when given a historical corpus.

The Proceedings of Old Bailey (Hitchcock et al., 2012b) is a particularly interesting corpus because of the population it represents. Unlike many historical records which focus on the elite, or at least the part of the population with jobs in areas thought important enough to collect data on, the Proceedings depict the lives of a wide range of the London population with an emphasis on the lower classes. Thus the text provides a window into the lives of a unique subset of the London population that present a different set of experiences related to social and linguistic change. There is work to be done to understand not only how linguistic change is reflected in the corpus but also thematic change. This project will contribute to the understanding of this historical corpus and the capacity of computational tools to expand this understanding. Additionally, this work will help determine computational tools that can be applied to other historical corpora with similarly interesting sociolinguistic implications. One example of an additional historical corpus that can provide valuable information into our history is Documenting the American South (DocSouth), a collection of text, images, and more historical documents relating to history in the Southern states. This corpus can be examined using the tools developed in this research to inform ourselves further about African American history, among other things.

This paper first describes other work related to this corpus and work using similar methods in Chapter 2. Then, Chapter 3 explores the data in the Proceedings of Old Bailey. Chapter 4 introduces topic modeling and shows results from some models. Next, Chapter 5 presents findings from two vector-space models run over the Proceedings. In Chapter 6, results are reported for methods combining topic modeling and vector-space models. Finally, Chapter 7 explores the shortcomings of this project, conclusions, and the possibilities for future work.

Chapter 2

Related work

This section describes research with areas similar to this project including work that applies similar methods, namely topic modeling, to the Proceedings of Old Bailey and projects using topic modeling techniques on other historical corpora.

2.1 The Proceedings of Old Bailey

Several projects have been done investigating the Proceedings of Old Bailey (Hitchcock et al., 2012b). The work done by Klingenstein et al. (2014) aims to explore two distinct "genres" of trials in the Proceedings of Old Bailey: the non-violent trials and the violent trials. This work focuses primarily on how these two types of trials differ semantically and the ways in which they emerged as distinct using a subset of the whole corpus. This concentration on the semantic differences motivates the methods used in the work and the reason why topic modeling was not used.

The work of Degaetano-Ortlieb (2018) focuses on the lexical and stylistic differences over time between witnesses and victims of different genders and socioeconomic backgrounds present in the Proceedings of Old Bailey. The data is not explicitly annotated with each speaker's economic class, so Degaetano-Ortlieb (2018) determines if a speaker should be considered "higher class" or "lower class" based on the speaker's transcribed testimonies. Like this project, the research by Degaetano-Ortlieb (2018) focuses on change over time by examining the language of interest over the course of the transcribed range of years. The methods in the work of Degaetano-Ortlieb (2018) differ from those of this project, namely in the use of unsupervised topic models. Instead of using topic modeling, Degaetano-Ortlieb (2018) measures the lexical and stylistic divergence between each category of speaker with relative entropy.

2.2 Topic Modeling Historical Corpora

There has also been work with legal corpora other than the Old Bailey that use topic modeling as a tool for exploring and reasoning about the data. One of the first examples is the research of Carter et al. (2016) in examining the subject matter and decisions of the High Court of Australia. As an early example of using topic modeling to examine legal proceedings, it presents data to justify the benefit of applying these methods to corpora of this type.

In their research, Jockers and Minmo (2012) focus on using latent Dirichlet allocation (LDA) (Blei et al., 2003) to identify topics from a corpus of 19th-century fiction from a corpus of 19th-century British, Irish, and American fiction (Jockers and Minmo, 2012). This work does not consider the element of evolution over time, but does present an interesting evaluation of the effect of an author's gender on the types of topics present in the text.

Mimno (2012) uses unsupervised topic modeling to examine text from 24 journals on JSTOR in the area of classical philology and archaeology. In particular, he explores the text by decade. He runs LDA over the corpus and studies the number of words per year published in each topic to show how that topic evolved over time. This method of using non-dynamic LDA to examine text dynamically is something explored in this project as well. The work of Mimno (2012) places more focus, however, on exploring topics in scholarship in languages other than English. In his work text mining the diary of an 18th century midwife named Martha Ballard, Blevins (2010) also uses LDA to examine temporal trends. He looks at how present a topic is over each month in a year as well as over the course of the diary as a whole.

Momeni et al. (2018) propose methods for dynamic topic modeling and use two datasets including a collection of legal texts to evaluate these methods. In this work, they aim to consider various behaviors of topics in corpora that span many years, such as growing, splitting, merging, and dying. The methods provide a way to track the evolution of a particular topic over the course of the data, particularly interesting when the topic centers around a social or political issue, such as race, that grows or shrinks in importance during a given time period. The emphasis in this paper was less on the particular datasets used, but on the actual methods developed.

Chapter 3

Data

This chapter first presents some background and historical information on the corpus of interest, the Proceedings of Old Bailey, in Section 3.1. Then, in Section 3.2, the steps to process the text are described along with details about the idiosyncrasies of the text. Then, in Section 3.3, relevant statistics about the corpus are presented. Section 3.4 introduces a corpus used to supplement the Proceedings text data. Finally, Section 3.5 shows some word frequency data for the corpus.

3.1 The Proceedings of Old Bailey

The Proceedings of the Old Bailey, 1674–1913, is an online collection of transcribed criminal trials held at London's central criminal court. The collection contains the text from 197,745 criminal trials. The transcribed trials range in length from 8 words to more than 150,000 words. The criminal trials span 2,163 different sessions in court. The corpus also includes 475 ordinary's accounts, all of which are augmented with annotations indicating names, offenses, verdicts, and sentences relevant to the given trial.

The ordinary's accounts comprise a collection of documents published between 1676 and 1772 which include biographical information about prisoners executed in Tyburn, the location at which most London prisoners were executed, along with the prisoners' last words and actions before execution. Over the course of the nearly 100 years spanned by these accounts, the focus shifted from confessions to the actual crimes and trials committed. All accounts were written by the Ordinary of Newgate, who provided spiritual guidance to prisoners with death sentences at the Newgate prison. In these 475 accounts, one for nearly each day in which prisoners were executed at Tyburn, about 2,500 prisoners are depicted. Within the collection of accounts, there were at least three different Ordinaries of Newgate. Publishing these accounts provided a source of income for these ordinaries, a lesson to readers about the consequences of crime, and an opportunity for convicts to explain their actions. The accounts were subject to some doubts about the morality and eventually stopped being published because of diminished demand (Emsley et al.).

3.2 Pre-processing

There are several idiosyncrasies of this corpus that are not necessarily present in corpora of other types. For example, due to the nature of the context (i.e., transcriptions of individual trial proceedings), the corpus has an abundance of proper nouns. Each different session in court includes

new names for each witness, defendant, and juror along with other names and locations relevant to the testimony. To see an example of the text included before each session, see Figure 2 in the appendix. Also included in the appendix is Figure 1, an example front page to a session in court, and Figure 3, an example trial transcript. In comparison with a corpus of literature, which may have a handful of character names and places that appear frequently in a given document, this corpus contains many proper nouns in each document that appear just a few times. This would suggest that proper nouns are a more significant portion of the text than in many other corpora, and that there is inherently more variation. To provide information about how the number of proper nouns in Old Bailey compare to a more standard corpus, the Berkeley Neural Parser (benepar) (Kitaev and Klein, 2018) was used to parse both the Proceedings of Old Bailey and the British National Corpus (BNC) (BNC Consortium, 2007). The BNC was chosen to compare with the Old Bailey corpus because it contains a large number of words in British English not related to any particular topic or area of study. It provides a baseline for the composition of a more typical corpus than the Old Bailey which has such a narrow focus. Once both datasets were parsed, the number of words identified as proper nouns were counted in each. Table 3.1 shows some statistics about the presence of proper nouns in each corpus.

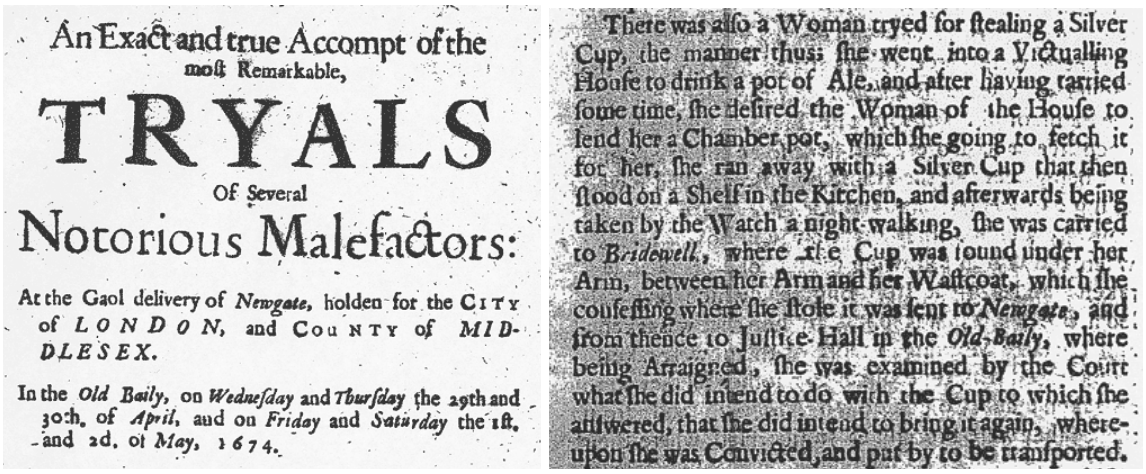
	BNC	**Old Bailey**
% Nouns in corpus	24%	19%
% Proper nouns out of nouns	2%	3%
% Proper nouns in corpus	0.5%	0.5%

Table 3.1: Percentage of proper nouns

Contrary to the intuition that the Proceedings of Old Bailey would have significantly more proper nouns than the BNC, Table 3.1 shows little difference between the presence of proper nouns in these two corpora. However, the large variation in capitalization and spelling is a particularly widespread issue in the corpus, especially in the earlier documents. Figure 3.1 shows an except from a trial in one of the earliest sessions in 1674. Words like "*woman*", "*shelf*", and "*convicted*" are all capitalized in the middle of sentences. Additionally, Figure 3.1 shows the front page from that same session in 1674. In this page alone, the text includes alternative spellings "*trial*" and even "*Old Bailey*". The text also uses the word "*accompt*", an archaic synonym for "*account*". These irregularities in capitalization make the benepar parser less reliable on the Proceedings than it is for the BNC. Therefore, the results in Table 3.1 may not be accurate.

While some variation in spelling is likely the result of less strict spelling standards and simple errors in the original document, there is a more consistent spelling issue in the corpus that needed to be fixed. The way in which the original documents were converted to XML created an issue with merging words. Although all the text in the Proceedings were transcribed manually at least once, OCR was also employed in the digitization of documents (Hitchcock et al.). There are many instances where words are merged together unnecessarily, sometimes because the transcribed data does not properly recognize line breaks. For example, an Ordinary's account from October 25, 1676 transcribed words like "*deserveddeath*" and "*aMinister*" by merging the last word on a line with the first word of the next line.

In order to address this issue, two lists were constructed: a list of bigram frequencies and a list of unigram frequencies. Both lists were compiled using the trial transcripts and Ordinary's accounts dated between 1674 and October 1834, the subset of the corpus that was manually typed



(a) Front page of session (b) Excerpt from trial transcript

Figure 3.1: Data from session on April 29, 1674

twice and then merged automatically. This subset was chosen to minimize the number of errors in digitization present in the list of bigrams. The Python package PyEnchant was used to split merged words. Each word in the data was determined to be a valid word or not. Words are considered valid if they are present in the MySpell dictionary of British English words or are capitalized (assumed to indicate a proper noun). If the word was a valid word, it was not changed. If the word was not valid, PyEnchant suggests possible corrections to the word using the same British English dictionary augmented with the list of unigrams. Including the unigrams in the dictionary allows PyEnchant to suggest word splits containing words unique to the corpus, either because the word is a proper noun or is spelled differently from a modern day spelling. All suggestions that indicate a possible split of the original word were considered possible corrections to the split word. The original merged word was then replaced in the text with the option that had the highest frequency according to the compiled list of bigram frequencies.

However, this approach to the issue of merged words does not successfully address all necessary splits. Since all capitalized words are ignored, any proper nouns merged with the following word will remain unchanged. Additionally, the inconsistent capitalization in the text means that many words, not just proper nouns, are capitalized. If any of these words are merged, they will not be fixed in the processed text.

Note that the original XML data also contained a handful of words that were split into two words. For example the word "*prisoner*" turned into "*pri soner*" in the OCR process. There are several other examples of this issue, however, correcting the problem frequently caused the unnecessary merging of words. These split words were left as is.

3.3 Relevant Statistics

In general, the quantity of data transcribed about a trial increased over the course of the 239 years, resulting in some more detailed data from the late 18^{th} century, 19^{th} century, and early 20^{th} century. To gain an understanding of the language used in the Proceedings, some descriptive statistics have

been calculated. Table 3.2 shows some details about the corpus at different phases of processing in addition to the same statistics for the entire BNC for comparison.

The corpus as a whole was considered when calculating these statistics and was also divided into three subsets based on date. The first includes all transcripts from the years 1674 to 1773, the second includes those from 1774 to 1873, and the last includes the remaining transcripts (1874 to 1913). This particular division of the data was chosen to allow for a relatively even split while taking into account the evolving characteristics of the data over time. These differences are described in Table 3.2. For each of these subsets, the number of documents, trials, tokens, and types (i.e., unique tokens) were calculated to show how the corpus changed throughout processing. A document is defined as the proceedings (or ordinary's accounts) for one day. Within each day, this could include from 1 to 482 trials. Note that the number of trials includes all the text that appears in each document before the trial starts. This often consists of an introduction of the Proceedings of Old Bailey (e.g., Figure 1), a list of jurors (e.g., Figure 2), and similar information that applies to each trial in the document.

	Start Year	Docs	Num. Trials	Num. Tokens	Num. Types
Original	**1674**	1192	37 773	15 502 024	250 877
	1774	972	131 564	80 457 401	958 208
	1874	474	34 232	32 368 114	609 837
	Total	2638	203 569	128 327 539	1 818 922
Tokenized	**1674**	1192	37 773	6 897 523	124 181
	1774	972	131 564	35 852 853	299 332
	1874	474	34 232	15 007 878	146 705
	Total	2638	203 569	57 758 254	570 218
Spell-checked	**1674**	1192	37 773	6 907 029	123 726
	1774	972	131 564	35 905 765	296 662
	1874	474	34 232	15 026 426	145 523
	Total	2638	203 569	57 839 220	565 911
BNC	**Total**	4049	*N/A*	97 087 701	1 530 235

Table 3.2: Basic statistics

There are three phases of processing reflected in Table 3.2. First, values are reported for the original data. This refers to text data converted from the XML documents that comprise the corpus. No other modifications are made to the data at this point. The second phase of processing is the tokenized data. This data has been separated into words by NLTK's word tokenizer and has been further modified to get rid of words from NLTK's list of stop words (Bird et al., 2009) and any unwanted symbols in or around words. Once tokenized, all tokens shorter than two characters are removed along with any token that does not contain at least one letter (i.e., all punctuation and numbers are removed). Additionally, contractions split by the tokenizer are replaced by the corresponding words (i.e., "*don't*" is tokenized to "*do n't*" and then replaced with "*do not*"). The table shows that the process of tokenizing greatly reduces the number of tokens and types in the corpus. The final stage of processing the data is spell-checking. This stage involves processing the text as described in Section 3.2. The number of tokens increases after spell-checking because it involves turning single words into two words. However, the number of types decreases because

many of the merged words appear rarely in the corpus. When split, they become words frequently recognized by the corpus and thus don't contribute to the count of unique words.

The statistics for the BNC can help put this information about the Old Bailey Corpus into perspective. The text from the BNC was not processed the same way the Proceedings were. Instead, the content of each file in the BNC was split by whitespace to calculate the number of tokens and types. This means it is at the same stage as the files in the row labeled "*Original*". While the BNC has about 1.5 times as many documents as the Proceedings of Old Bailey, it has far fewer tokens. However, while the Old Bailey has more than 100 times fewer types than tokens after processing, the BNC has about 65 times fewer types than tokens. This difference is unsurprising given the repetitive nature of the Old Bailey corpus. Since the BNC does not have such a narrow area of focus, it has a higher percentage of unique words.

It is evident from Table 3.2 that this particular split of years is not ideal. Clearly, there are significantly different numbers of documents in each subset of the corpus. The first subset has the most documents primarily because all of the ordinary's accounts were published within this time period. The final section of documents spans a shorter period of time than the first two, 39 years versus the typical 100. More revealing than the number of documents, however are the numbers of tokens and types in the subsets of the corpus. Even though the first section has the most documents, it has the fewest tokens and types by a large margin. Even with the extra ordinary's accounts, the transcripts were significantly less verbose in the earlier years. This split of the corpus was selected as a way to balance the imbalance in number of words per document over time.

3.4 London Lives

To augment the data of the Proceedings for use in topic modeling (section 4.1), the London Lives corpus (Hitchcock et al., 2012a) was also processed. The corpus combines data from various archives and datasets of text published between 1662 and 1800. The text data, like the Proceedings of Old Bailey, focuses on representing the lives of the non-elite in London. The London Lives corpus is composed of data from Old Bailey and the Ordinary's accounts, lists of names of prisoners from Newgate prison, records of several parishes (responsible for relieving poverty), records of a London guild responsible for charity, hospital records, and several other records that provide insight into plebeian life in the 18^{th} century.

Because this data overlaps with the Proceedings of Old Bailey in terms of time period and emphasis on providing data reflecting the lives of lower class Londoners, it can add useful information about the time period in general. Comparing this data with Old Bailey data can reveal what information is particular to Old Bailey and what is characteristic of the time period. The London Lives corpus will also be used to supplement the data used when running topic models to avoid sparsity. While the corpus is slightly adjacent to the Proceedings Old Bailey in that it does not contain only trial transcripts, it serves the same purpose as the collection of Ordinary's Accounts, which are closer to essays centering on sin and religion than to trial transcripts.

Table 3.3 shows some statistics about the London Lives corpus after going through the same processing as the Old Bailey corpus described in section 3.2. Because the corpus will be integrated with the Old Bailey corpus, it is split up into similar subsets as discussed in 3.3. The London Lives corpus includes documents starting in 1662, so the first subset of 100 years includes documents published between 1662 and 1762. There are only two time slices in this corpus since it does not include documents published after 1800. It is clearly a smaller corpus than the Proceedings given

Start Year	Docs	Num. Tokens	Num. Types
1662	884	8 192 745	213 847
1762	740	12 218 962	245 094
Total	1624	20 411 707	458 941

Table 3.3: Basic statistics for London Lives

that it has around 20 million tokens in comparison to the 57 million in the Old Bailey. Just like the Old Bailey, however, the data published in the later years includes many more words than the earlier documents. This may be because the documents themselves are more verbose, more data is available for the later years, or both. Even though the second time slice spans 38 years, less than half the time included in the first time slice, it contains nearly as many documents indicating that there is much more data available for the later years. Additionally, this second time slice contains more than 1.5 times as many tokens as the first which suggests that the data itself is significantly more verbose.

3.5 Word Frequencies

To further understand the composition of this corpus, word frequencies were calculated for the BNC, the Old Bailey corpus as a whole, and the same three temporal subsets discussed in section 3.3. Figure 3.2 shows word frequencies for the BNC for the purpose of comparing this unique corpus with a more standard one. In Figure 3.2, the top 30 words in the BNC corpus are along the x-axis and their frequencies within the corpus are plotted along the y-axis. The shape of the curve reflects the properties stated by Zipf's Law, namely that a frequency plot like Figure 3.2 will decrease roughly exponentially. There is a very steep decline from the most common words to the others within the top 30. Additionally, it is important to note that none of the top 30 words is one that adds much meaning to the corpus.

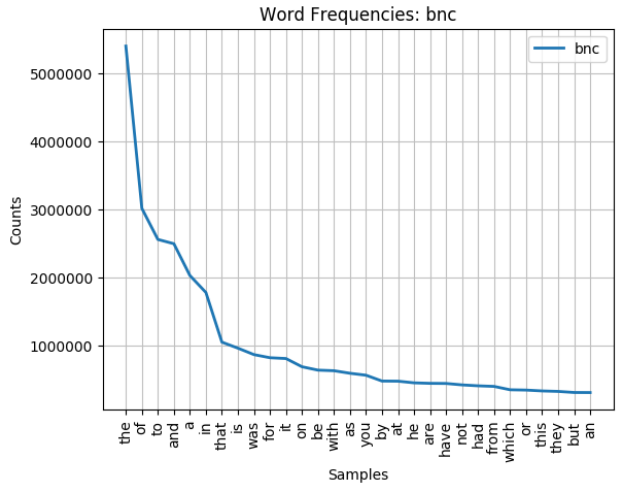
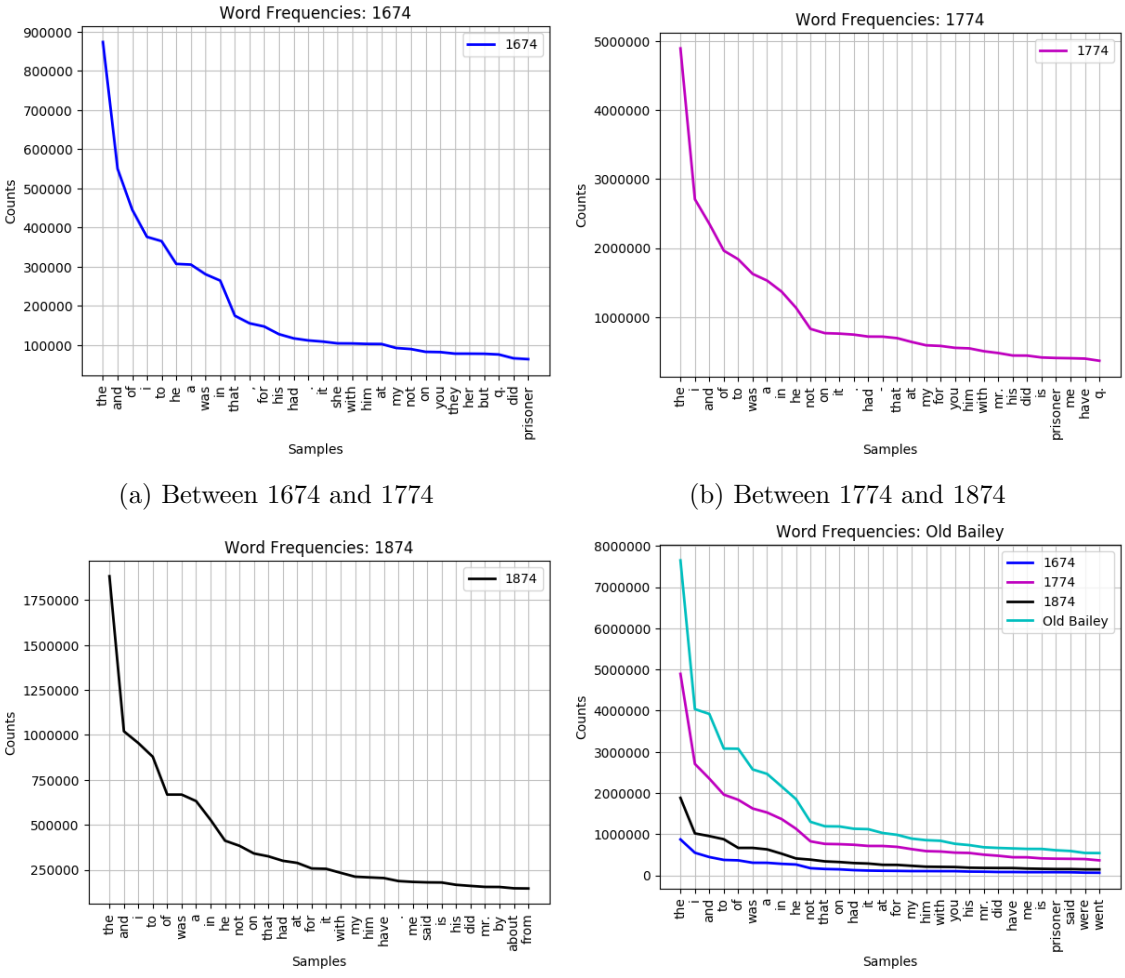


Figure 3.2: Word frequencies over British National Corpus

Figure 3.3 shows four different word frequency graphs. First, Figure 3.3 (a) plots the counts of the top 30 words in the subset of the corpus spanning 1674-1774. Figures 3.3 (b) and 3.3 (c) show the same measure for subsequent subsets of the corpus. Finally, Figure 3.3 (d) shows the word frequency counts of the top 30 words in the corpus as a whole. The same plots from the previous three figures are included to show how they compare to the corpus' measure. Interestingly, there are words in the set of 30 most frequent words in each of the subsets that are not stopwords. The content word "*prisoner*" appears in this set for the 1674 subset, the 1774 subset, and the corpus as a whole. It makes sense that this word would appear so frequently given the context of the corpus. Additionally, the word "*q.*", indicating a question, is included in the set for the first two subsets of the corpus. This is also understandable when considering the Old Bailey corpus which is filled with questions being asked during trials. The last content word is "*mr.*" which appears in the top 30 most frequent words in three of the four plots shown in Figure 3.3. This ties back to the prevalence of proper nouns in this particular corpus, as discussed in section 3.2.



(a) Between 1674 and 1774

(b) Between 1774 and 1874

(c) Between 1874 and 1913

(d) Whole corpus

Figure 3.3: Word frequencies over Old Bailey Corpus

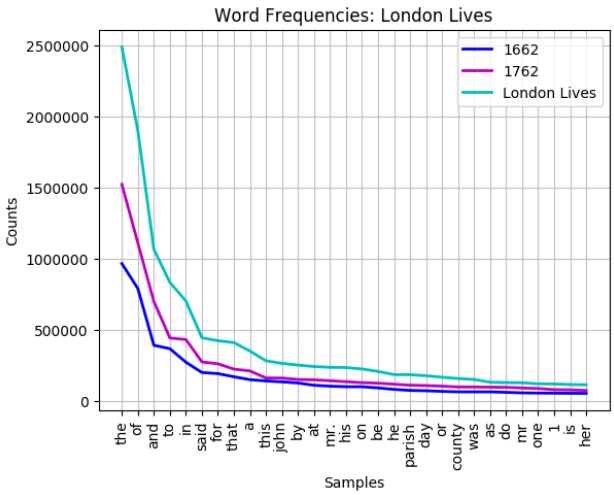


Figure 3.4: Word frequencies over London Lives corpus

The same calculation of word frequency was also computed for the London Lives corpus to gain a similar understanding of the data that will be used to augment the Proceedings of Old Bailey. Figure 3.4 shows the result of this computation. The 30 most frequent words in the London Lives corpus include even more content words than the top 30 words in the Old Bailey corpus. Notably, these content words include the name "*John*" which makes sense because of the corpus' focus on representing lives of Londoners and thus containing many names.

Chapter 4

Topic Modeling

This chapter presents three main topic modeling techniques. First, Section 4.1 introduces latent Dirichlet allocation (LDA) and explores the results of running LDA over various forms of the corpus. Next, Section 4.2 presents the results of running Dynamic Topic Modeling (DTM), a method that takes time into account, over a few time slices of the corpus. Finally, Section 4.3 demonstrates a manual alternative to DTM and explores the results.

4.1 Latent Dirichlet Allocation

Beyond corpus statistics, the corpus itself can be understood better through the use of unsupervised topic modeling. Latent Dirichlet allocation (LDA) (Blei et al., 2003) in particular can be useful in revealing latent themes and topics in the text. In other words, LDA identifies themes that are not explicitly stated in the text, but are instead implied and reflect the content and language in the text. Each LDA model in this project was run using the MALLET (McCallum, 2002) implementation.[1] Unless otherwise stated, every model described in this section defines its documents as defined in Section 3.3 (i.e., one document consists of all trials from a given day).

4.1.1 Topic Coherence

When determining which models were the best performing, a measure of topic coherence was calculated for each topic found by a given model. Then, the average topic coherence was computed and used to choose the best model. The method used to calculate topic coherence uses Word2Vec (Mikolov et al., 2013) similarities. To determine coherence for a specific model, the vocabulary used to train that model was first used to train a Word2Vec model. This ensures that the LDA model and the Word2Vec model have identical vocabularies. For each topic found by the LDA model, the average topic coherence was calculated by iterating over each pair of words in the topic, summing up the value of the Word2Vec similarities between the pair, and dividing the sum by the total number of pairs. The LDA model with the highest average topic coherence over all the topics identified by the model was determined to be the best-performing model.

This method of calculating topic coherence was chosen to avoid needing a different reference corpus. Other methods like those described in Lau et al. (2014) and Aletras and Stevenson (2013) use reference corpora of New York Times data or Wikipedia data to determine how coherent a topic

[1] All models were run using a wrapper for MALLET implemented by Gordon et al. (2016).

is. However, given the subject matter of the Proceedings of Old Bailey, using a reference corpus with modern American news data or Wikipedia articles would not be a good measure by which to judge topic coherence. Instead, this method uses the word similarities in the Old Bailey corpus itself to measure topic coherence which helpfully utilizes the actual relationships in the text.

4.1.2 Unigrams

First, LDA was run over the corpus as a whole to get a general understanding of the topics present over the entirety of the corpus. Table 4.1 shows some hand-selected topics found by running LDA with 50 topics and 1000 iterations. This model had the best average topic coherence in comparison to models with 30, 75, and 110 topics.

Topic ID	Top 6 Words
0	deceased, prisoner, wound, man, body, murder
1	examined, house, mrs, paid, money, letter
2	prisoner, o'clock, found, aged, street, house
...	...
7	prisoner, examined, cross, gave, station, road
...	...
9	found, guilty, summary, indicted, prisoner, goods
10	prisoner, goods, indicted, house, john, stealing
11	paid, account, examined, book, goods, money
...	...
21	house, room, door, found, street, time
22	god, life, years, death, lord, great, sins
...	...
47	november, october, december, september, shop, mrs
48	captain, ship, board, boat, time, vessel
...	...

Table 4.1: Results of LDA run over entire corpus

Much of the sparse data includes repetitive words and information that does not contribute very meaningfully to the topic model, namely details about date at the start of each trial and within the testimonies, like those in topic 47 in Table 4.1. Additionally, as Table 4.1 shows, several common words appear heavily in several topics. Just in the few topics displayed in the table, the word "*prisoner*" is the strongest word in three of the topics. Additionally, out of the 50 topics found, "*prisoner*" appears in the top 20 words of 30 of them and is the strongest word in 19.

Some of the topics in Table 4.1 are promising. Topics like 0, 9 11, 22, and 48 appear to be coherent and more meaningful than many others. However, the repetitive nature of the topics shown in Table 4.1 is also reflected in the topics excluded from the list. Many of the topics not shown are nearly identical to those in the table or have no clear theme connecting the words that comprise it.

4.1.3 Bigrams

To avoid having topics that are so similar, LDA was also run over bigrams of the corpus. This provides more data to the model and prevents some very common words from being present in every topic. Table 4.2 reports some hand-chosen topics found by LDA when run over the corpus bigrams with 50 topics. This model had the best average topic coherence in comparison to models with 30, 75, 110, and 125 topics.

Topic ID	Top 6 Bigrams
0	verdict guilty, police constable, prisoner oath, sentence months, police station, months hard
1	guilty aged, indicted stealing, cross examined, aged confined, confined months, police constable
...	...
7	mrs emsley, barnsley street, mrs haggerstone, mrs blackburn, mrs wheeler, miss harvey
...	...
18	stealing october, stealing september, st october, october clock, october prisoner, september prisoner
...	...
20	justice fielding, cross examination, spinster indicted, general character, produced court, mrs rudd
...	...
40	montagu williams, williams prosecuted, poland montagu, lloyd prosecuted, messrs lloyd, examined geoghegan
...	...
42	coal heavers, riotously tumultuously, begin demolish, demolish pull, unlawfully riotously, pull dwelling
...	...

Table 4.2: Results of LDA run over corpus bigrams

Clearly, many of the terms that dominate the topics in Table 4.1 do not overwhelm the results of a model using bigram data. While "*prisoner*" is still present in the topics shown by Table 4.2—as it should be—it is to a much lower degree and always put into more context, thus the topics show more variety from one another. Topics like 7 and 18 are not particularly exciting, but they indicate the model's ability to identify the common theme of proper nouns and months. Some other topics, on the other hand, are somewhat interesting. Topic 40 in particular identifies the connection between Montagu Williams, a prosecutor at the Old Bailey in the 1800s, and prosecution.

Many of the topics not included in Table 4.2 are again very similar to those in the table or have no obvious common thread to justify its designation as a topic.

4.1.4 Combining Old Bailey with London Lives Data

To continue exploring bigram data, the Old Bailey Corpus was combined with the London Lives corpus described in section 3.4. With this combined corpus, LDA was run over the bigrams of the data. This model found 100 topics in the data instead to adjust for a larger amount of varied data.

Table 4.3 shows a sample of topics found by this model. These particular topics shown are the five topics most similar to any topic found by the model discussed in section 4.1.3. The method for selecting an alignment of this kind is discussed in detail in section 4.3.1. The topics were selected this way to identify the subset of the 125 topics found that reflect the themes most similar to those in the Old Bailey corpus.

Index	Top 6 Bigrams
0	lord mansfield, unlawfully riotously, sir william, tap room, number people, riotously tumultuously, coal heavers, demolish pull, begin demolish, pull dwelling
1	madame valentin, mrs emsley, welbeck street, mrs foote, barnsley street, richardson webster, brown ross, mrs johnson, bretten terrace, examined gill
2	unsound mind, state mind, younger prisoner, sound mind, prussic acid, mind time, attorney general, medical men, commit suicide, act committed
3	bull inn, mrs stephenson, mrs roper, mrs sammons, mrs harrisson, william gillard, mrs brown, serjeant pigott, chelsea omnibus, henry lovesay
4	mrs watt, electrical accessories, miss toovey, mrs lewis, finsbury pavement, lady violet, pitt street, mrs foster, mrs roberts, miss nix
5	stamp office, aces spades, mrs lloyd, toll house, ace spades, ralph holden, mrs mattingley, packs cards, card maker, madame denis
...	...

Table 4.3: Results of LDA run over Old Bailey and London Lives bigrams

The majority of the topics shown in Table 4.3 and many of the remaining topics are predominantly composed of names. This is certainly an increase over the percentage of topics containing mostly names in the model described in section 4.1.3, which does not include the London Lives data. This might suggest that the extra data does not add valuable information to the corpus when used in the context of topic modeling.

4.2 Dynamic Topic Modeling (DTM)

Since LDA does not explicitly consider the temporal differences between the documents in the Proceedings, the corpus was split up into the same three sub-corpora as was done in section 3.3, each containing documents spanning approximately 100 years. This allows for a comparison between topics found in documents in one time period to documents in the others, potentially revealing key differences.

Blei and Lafferty (2006) developed an implementation of dynamic topic modeling that finds topics and identifies how they change over time. This model was run over the Proceedings of Old Bailey using the same subsets of the corpus analyzed in section 3.3 using Gensim's DtmModel wrapper (Řehůřek and Sojka, 2010) and identifying 30 topics. Some hand-selected topics are shown in Table 4.4. Each topic ID in the first column corresponds to three topics, one from each time slice.

The topics in Table 4.4 have many of the same issues present in Table 4.1. Primarily, the repetitive nature of the topics make these results generally uninformative. Again, the word "*prisoner*" appears in all three time slices of eight of the topics found by this model. Additionally, the remaining topics use mostly common words that don't reveal much about the text itself. The topics

ID	Year	Topic
0	**1674**	value, indicted, o'clock, handkerchief, live, john
	1774	value, indicted, said, o'clock, aged, live
	1874	said, september, asked, august, live, street
1	**1674**	one, two, three, half, got, could
	1774	one, two, three, got, half, could
	1874	one, two, three, got, could, came
2	**1674**	prisoner, deceased, head, hand, court, side
	1774	prisoner, head, hand, court, side, knife
	1874	prisoner, head, knife, court, drink, hand
3	**1674**	sept, aug, think, child, time, would
	1774	think, child, would, saw, time, say
	1874	child, think, would, examined, saw, anything
...	...	...
10	**1674**	one, captain, sir, ship, time, would
	1774	one, recollect, sir, would, captain, time
	1874	one, case, think, sir, would, made
11	**1674**	prisoner, went, came, took, told, davis
	1774	prisoner, went, came, took, told, davis
	1874	prisoner, went, came, took, told, could
...	...	...

Table 4.4: Dynamic Topic Modeling (Blei and Lafferty, 2006)

include few words that refer to the context of this corpus and instead favor relatively common and consistent words like months of the year.

These results suggest that DTM (Blei and Lafferty, 2006) is not a very effective way to reveal temporal differences in the Proceedings of Old Bailey. Note that while it might be beneficial to run this same DTM model over the corpus bigrams, the Gensim implementation is prohibitively time-consuming on a corpus this size given the parameters.

4.3 Manual Dynamic LDA

While the topics found through DTM weren't very revealing, there is still value in a version of topic modeling that considers the temporal aspect of a corpus. As an alternative to DTM, LDA can be run over manually separated sections of the corpus. This section demonstrates this method for identifying temporal change. Additionally, to address the issue of frequent words like "*prisoner*" dominating the topics, the data was modified so that the model was only run on the words that were not too common and not too rare. This means words that appeared in more than 80% of all documents in a given time slice were excluded along with words that appeared in less than n documents where n is 5 for a unigram model and 10 for a bigram model.

4.3.1 Topic Alignment Method

When running LDA like this, it is useful to have a method of determining which topics from each time slice correspond to topics from the others. The method used in this project to automatically align topics from the different models is to calculate the difference between each topic and consider the most similar topics as an alignment. Finding the difference between two topics from two different models is shown in Algorithm 1.

Algorithm 1 Topic Differences

```
for t_1 in T_1 do
    for t_2 in T_2 do
        diff[t_1, t_2] = 0
        for w in vocab do
            diff[t_1, t_2] += abs(weight of w in t_1 − weight of w in t_2)
```

Since it makes sense to split the Proceedings of Old Bailey into three different sub-corpora based on time, thus creating three topic models to align, this difference is calculated for each pair of topics in each pair of models. Then, the topic differences are ranked from least different to most different. Some topic alignments will include topics from all three models. For example, consider three models run over each subset of the corpus, m_1, m_2, and m_3 with lists of topics T_1, T_2, and T_3 where $T_n(i)$ refers to the i^{th} topic of model m_n. If $T_1(x)$ is least different from $T_2(y)$, $T_2(y)$ is least different from $T_3(z)$, and $T_3(z)$ is least different from $T_1(x)$, this creates an alignment for all three models between $T_1(x)$, $T_2(y)$ $T_3(z)$. If this relationship does not exist for some $T_n(i)$, it will not be included in an alignment of three topics, and instead be aligned with whichever topic from the remaining two models it is least different from.

After creating all alignments, each one is given an index. The lower indices indicate that the alignment has the lowest difference sum. Higher indices mean that the alignment includes more variation and may suggest a set of topics that evolve over time.

Since this alignment method allows for some alignments to only span two topics, any remaining topics that are not included in any of the alignments are determined to be unique to their time slice. These will have the highest indices, indicating that they are the most different alignments.

4.3.2 Unigrams

Table 4.5 shows a selection of the 15 alignments found from topics found when running LDA over the three time slices of the Proceedings of Old Bailey discussed in section 3.3. Since there were three individual LDA models, average topic coherence was calculated for each model and then averaged together to get the overall average topic coherence for each number of topics. In comparison to 10, 20, and 30 topics for each time slice, the highest overall average topic coherence was found in the model with 15 topics per time slice.

Out of the 15 alignments found, only 5 of them included the first time slice, meaning there were 10 topics determined to be unique to the first 100 years of the Proceedings. The remaining alignments are similar to the ones hand-selected to appear in Table 4.5. Some of the examples show little variation between topics, typically in contexts that haven't varied much in language. For example, the second alignment shows that the topics are not meaningfully distinct based on time slices, however that is understandable when considering the content (nautical language) and

Index	Year	Topic
0	**1674**	*N/A*
	1774	burnt, hayes, smoke, pearce, barry, bricks, turpentine, burning
	1874	insurance, paraffin, policy, smoke, oil, burning, cellar, fires
1	**1674**	*N/A*
	1774	palmer, stomach, symptoms, poison, arsenic, disease, medicine, poole
	1874	stomach, disease, bartlett, medicine, poison, patient, symptoms, brain
2	**1674**	*N/A*
	1774	boat, vessel, deck, cabin, barge, steam, smell, crew
	1874	captain, deck, mate, cabin, boat, vessel, port, collision
...	...	...
7	**1674**	mob, green, allen, heard, ramsey, tyrrell, clarke, murphy
	1774	meeting, committee, pistol, brunt, thistlewood, caspar, mullins, election
	1874	music, ross, flour, machine, walsh, paget, osborn, association
...	...	...
12	**1674**	deceased, prisoner, murder, wound, heard, mrs, room, asked
	1774	sovereigns, sovereign, policeman, trousers, pairs, clarkson, adolphus, cheque
	1874	bonds, mortgage, trustee, estate, ledger, loan, deed, creditors
...	...	...
20	**1674**	note, letter, prisoner, writing, bill, office, bank, wrote
	1774	*N/A*
	1874	*N/A*
...	...	...

Table 4.5: Results of LDA run over sub-corpora

the fact that all words used in the topic are still used in modern English. It is important to note that this topic roughly corresponds to topic 10 in Table 4.4 which also includes some nautical terms. However, the topics composing this alignment are much more coherent and contain almost exclusively nautical-themed words while topic 10 in Table 4.4 contains many other words. This is certainly an improvement from the performance of DTM over the Old Bailey.

The issue of repetitiveness discussed in section 4.1 is not a big issue with this model. Because of how the data was modified to exclude words that were too common or too rare, words like "*prisoner*" do not dominate the topics. However, this does mean that alignments like the 20$^{\text{th}}$ are created. Since this alignment does not contain topics for the second two time slices, this implies that the topic from the first time slice is unique to that period of 100 years. However, these words are not unique, and in fact are likely too common in the other time slices to be included. Thus this assumption does not hold.

It's important to note that Table 4.5 demonstrates how the alignment method can succeed and how it can fail. The first three alignments, for example, appear to be very logical and well-matched. Some of them can even be assigned descriptive titles like "*words relating to poison*" for alignment 1 and "*nautical terms*" for alignment 2, as mentioned before. Other alignments, however, do not show much similarity between the aligned topics. Alignment 7, for example, does not appear to be very coherent.

4.3.3 Bigrams

Just as bigrams were incorporated to improve the output of LDA over the entire corpus in section 4.1.3, the same method was employed to try to get more interesting results in the manual dynamic LDA run over the three different time slices by reducing the repetitive nature of the topics. Table 4.6 shows a selection of topic alignments ranked by similarity for some of the 55 topics found for each time slice. This model performed best in terms of average topic coherence in comparison to models with 35 and 45 topics per time slice.

Table 4.6 shows an interesting change brought by the use of bigrams. Of the 55 topic alignments found between at least two topics, there are several alignments that appear to be very similar to the alignment with index 0, meaning alignments that feature mostly proper nouns. Additionally, many of the other alignments, such as the alignment with index 4, consist of topics containing proper nouns despite more prominent themes. The two topics in this alignment seem to explore post-mortem examinations and other medical examinations, yet the topic from the third time slice includes several proper nouns. While the actual proper nouns that compose these topics vary somewhat between alignment, the fact that proper nouns became so much more present when using bigrams despite pruning the data to remove words that are too rare or too common is interesting. This is likely because names and streets, for example, have more consistent bigrams than other words since they are frequently composed of two unigrams in the form "*firstname_lastname*" or "*title_lastname*". The proper nouns that appear in these topics are most likely common names and locations mentioned across many sessions, but not with enough frequency to get removed from the data input to Mallet. This pattern motivated the decision to exclude words appearing in less than 10 documents for bigrams as opposed to 5 documents as was done for the unigram model.

Some of the alignments show some promise, such as alignment 10. From a human perspective, this appears to be an alignment consisting of two very similar topics. The alignment can be classified as "*fire-related words*", likely the result of trials involving arson. Additionally, alignment 31, for the most part, seems to consist of money and bank-related words. The fact that these two alignments are given indices as low as they are reveals a flaw in the alignment method. Considering how almost all the words in the alignment 10 contain the word "*fire*", it might be wise to instead allow the alignment method to consider unigrams, even when the models being compared use bigram data. In that case, alignments like this one with a strong theme throughout would be given a lower index despite having only a few exact bigrams in common. Another option would be to implement a method similar to the method for topic coherence described in section 4.1.1. Instead of just measuring the difference in weights of words in different topics, somehow incorporating Word2Vec similarity may allow for a less strict method of aligning topics.

Next, LDA was run with the combined London Lives and Old Bailey data. This model also included bigrams and was run over the three subsets of the corpus. Table 4.7 shows hand-chosen topics from this model which found 60 topics. In comparison to models with 50 and 70 topics, this one had the highest topic coherence.

Table 4.7 reveals something interesting about the alignment method. Many of these alignments are composed of topics that are almost completely different from each other. This is likely because the topics are so distinct that during the alignment process, out of about 432,000 comparisons made between the weight of a given word in two topics, only 1,335 of the comparisons involved two non-zero weights, meaning the two topics shared the word. Note that the number of comparisons was calculated by multiplying the number of loops done in Algorithm 1 for the three models which each have 60 topics and 20 words per topic, making the vocab—a combination of the words in the

Index	Year	Topic
...	...	...
3	**1674**	ann jones, bill exchange, william smith, george taylor, stamp office, david evans
	1774	henry lee, whit monday, southampton buildings, james pratt, pay sheets, line road, time keeper
	1874	alfred day, finsbury park, trafalgar street, divorce court, divorce case, osborn osborn
4	**1674**	*N/A*
	1774	post mortem, attorney general, prussic acid, mortem examination, spinal cord, medical man
	1874	edward thompson, walter thompson, mrs powell, medical man, mortem examination, mrs chapman
5	**1674**	*N/A*
	1774	goods sold, bankruptcy court, court bankruptcy, coach house, official assignee, day book
	1874	day book, goods sold, goods bought, sold goods, stock sheets, balance sheet
6	**1674**	*N/A*
	1774	chief mate, long boat, board ship, log book, left ship, put irons
	1874	william shaw, euston square, starboard side, time collision, bill sale, chief officer
...	...	...
10	**1674**	*N/A*
	1774	house fire, fire house, fire office, time fire, set fire, alarm fire
	1874	fire brigade, setting fire, set fire, fire house, night fire, insurance company
...	...	...
31	**1674**	hand writing, coffee house, bank note, bill exchange, bank notes, intent defraud
	1774	bank england, bank notes, bank note, note note, power attorney, custom house
	1874	bank england, england notes, scotland yard, circular notes, bank notes, gloucester road
...	...	...
72	**1674**	jury found, jury acquitted, found guilty, privately stealing, indicted privately, indicted assaulting
	1774	*N/A*
	1874	*N/A*
...	...	...

Table 4.6: Results of LDA run over sub-corpora using bigrams

two topics being compared—a maximum of 40 words. Because of this aspect of the bigram model, many of these alignments are simply collections of topics and, unfortunately, not much weight should be put into the evolution of the topics over time.

Index	Year	Topic
0	**1674**	*N/A*
	1774	cash book, prisoner writing, day book, money paid, time time, money received
	1874	mrs reed, bills lading, sales book, bill lading, edward harris, westminster bridge
...	...	...
3	**1674**	*N/A*
	1774	hand writing, house fire, set fire, back room, oxford street, fire office
	1874	mrs chapman, bills lading, bill lading, mrs levy, henry brown, bank england
...	...	...
10	**1674**	aged years, bad company, years born, condemned felony, condemned criminals, give account
	1774	mutton mutton, twelve clock, cato street, cleaning repairing, riotously tumultuously, unlawfully riotously
	1874	examined marshall, jermyn street, mrs robertson, property book, george mead, newgate street
...	...	...
28	**1674**	*N/A*
	1774	conducted prosecution, guilty confined, penal servitude, william webster, police court, confined twelve
	1874	conducted prosecution, guilty confined, confined months, confined twelve, william webster, confined eighteen
...	...	...
54	**1674**	*N/A*
	1774	overseers poor, parish saint, poor parish, justices peace, churchwardens overseers, church wardens
	1874	months hard, montagu williams, messrs poland, bad half, guilty conviction, conviction felony
...	...	...

Table 4.7: Results of LDA run over sub-corpora of Old Bailey and London Lives using bigrams

Alignment 3 presents an interesting issue. This alignment appears to consist of topics very similar to the ones in alignments 10 and 31 in Table 4.6. However, the alignment in Table 4.7 seems to combine these topics and make two less coherent topics. This suggests that the addition of the London Lives corpus did not help the bigram model generate coherent topics.

Additionally, clearly the measure for ordering these alignments based on similarity is skewed since the 28^{th} alignment appears to be the most similar of the ones displayed in Table 4.7. Since alignments are sorted by the sum of the word weight differences between topics, as explained in section 4.3.1, this suggests that the topics with lower indices simply contain words with lower weights. Regardless, this reveals another flaw in the alignment method.

Chapter 5

Word Vector Models

In this chapter, two different vector-space models are presented: Word2Vec in Section 5.1 and FastText in Section 5.2. The results for each model are compared and the shortcomings of each are discussed.

5.1 Word2Vec

Word2Vec (Mikolov et al., 2013) was also explored as an alternative to LDA. Word2Vec is an unsupervised, efficient, and easy to train way to represent words as short, dense vectors. Unlike LDA's ability to explore thematic trends, Word2Vec provides insight into semantic relationships between words in the text, so it was trained over the Proceedings of Old Bailey to gain an understanding of the corpus' semantics.

5.1.1 Over Entire Proceedings

Before running Word2Vec, the words in the corpus were all converted to lowercase and were lemmatized by NLTK's WordNetLemmatizer (Bird et al., 2009). Modifying the case was done so that identical words with different capitalization did not dominate the list of a word's nearest neighbors. Additionally, lemmatizing prevents the nearest neighbors from being different forms of the same word, allowing more meaningful neighbors to be found. Gensim's Word2Vec (Řehůřek and Sojka, 2010) wrapper was then run over the corpus using 100-dimensional word vectors and then used to calculate the most similar words to a list of hand-picked words of interest. These words relate to crimes, London, or reference concepts or groups that may reveal interesting information about the time period. Table 5.1 shows all the nearest neighbors found by Word2Vec for these words of interest. The table splits the words into thematic groupings.

For the most part, Table 5.1 shows Word2Vec's capacity to show semantic similarities between words. For example, the model recognizes that "*murther*" is an alternate or old-fashioned spelling for "*murder*". Since the model was run over the entire corpus, it does not show whether "*murther*" is the nearest neighbor to "*murder*" at all points in time. For some words, there aren't prominent alternate spellings reflected in the neighbors. Instead, the neighbors tend to show some logical words used in similar contexts. The word "*innocent*", for example, has logical neighbors like "*untrue*" and "*deny*", words one would expect to see in a transcript of someone found innocent.

Word	Nearest Neighbors
Words about crimes	
burglary	felony, robbery, fellony, rape, accessary, entring, breaking, larceny, adultery
crime	adultery, guilt, outrage, theft, offence, fact, heinous, deserve, sin
larceny	felon, felony, fellony, sentenced, trespass, fraud, bigamy, acquittal misdemeanour
murder	murther, murdering, murdered, rape, kill, manslaughter, barbarous, wilful, killing
rape	murther, bigamy, wilful, accessary, sodomy, murder, inquisition, perjury, manslaughter
robbery	robbing, burglary, fact, suicide, outrage, accomplice, theft, adultery, murther
theft	crime, heinous, fraud, acknowledg, adultery, unnatural, outrage, dishonesty, horrid
Words related to guilt	
guilt	motif, unjust, presumption, conscience, crime, innocence, horrid, denial, motive
hanging	hung, pinned, sticking, underneath, undid, tying, tucked, ripped, concealed
innocent	untrue, deny, innocence, denying, ignorant, alledged, sorry, denyed, solemnly
prison	jail, goal, gaol, turnkey, confinement, gate, gatehouse, asylum, bailed
punishment	fate, pronounce, deserved, punished, deserve, folly, misery, shameful, punish
sentence	judgement, sentenced, reprieve, punishment, condemnation, pregnancy, convict, pronounced, felon
Words about people	
foreign	spanish, current, turkish, american, french, spain, consul, dutch, indian
foreigner	interpreter, englishman, english, german, frenchman, french, italian, speaks, interpreted
london	lon, dublin, edinburgh, norwich, canterbury, bristol, england, cardiff, battle
man	woman, gentleman, person, lad, fellow, men, prisoner, boy, prosecutor
woman	girl, man, female, gentlewoman, lady, prisoner, creature, gentleman, fellow

Table 5.1: Results of Word2Vec over entire corpus

5.1.2 Temporal Word2Vec

Since Word2Vec run over the whole corpus cannot reveal any insights into the temporal change in the corpus, it was then run on subsets of the corpus based on year, just as LDA was done in Chapter 4.3. The same subsets of the corpus were used along with the list of hand-chosen words to explore. Table 5.2 shows the 10 nearest neighbors in each time slice for a particularly interesting selection of the words of interest.

The temporal approach reveals what words change in meaning and usage between 1674 and 1913 and which words remained relatively consistent. The word "*innocent*" is an example of a word that does not evolve much over time. Each list of neighbors for the word is clearly semantically related in the context of a trial. The word "*murder*", on the other hand, shows some interesting change as time progresses. As shown in Table 5.1, "*murther*" was the nearest neighbor to "*murder*" in the corpus as a whole. However, when looking at Table 5.2, the section of the corpus published between 1674 and 1773 is the only subset of the corpus where "*murther*" even appears in the top 10 most

Year	Word	Nearest Neighbors
1674	female	male, bastard, infant, childe, sex, strangling, unnatural, inhuman, choaking, murdering
	foreign	spain, curious, satyr, portugal, revenue, europe, modern, germany, british, italy
	innocent	wrongfully, unborn, accused, denying, protested, accuse, denial, sorry, deny, blame
	murder	murther, murdering, killing, murdered, barbarous, murthering, rape, wilful, manslaughter, inquest
	prison	gate, jail, goal, bury, gatehouse, gaol, castle, stead, compter, mint
	slave	comfortable, practise, sinning, humane, profession, preservation, acceptable, exercised, covetous, sensual
1774	female	male, woman, prosecutrix, girl, gentlewoman, child, landlady, sister, companion, lady
	foreign	current, spanish, geneva, british, american, russia, french, engraving, base, realm
	innocent	deny, declare, contrary, untrue, unborn, alledged, solemnly, ignorant, innocence, declared
	murder	murdering, murdered, loudly, screamed, kill, scream, wilful, wretch, hallooing, stabbed
	prison	turnkey, confinement, gaol, clerkenwell, bail, shanley, committed, jail, bailed, summoned
	slave	cargo, africa, coast, sailing, island, sail, failed, government, provision, vessel
1874	female	male, woman, prosecutrix, searcher, landlady, niece, daughter, deceased, wife, sister
	foreign	american, spanish, colonial, french, correspondent, advertising, anglo, stationery, manufacture, mercantile
	innocent	deny, untrue, admit, denied, true, wicked, sober, yon, asserted, innocence
	murder	murdering, kill, shooting, murdered, shoot, killing, manslaughter, injure, threatening, accuse
	prison	asylum, lunatic, warder, infirmary, insane, workhouse, bail, patient, released, committal
	slave	*not present in text*

Table 5.2: Results of Word2Vec over temporal subsets

similar words to "*murder*". Additionally, Table 5.3 reports the similarity measure of "*murder*" and "*murther*" for the three time slices and the corpus as a whole. It shows how drastically the similarity drops over time considering "*murther*" is not even present in the third time slice (i.e., after 1874).

Some words and their neighbors can reflect historical events. The word "*slave*", for example, appears in the first two time slices, but is not mentioned at all in the text published after 1874. This may reflect the fact that slavery was abolished in England in 1833, or perhaps show that the language used to describe slaves in England changed. The word "*foreign*" also evolves to have the

word "*american*" as a neighbor in the second two time slices, the majority of which are contained in the period after the American Revolutionary War.

Additionally, the neighbors found by Word2Vec for the word "*female*" have a distinctly different context in the first time slice than in the second two. In the first 100 years of the Proceedings of Old Bailey, words associated with "*female*" refer almost exclusively to children and violence of some kind. This might suggest that women were frequently discussed in the context of being mothers or victims of attacks. In the second and third periods of 100 years, "*female*" is associated with a wider variety of words, but the word "*prosecutrix*", meaning a female victim of some crime, is introduced, maintaining the sentiment from the first time slice that women were often victims. A final word of interest is "*prison*". In the first two time chunks, there is nothing particularly surprising about the word's neighbors, however, after 1874, the words Word2Vec identifies as most similar to "*prison*" place more emphasis on the mental stability of a prisoner. Several of the word's neighbors refer to health rather than guilt.

5.1.3 Pre-trained Word Embeddings

Because data sparsity may be an issue in the Old Bailey corpus, the text was also used to supplement training of a pre-trained word embedding model initially trained on about 100 billion words of Google News dataset.[1] Table 5.3 shows how the pre-trained model compares to Gensim's Word2Vec model in the task of word similarity between "*murder*" and "*murther*". The pre-trained model appears to diminish the similarity because "*murther*" is not present in the data used to train initially train Google's model. Because the time period in which the Proceedings of Old Bailey were published has such drastically different language than that of the data that the pre-trained model was trained on, using the pre-trained model minimizes the importance of words commonly used in the Proceedings.

Year Range	**Word similarity between "*murder*" and "*murther*"**	
	Gensim Model	Pre-trained Model
1674-1773	0.8792	0.7686
1774-1873	0.6841	0.5253
1874-1913	*"murther" not in vocabulary*	*"murther" not in vocabulary*
1674-1913	0.8006	0.7119

Table 5.3: Word similarities between "*murder*" and "*murther*"

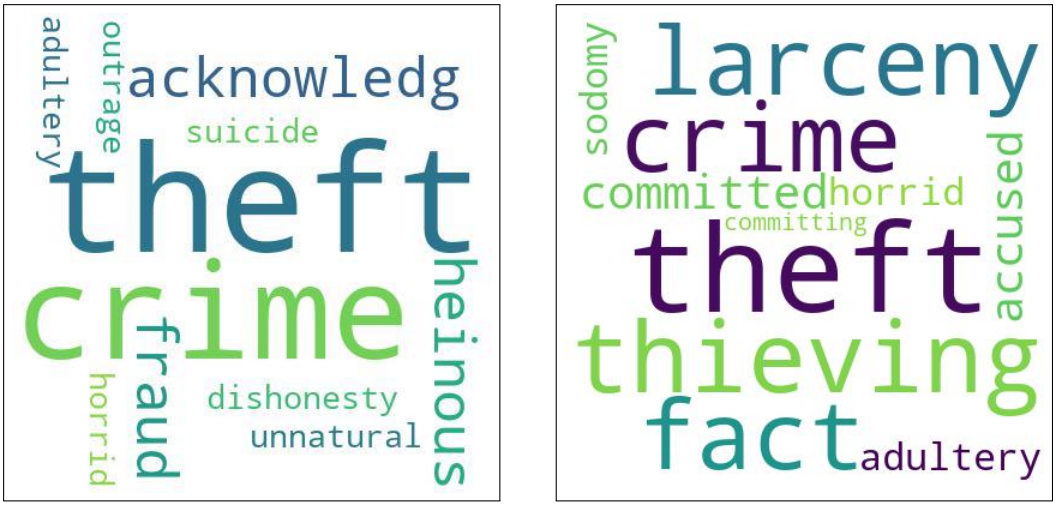
5.2 FastText

Tables 5.1 and 5.2 demonstrate Word2Vec's ability to reveal linguistic change, namely the evolution of word similarity over time, in this corpus. FastText (Bojanowski et al., 2016), an alternative vector space model and an extension to Word2Vec was also explored to determine if it had the capacity to reveal more information about the text. Instead of representing words as vectors like Word2Vec (Mikolov et al., 2013), FastText represents words as character n-grams that have associated vectors (Bojanowski et al., 2016). This means that the model recognizes sequences of characters within

[1]Model was obtained here: https://code.google.com/archive/p/word2vec/.

a word, such as a prefix or suffix, and can define word similarity based on having these character sequences in common.

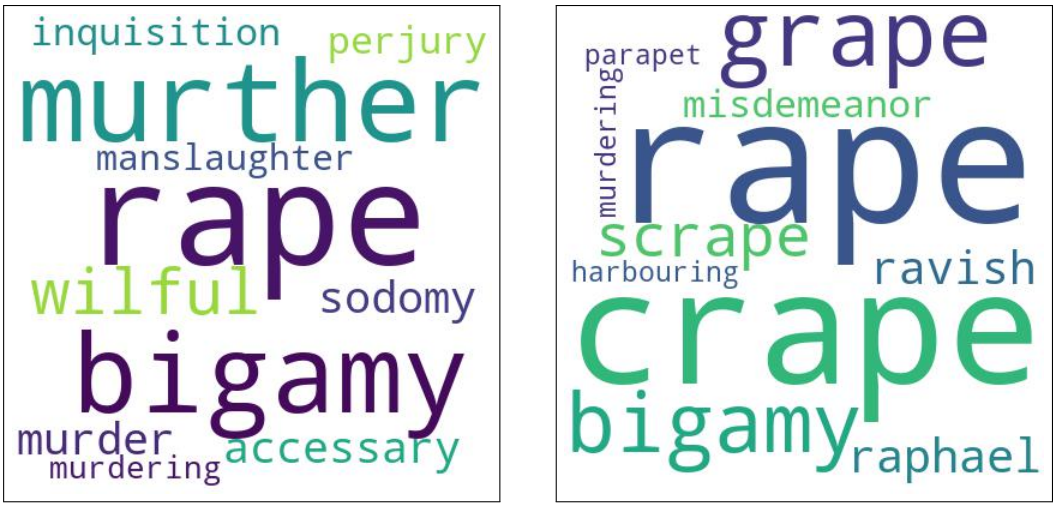
Gensim's FastText wrapper (Řehůřek and Sojka, 2010) was run on the same data as Word2Vec to demonstrate possible differences between the two models' results and to determine which model was best at capturing word similarities and relationships in this particular corpus. The results of running FastText were not drastically different from the Word2Vec results. Some words, "*theft*" for example, have very similar lists of nearest neighbors for both models. Figure 5.1 shows word clouds generated to represent the words Word2Vec and FastText identified as most similar to the word "*theft*". Similar words appear in both word clouds and those that differ only differ slightly.



(a) Word2Vec (b) FastText

Figure 5.1: Wordcloud of words most similar to "*theft*".

Other words, however, reveal an interesting difference between Word2Vec and FastText. Figure 5.2 is a visualization of words most similar to "*rape*" according to the two models. While Word2Vec exclusively selects words related to similar crimes or words commonly used in a trial setting, FastText identifies the closest neighbors to "*rape*" seemingly according to spelling. Three out of four of the closest words picked by FastText appear to have little to no relation to "*rape*" except for the last three letters. This emphasis on shared sequences of characters, as shown in Figure 5.2(a), reflects the model's focus on character n-grams, something that, in this context, is unwanted behavior. Figure 5.2(b), however, shows that Word2Vec consistently reflects word similarities based on meaning and usage. In Chapter 6 and in future work, Word2Vec is preferred over FastText because of this shortcoming.



(a) Word2Vec (b) FastText

Figure 5.2: Wordcloud of words most similar to "*rape*".

Chapter 6

Combined Topic Modeling and Vector Space

Sections 4 and 5 demonstrate value in both applying LDA and Word2Vec to the Proceedings of Old Bailey. LDA primarily focuses on revealing more abstract thematic patterns in the text while Word2Vec reveals more semantic changes and similarities within the corpus. Since both of these methods contribute something meaningful, this section focuses on combining the two methods. Specifically, section 6.1 describes some changes made to the data, section 6.2 explores embedded topic modeling (Dieng et al., 2019b), and section 6.3 shows results obtained from an extension to this method, dynamic embedded topic modeling.

6.1 Data

In all previous sections, each document represented a full day of trial transcripts, as described in section 3.3. However, all experiments in section 6.2 and section 6.3 define a document as a single trial or the text from the front matter of a session in court. This is to avoid mixing together text of trials that may be distinct. Initially, ETM as described in the following section was run using the original definition of a document, but the model found only very generic topics. This new split into individual trials reduces the likelihood that a document will contain a wide range of different proper nouns and crimes and allowed the model to discover more interesting topics. Note that this text was processed the same way as the text used in previous sections (described in section 3.2)

One of the parameters varied in the following sections is whether the data was split by 100 years or by 10 years. When the data is split into time slices of 100 years, this is the same split used in previous dynamic models and described in section 3.3. If the data is split into time slices of 10 years, however, each slice is formed by all the data from a particular decade. This means that the data from the earliest decade, the 1670s, will not include data from all 10 years. This parameter was included because splitting data into sections spanning 10 years creates a more consistent distribution of text across the time slices.

6.2 Embedded Topic Modeling (ETM)

Embedded Topic Modeling (ETM) is a model developed by Dieng et al. (2019b) intended to combine traditional topic modeling with word embeddings. Putting together these two methods creates a

model capable of finding latent topics in the text as well as generating a representation of word meaning (Dieng et al., 2019b). In other methods of topic modeling, such as LDA, a topic is a distribution over the vocabulary. ETM, on the other hand, extends LDA and represents topics as vectors in the embedding space, called topic embeddings. Words are assigned high probability in a given topic using a measure of similarity between the embeddings of the word itself and the topic as a whole. Dieng et al. (2019b) show improved performance with large vocabularies in comparison to traditional LDA. Specifically, ETM shows better topic coherence and diversity in general along with better performance over a corpus containing many stop words.

To find the best possible ETM model, several parameters were varied. The best model is defined as the one that achieved the lowest perplexity on the validation set. As described by Dieng et al. (2019a), the perplexity is calculated by computing the probability that each word in the second half of a test document will appear given the words in the first half of the document.

Table 6.1 shows the parameters varied and the values chosen for the best model. The total number of topics for the model in Table 6.2, 55, was selected by choosing the value used by the best, unigram LDA model with the same corpus run in section 4.1. Note that the minimum document frequency refers to the minimum number of documents that must contain a word in order for it to be included in the vocabulary. Maximum document frequency refers to the maximum percentage of documents a word can appear in for it to be included in the vocabulary (i.e., 0.8 means 80% of documents).

Parameter	**Chosen Value**
Embedding Data	Old Bailey only
Epochs	300
Learning Rate	0.005
Number of Topics	55
Min. Doc. Frequency	100
Max. Doc. Frequency	0.8
Train/Test/Validation document split	85%/10%/5%

Table 6.1: Parameters for best ETM model

Table 6.2 shows some hand-selected topics from the ETM model run with the parameters in Table 6.1 over the entire Proceedings of Old Bailey.

The selection of topics shown in Table 6.2 show some of the same general themes found by LDA in section 4.1, such as the topics consisting mostly of proper nouns, but also show some more meaningful, coherent topics. Nearly all of the topics in this table can be easily given a title that can clearly describe the topic. For example, topic 4 describes dishware and other household items, possibly items that could be stolen, given the context of this corpus. Additionally, topic 14 refers to time and topic 22 consists of words relating to drinking. This is an improvement over the topics found by LDA which were inundated with common words like "*prisoner*", making each topic less coherent. While not all of the topics found by the model run with the parameters in Table 6.1 were so coherent or interesting, the model's capacity to identify this selection of topics shows its value when run using the Proceedings of Old Bailey.

Topic ID	Top 9 Words
0	john, thomas, richard, william, robert, charles, edward, chief, george
1	shop, back, asked, door, green, work, clock, evening, minutes
2	francis, lewis, kelly, moore, morris, collins, morgan, clarke, lee
3	note, person, case, notes, officer, number, bank, warrant, information
4	brass, candlesticks, plates, pewter, blankets, missed, screw, irons, plate
...	...
12	police, examined, station, policeman, custody, charge, sergeant, inspector, statement
...	...
14	clock, morning, night, half, past, saturday, hour, minutes, evening
15	conducted, prosecution, messrs, porter, employed, named, produce, employ, clerk
...	...
22	public, beer, bar, drink, pot, glass, water, called, asked
...	...
33	death, life, dying, god, hath, account, behaviour, murther, die
...	...
53	cab, coach, pistol, stopped, coachman, stood, carriage, happened, stick
...	...

Table 6.2: Results of ETM

6.3 Dynamic Embedded Topic Modeling (D-ETM)

ETM, like LDA, does not consider change over time. Dieng et al. (2019a) also implemented a version of ETM that does consider temporal change. This method, Dynamic Embedded Topic Modeling (D-ETM) (Dieng et al., 2019a), is ETM's equivalent of LDA's dynamic model, DTM, described in section 4.2. D-ETM is an extension of dynamic LDA (Blei and Lafferty, 2006), again representing each topic as a topic embedding. Unlike the representation of a topic in ETM, D-ETM's topic vectors vary over time. The probability of a word in a given topic is calculated the same way as it is in ETM.

The best D-ETM model is defined the same way as the best ETM model (i.e., lowest validation perplexity). To achieve the best model, the same parameters were varied and the values used in the best-performing D-ETM model are shown in table 6.3.

Some hand-selected topic alignments from a D-ETM model run using the parameters described in Table 6.3 are shown in Table 6.4. Note that the topics were not aligned using the method described in section 4.3.1, but were instead found by the D-ETM model itself.

The topics found by the best D-ETM model show some interesting changes over time. In the first time slice of alignment 11, for example, the word "*murther*" is present among other words describing violent crimes. As discovered in section 5.1.3, "*murther*" was only present within the first two time slices, and much more prominent in the first. This is reflected by alignment 11.

Additionally, alignment 13, which appears to describe animals and transportation, shows an increase in the importance of carts, vans, and roads but a decrease in the prominence of horses. This suggests a shift in modes of transportation. Alignment 16 is similar to 13 in that it also describes some modes of transportation, while also including some words related to mail. This

Parameter	Chosen Value
Embedding Data	Old Bailey only
Epochs	300
Learning Rate	0.004
Years Per Time Slice	10
Number of Topics	30
Min. Doc. Frequency	100
Max. Doc. Frequency	0.8
Train/Test/Validation document split	85%/10%/5%

Table 6.3: Parameters for best D-ETM model

alignment is especially interesting in that the words "*train*", "*railway*", and other related words appear in the third time slice.

Finally, it is interesting to note that the word "*prisoner*" appears in two of the topics in alignment 5. These seem to not be very related to the other words in the topic which mostly relate to money and currency. This is similar to an issue found with LDA, although the the word "*prisoner*" is not as dominant in the topics found by the D-ETM model as it is in the topics found by the LDA model. While these topics certainly appear to be more coherent than previous versions of dynamic LDA, many of the topics not shown in Table 6.4 are less informative than those in the table.

Figure 6.1 presents the word evolution of three different hand-selected topics from 1674 to 1913. These visualizations were generated using slight modifications to methods created by Dieng et al. (2019a) to visualize their D-ETM topics. The topics in this figure were found by the same model described in Table 6.3. The words shown in each plot reflect the top words of each given topic at three distinct time slices. These time slices are 1670, 1770, and 1870, intended to align as closely as possible with the way data was split in previous sections, and described in section 3.2. The top three words were chosen from the topic at these distinct time slices and then the probability of each word at every single time slice (i.e., every decade from 1670 to 1910) was plotted to create Figure 6.1.

The first sub-figure in Figure 6.1 reflects this evolution of the second topic found by the model. This topic which includes nouns describing people (e.g., "*child*", "*woman*", "*girl*", etc.) and words related to death (e.g., "*death*" "*found*", "*body*", etc.), appears to relate to murders. What is interesting to note in the plot of this topic's evolution is first how much higher the probability of the word "*child*" is in the earlier decades than all the other words. This may indicate a high rate of death among children. Secondly, the probability of this word drops drastically over the observed decades, likely because the mortality rate for children has declined over time.

The second sub-figure shows the evolution of the 24^{th} topic. From the words chosen by the visualization, this topic seems to relate to items in a home, possibly items that can be stolen, similar to topic 4 in Table 6.2. It's notable that the probability of almost every word declines over time. This may be caused by the fact that the later years contain much more data than the earlier years. The plot might also suggest that "*silver*" was stolen more frequently in the late 1600s and early 1700s than it is in the more recent history.

Index	Year	Topic
0	**1674**	court, law, words, lord, made, great, fine, judgment, offence, till
	1774	court, time, case, bond, made, adair, evidence, hand, law, sir
	1874	court, time, letter, defendant, made, case, examined, day, office, writing
...	...	...
4	**1674**	house, drink, silver, tankard, woman, watch, maid, brought, ale, prisoner
	1774	watch, house, silver, prosecutor, pocket, money, asked, gold, found, woman
	1874	house, public, bar, man, beer, drink, glass, money, left, cross
5	**1674**	money, pounds, found, prisoner, indicted, pieces, shillings, half, ten, gold
	1774	money, half, guineas, shillings, found, guinea, shilling, pence, piece, put
	1874	gave, bad, half, florin, found, shilling, crown, prisoner, money, put
...	...	...
11	**1674**	prisoner, man, deceased, died, murther, killed, manslaughter, wound, kill, sword
	1774	deceased, man, heard, wound, head, murder, door, blood, time, hand
	1874	struck, knife, head, deceased, hand, prisoner, hospital, blood, fell, examined
...	...	...
13	**1674**	horse, mare, found, gelding, stealing, bought, lead, death, horses, indicted
	1774	horse, stable, man, lead, found, cart, hay, field, morning, sack
	1874	cart, horse, examined, van, man, cross, road, coals, car, yard
...	...	...
16	**1674**	coach, man, time, inn, riding, horse, coming, men, horses, aforesaid
	1774	coach, chaise, pistol, man, horse, robbed, box, coachman, sir, side
	1874	office, letter, post, train, letters, ticket, railway, tickets, station, box
...	...	...
26	**1674**	execution, god, death, sentence, ordinary, great, time, men, years, day
	1774	captain, ship, board, time, shore, boat, john, roche, god, life
	1874	board, captain, ship, time, collision, mate, vessel, barge, water, port
...	...	...

Table 6.4: Results of D-ETM

The third and final plot in figure 6.1 shows the evolution of topic 29. While the spike in the probability of "*summary*" sticks out the most in this plot, what is more interesting is the trajectory of "*cheque*" and "*bank*". These two words started off in the 1670s with a probability of zero or near zero. The first forms of banking as we know it today appeared in London in the 17^{th} century and grew in popularity, which is potentially reflected in the plot. The same pattern is true for checks, or cheques, which were formalized in London in the late 1700s and became more widely used over time.

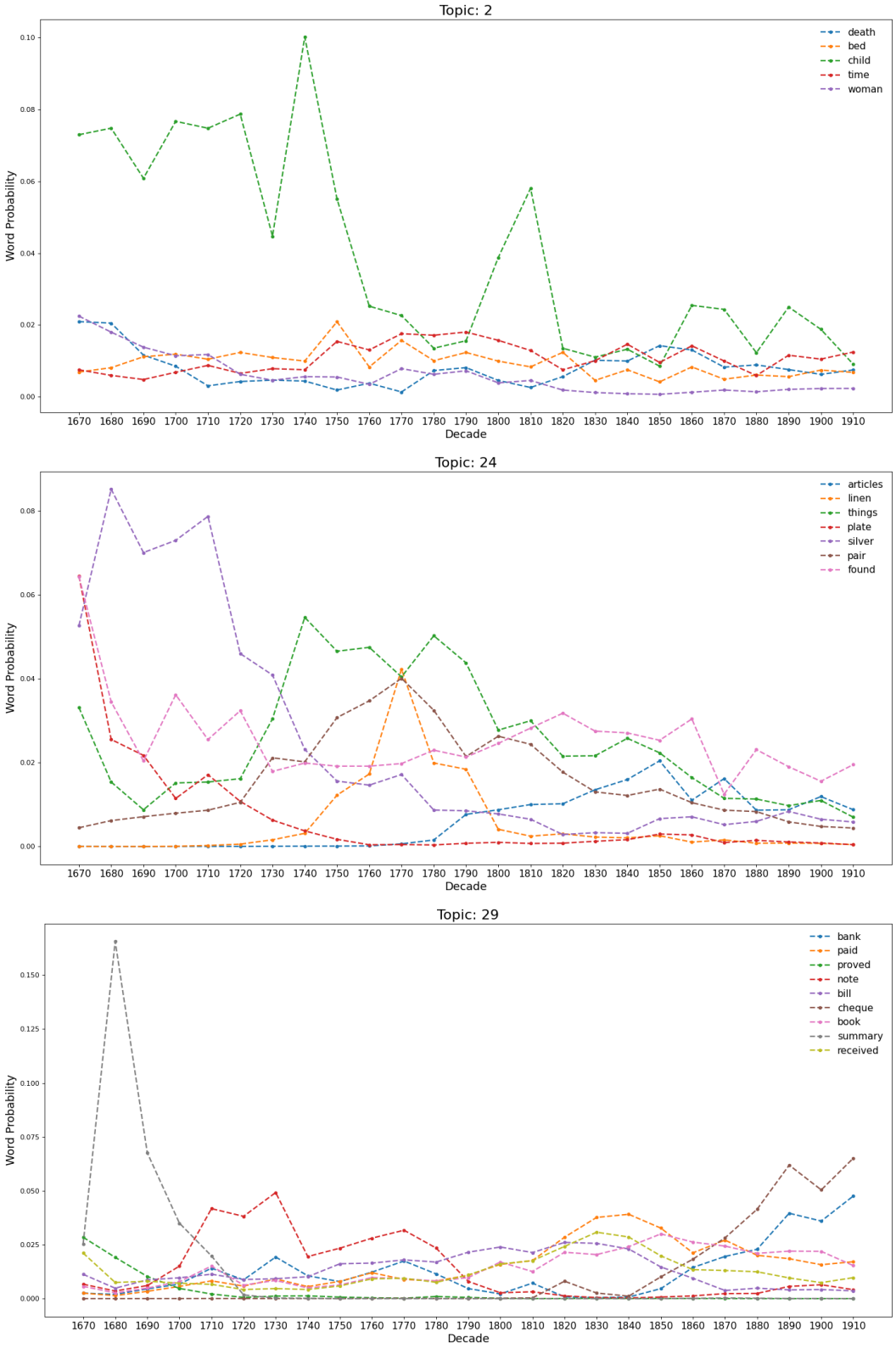


Figure 6.1: Plot of word evolution over time for three topics found by D-ETM.

Chapter 7

Conclusion

Topic modeling and vector space models each highlight different aspects of the Proceedings of Old Bailey. However, because the corpus itself is so unique and idiosyncratic both thematically and in terms of the text itself, many of the methods applied in this project did not reveal as much about the corpus as they might be able to reveal for other, more standard corpora. The variability of the text that comprises the Proceedings coupled with the repetitive nature of the data means that there are many obstacles for LDA, Word2Vec, and the other methods as well. In the future, other vector-space models, like BERT (Devlin et al., 2018) and ELMo (Peters et al., 2018), may be applied to this corpus to see if they can offer some other insights.

A possible modification to the methods described in section 4.3 would be to not enforce that the model for each time slice find the same number of topics. When this is done, it can be harmful for corpora like the Proceedings of Old Bailey in which the amount data in each time slice is not evenly distributed. Instead, for dynamically running LDA over n time slices, it may be beneficial to use the n models with the highest individual topic coherence (i.e., not computing the average over all n models). Doing this would allow each time slice to find the number of topics that produces the highest topic coherence. This may also improve topic alignment which might be able to more accurately output topics that are unique to a given time slice.

The methods described in Chapter 6, however, showed much more promise in identifying interesting topics and patterns in the corpus. Something that ETM and D-ETM do not do, however, is recognize topics that are not present in all time slices. All topics found by these methods span each time slice, thus ignoring topics that may fade out over time or develop in the later documents of the corpus. While the method for identifying these unique topics in section 4.3.1 is not ideal in practice, it is still important not to ignore these topics. Future work may include implementing a way to recognize these topics effectively using ETM and D-ETM.

There are also other possibilities for combining topic modeling and vector-space modeling that may be worth trying. For example, there is a method implemented by Nguyen et al. (2015) which is specifically aimed at working on corpora with relatively small datasets. Another option is to implement a new method of combining topic modeling and vector-space models.

Finally, in future work it may help to explore the annotations included in the original XML data. This includes information about many people mentioned in the corpus such as occupation and gender. Incorporating these annotations into the text may help control the often overwhelming presence of topics found by LDA and other topic models consisting only of proper nouns. Other possible annotations to explore include verdicts and descriptions of punishments.

Bibliography

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating Topic Coherence Using Distributional Semantics. pages 13–22.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

David M. Blei and John D. Lafferty. 2006. Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, USA. ACM.

David M. Blei, Andew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Cameron Blevins. 2010. Topic Modeling Martha Ballard's Diary.

BNC Consortium. 2007. British National Corpus, XML edition. Oxford Text Archive.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information.

David Carter, James Brown, and Adel Rahmani. 2016. Reading the High Court at a Distance: Topic Modelling the Legal Subject Matter And Judicial Activity the High Court of Australia, 1903–2015. *The University of New South Wales Law Journal*, 39:1300.

Stefania Degaetano-Ortlieb. 2018. Stylistic Variation Over 200 Years of Court Proceedings According to Gender and Social Class. In *Proceedings of the Second Workshop on Stylistic Variation*, pages 1–10, New Orleans. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2019a. The Dynamic Embedded Topic Model.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2019b. Topic Modeling in Embedding Spaces.

DocSouth. Documenting the American South. University Library, The University of North Carolina at Chapel Hill.

Clive Emsley, Tim Hitchcock, and Robert Shoemaker. The Proceedings – Ordinary of Newgate's Accounts.

Jonathan Gordon, Linhong Zhu, Aram Galstyan, Prem Natarajan, and Gully Burns. 2016. Modeling Concept Dependencies in a Scientific Corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–875, Berlin, Germany. Association for Computational Linguistics.

Tim Hitchcock, Robert Shoemaker, Sharon Howard, Jamie McLaughlin, et al. 2012a. London Lives, 1690-1800.

Tim Hitchcock, Robert Shoemaker, et al. The Proceedings of the Old Bailey – Methods.

Tim Hitchcock, Robert Shoemaker, et al. 2012b. The Old Bailey Proceedings Online, 1674-1913.

Matthew L. Jockers and David Minmo. 2012. Significant Themes in 19th-Century Literature. pages 750–69.

Nikita Kitaev and Dan Klein. 2018. Constituency Parsing with a Self-Attentive Encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics.

Sara Klingenstein, Tim Hitchcock, and Simon DeDeo. 2014. The Civilizing Process in London's Old Bailey. *Proceedings of the National Academy of Sciences*, 111(26):9419–9424.

Jey Lau, David Newman, and Timothy Baldwin. 2014. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. *14th Conference of the European Chapter of the Association for Computational Linguistics 2014, EACL 2014*, pages 530–539.

Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit.

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space.

David Mimno. 2012. Computational Historiography: Data Mining in a Century of Classics Journals. *J. Comput. Cult. Herit.*, 5(1):3:1–3:19.

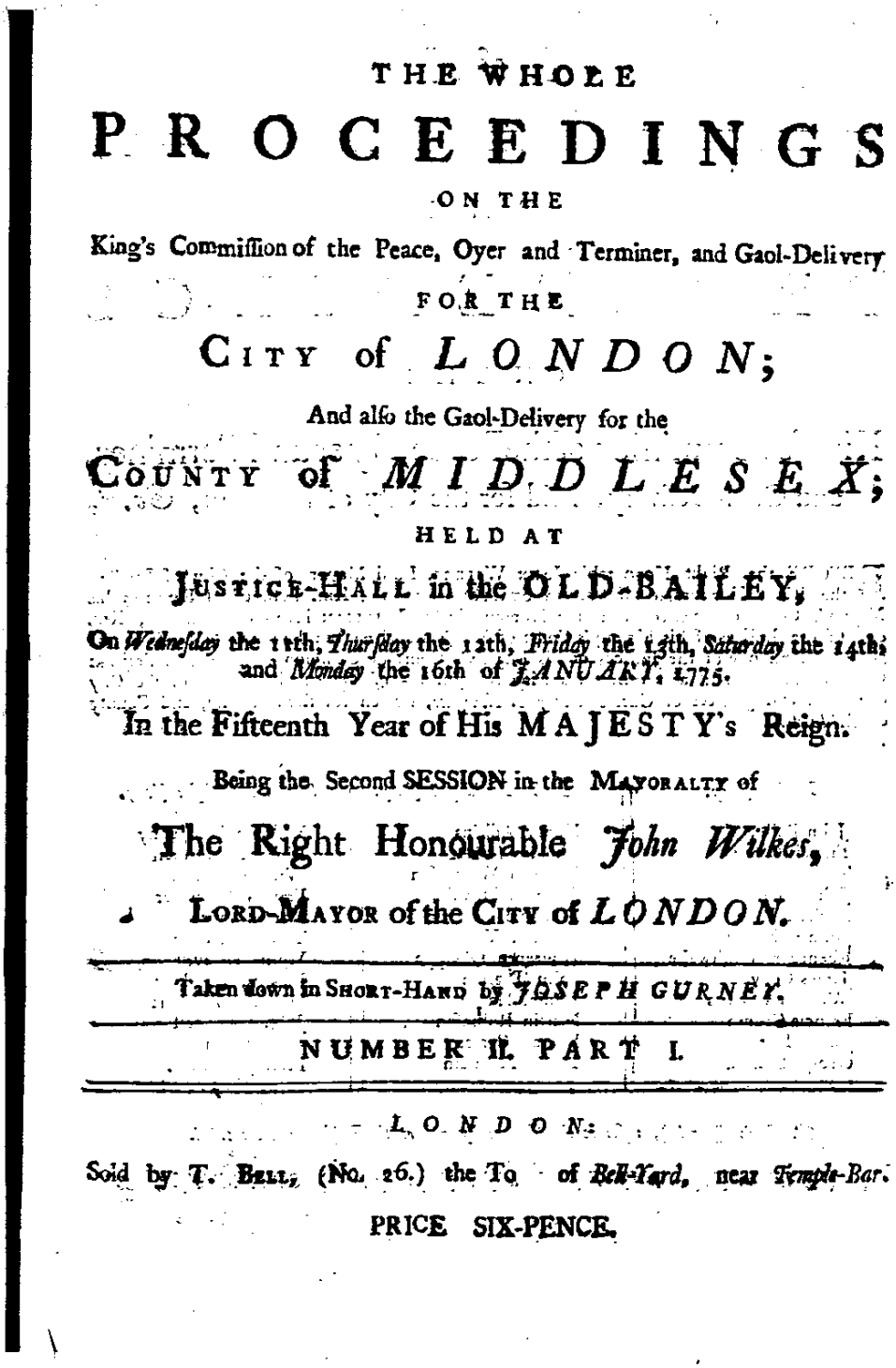
Elaheh Momeni, Shanika Karunasekera, Palash Goyal, and Kristina Lerman. 2018. Modeling Evolution of Topics in Large-Scale Temporal Text Corpora. In *International AAAI Conference on Web and Social Media*.

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015. Improving Topic Models with Latent Feature Word Representations. *Transactions of the Association for Computational Linguistics*, 3(0):299–313.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Appendices



THE WHOLE

PROCEEDINGS

ON THE

King's Commiſſion of the Peace, Oyer and Terminer, and Gaol-Delivery

FOR THE

CITY of *LONDON*;

And alſo the Gaol-Delivery for the

COUNTY of *MIDDLESEX*;

HELD AT

JUSTICE-HALL in the OLD-BAILEY,

On *Wedneſday* the 11th, *Thurſday* the 12th, *Friday* the 13th, *Saturday* the 14th, and *Monday* the 16th of *JANUARY*, 1775.

In the Fifteenth Year of His MAJESTY's Reign.

Being the Second SESSION in the MAYORALTY of

The Right Honourable *John Wilkes*,

LORD-MAYOR of the CITY of *LONDON*.

Taken down in SHORT-HAND by *JOSEPH GURNEY*.

NUMBER II. PART I.

LONDON:

Sold by T. BELL, (No. 26.) the To of *Bell-Yard*, near *Temple-Bar*.

PRICE SIX-PENCE.

Figure 1: Example front matter from January 11, 1775

THE

PROCEEDINGS

ON THE

King's Commission of the Peace, Oyer and Terminer, and Gaol-Delivery, held for the City of LONDON, &c.

BEFORE the Right Honourable JOHN WILKES, Lord Mayor of the City of *London*; the Right Honourable Sir SYDNEY STAFFORD SMYTHE, Knt. Lord Chief Baron of his Majesty's Court of Exchequer*; the Honourable Sir WILLIAM HENRY ASHHURST, Knt. one of the Justices of his Majesty's Court of King's Bench†; the Honourable Sir GEORGE NARES, Knt. one of the Justices of his Majesty's Court of Common Pleas‖; Mr. Serjeant GLYNN, Recorder‡; and others his Majesty's Justices of *Oyer* and *Terminer* of the City of *London*, and Justices of *Gaol-Delivery* of *Newgate*, holden for the said City, and County of *Middlesex*.

London Jury.	*First Middlesex Jury.*	*Second Middlesex Jury.*
John Bailey	Henry Adkins	William Halfpenny
Francis Carey	Thomas Lovett	Richard Marshall
Thomas Oades	William Wilkinson	William Leader
John Jacobs	John Braithwaite	Francis Crowther
Richard Rice	Nathaniel Morgan	Norton Tyler
John Robinson	Richard Wall	Thomas Clarke
Ebenezer Braithwaite	Lawrence D'Rippe	William Cavill
James Coby	John Pearson	Edward Hawkins
William Roberts	Daniel Hardy	William Lambert
Francis Crump	Martin Robinson	John Campbell
James Harris	Francis Carter	John Smith
George Seaton.	David Fountain	Peter Johannett.

*The *, †, ‖ and ‡, refer to the Judges by whom the Prisoners were tried.*

(L.) *London Jury.* (M.) *First Middlesex Jury.* (2d M.) *Second Middlesex.*

Figure 2: Example introduction to a session in the Old Bailey courthouse from January 11, 1775

at another lodging with all the goods I loft, in her poffeffion.

(*They were produced in court, and depofed to by the profecutor.*

Prifoner's Defence.

I did not ſteal the goods, another of the profecutors lodgers brought them to me, and he deſired me to let them ſtand in my lodgings, I did not know what they were.

Guilty T.

---

96. (*M.*) WILLIAM HOLLAND, was indicted for ſtealing ſix yards of black ſilk lace, value thirty-ſhillings, the property of Ann Sowerby, ſpinſter. December 10th.

Ann Sowerby. I keep a milliner's ſhop in Round-court. On the 10th of December, I juſt ſtept out of the ſhop to call the maid; as I came up ſtairs, I ſaw the priſoner in the ſhop with a card of black ſilk lace in his hand, I laid hold of him, and he threw the lace down; he had taken it out of the window, then he aſked me for a half-pennyworth of what he called ſilk galloon.

Q. What account did he give of himſelf?

Sowerby. When he was before the juſtice he ſaid he had worked a week for a paper ſtainer, and that he lodged in a two-penny lodging at St. Giles's.

Priſoner's Defence.

I am not thirteen years old.

"The priſoner called his father, who ſaid he "is a green-grocer in Covent garden, that he "had the ſmall pox in his family, and therefore "took a lodging for the priſoner, in St. Giles's, "and that the priſoner always behaved well.

Guilty T.

---

97. (*M.*) WILLIAM MATTHEWS was indicted for breaking and entering the dwelling houſe of Samuel Lyon, on the 12th of December, about the hour of twelve at noon, (no perſon being in the ſaid dwelling houſe) and ſtealing a woollen cloth coat, value ten-ſhillings, a woollen cloth waiſtcoat, value ſix-ſhillings, two pair of leather breeches, value ten-ſhillings, and a woollen ſurtout coat, value ten-ſhillings, the property of the ſaid Samuel Lyon in his dwelling houſe.

——— *Creſby.* I was collecting the poors rates, in New Inn on the 12th of December, while I was ſtanding at the bottom of No. 1. a perſon let himſelf out of a window, he brushed my back, got up and run away, this made me ſuſpect there was ſomething wrong, I went up the ſtairs, and upon the landing place I found the priſoner ſitting upon a ſack, we ſecured him, and upon examining the ſack, we found it contained the things mentioned in the indictment.

Mr. Creſby's evidence was confirmed by a Gentleman who was with him. (The goods produced in court.

Samuel Lyon. The cloths produced are my property, I left them in my chambers when I went out, I think there muſt be more than one concerned in the buſineſs, They got in at a window that is very high, I am a tall man, and when I ſtood upon the bannifters I could not reach the window, I apprehend one muſt ſtand upon another, they had taken an iron bar out of that window.

Priſoner's Defence.

The other perſon who is a chimney-ſweeper, deſired me to take care of his things, whilſt he went to No. 2. at Lyon's Inn; I waited a good while for him, but he not coming, I went up ſtairs, and there I found a ſack, I know nothing of ſtealing the goods.

He called three witneſſes who gave him a good character.

Not guilty of breaking and entering the dwelling-houſe, but guilty of ſtealing the goods, T.

---

98. 99. (*M.*) ABRAHAM DUGARD, and RICHARD BANFIELD, were indicted for wilfully, maliciouſly, and feloniouſly making an aſſault upon Dodding Jonathan Bruce, and with menaces, and in a forceable and violent manner, feloniouſly demanding his money, with intent the money of the ſaid Dodding Jonathan Bruce to ſteal, againſt the ſtatute, December 14.

Dodding Jonathan Bruce. I had been out with my wife to ſpend the evening, on Wed-neſday