Topic modelling on archive documents from the 1970s: Global policies on refugees

Philip Grant

Independent Researcher, UK

Ratan Sebastian

University of Edinburgh, UK

Marc Allassonnière-Tang

Lab Dynamics of Language UMR 5596, France

Sara Cosemans

KU Leuven Research Unit of History, Belgium

Abstract

This study conducts a historical analysis of global policies on refugees within typewritten and digitally-born documents (c. 55,000 pages) from international and national archives. The data originates from the 1970s and is stored in archives from the UK and US governments, plus the United Nations High Commissioner for Refugees (UNHCR). The overarching theme is to analyse the involvement of the UK, the US, and the UNHCR in different refugee cases that occurred during the 1970s. To do so, we (1) identify the main topics in each document; (2) investigate the transmission of topics horizontally (between organizations) and vertically (through time); (3) suggest targeted areas of the document set for further close reading by historians. Standard Optical Character Recognition (OCR) and object detection are used to extract information from documents and categorize them. Then, NLP methods like topic modelling and clustering are used to identify topics and the relationships between them across time. The results identify several main themes covered by different organizations and how the focus of each organization changes diachronically. Besides its academic contribution, this study also demonstrates how, through the use of existing techniques with limited customization, digital technologies in the hands of the historian can augment and complement qualitative methods in bringing to light the themes and trends demonstrated in large bodies of historical documents.

1 Introduction

International history based on bureaucratic and diplomatic source material often involves large scale multi-archival research across the world, resulting in large quantities of research pictures. One of the major challenges of this type of research is the processing of thousands of these images upon return from a research trip. One of the potential solutions is to use NLP methods on historical source materials, collected from the analogue archives by historians themselves. The current study provides preliminary results of this type of solution for archival source material, largely written on typewriters, from the 1970s.

The study is embedded in a larger international history project that focuses on refugee resettlement under coordination of the international refugee organization, the United Nations High Commissioner for Refugees (UNHCR), between 1972 and 1979[1]. For the purpose of this paper, we concentrate on the three major case studies of that period: (1) the expulsion of the Asian diaspora - a group of British passport holders with roots in the Indian Subcontinent - from Uganda by General Idi Amin in 1972 and their subsequent resettlement in more than 25 countries; (2) the exile of Chileans and Latin American refugees in Chile after the coup of General Augusto Pinochet in 1973, which led to UNHCR-assisted resettlement in 55 countries; and (3) the escape of Vietnamese 'boat people' from the fall of Saigon in 1975 onwards that resulted in the establishment of the international 'Orderly Departure Program' in 1979. The historical objective is to analyse the role of various (state and non-state) actors and their discourses, as well as the outcomes of resettlement programs.

The research rests on the premise that the 1970s formed the apogee of resettlement (as one out of the three durable solutions in the Statute of UNHCR) for refugees from the Global South. The 'success' of resettlement in terms of the numbers of resettlement countries, numbers of refugees resettled, and resources allocated to resettlement was unseen before and never repeated in later decades. Policymakers and scholars concerned with refugee protection have advocated for stronger resettlement programs (Noll and van Selm, 2003; Hashimoto, 2018), but it is actually not well understood why refugee resettlement in the 1970s became such a popular solution. In order to understand possible evolutions, and the nature, causes, and chronology of these resettlement programs in the 1970s, research was conducted on archival material from 18 archives in 9 countries in 12 different languages (predominantly English, Spanish, and French), resulting in 94,322 research pictures.

The objective from a digital humanities point of view is to develop an automated approach to help the historian deal with large quantities incompletely tagged, non-digitally-born data (i.e. research pictures with imperfect OCR and missing metadata) by (1) identifying the main topics in the dataset; (2) investigating the transmission of topics horizontally (between

[1] DETAILS PROJECT (will be added once review phase is completed)

organizations) and vertically (through time); and (3) suggesting targeted areas of the document set for further close reading by historians. Since the case studies were known to the historian at the moment of data collection, we expected to find topics related to these case studies. However, we also expected to find topics that would point towards evolutions within and across the case studies in order to better understand where policymaking on refugee resettlement originated, how actors communicated, and copied or adapted ideas and practices, and which overarching discourse emerged throughout these processes.

It is becoming increasingly common to use computational methods on historical data (Baierer *et al*., 2019; Riedl and Padó, 2018): particularly, textual Latent Dirichlet Analysis (LDA), a topic modelling method that enables working with natural language instead of metadata. This method was introduced by Blei (2003) and is interesting to humanities researchers asking questions about the vertical and horizontal transmission of ideas in bodies of text, ranging from literature to newspapers (Chandelier *et al*., 2018; Tangherlini and Leonard, 2013). These analyses commonly use existing large digitally-born data sets. Historians interested in that kind of research are increasingly gaining access to digital historical data. For instance, (Allen and Connelly, 2016) surveyed studies of digitized diplomatic cables using topic modelling methods to study the prevalence of various topics over time. In Gao *et al*., (2017) metadata about State Department documents were used to identify communications that warrant closer study and thus avoid the problem of dealing with natural language. Risi *et al*. (2019) explored how the use of topic models could allow a hypothetical 'artificial archivist' to identify historically significant documents. However, digitally-born original (i.e. unpublished) historical data exists only from the introduction of text processing computers onwards. The first data of this kind date back to the early 1970s and were produced in the American administration. For that reason, historians of earlier time periods conduct digital research on smaller manually transcribed (sometimes crowd-sourced) data sets (Barron *et al*., 2018; Roe *et al.*, 2016; Romein *et al*., 2020).

Besides its academic contribution in global policies of refugees during the 1970s, this study also contributes in terms of methodology by suggesting digital solutions for historians who manually select large quantities of non-digital typewritten data and store them as digital facsimiles, in the form of research pictures with underlying OCR text of varying quality. The aim is explicitly to keep 'the human in the loop' and to supply solutions explicitly geared towards the needs of the historian, as far as possible using standard existing tools without the need for extensive customized technology. We first describe the text corpus and its digitization before sharing the results of two NLP methods (topic modelling and clustering). Then, we share some preliminary results that can shed a light on how the NLP analysis assisted in resolving some of the historical objectives of the larger research project on refugee resettlement. Finally, we suggest work to develop our method further with the possibility of it becoming a standardized toolkit for historians working with typewritten sources from the end of the nineteenth century onwards.

2 The Text Corpus

From the 94,322 research pictures (the equivalent of the same number of pages) collected in the archives, we selected a sample from two archives: the British National Archives (hereafter TNA) at Kew and the Archives of the United Nations High Commissioner for Refugees (hereafter UNHCR) in Geneva. In these two collections we further selected only the documents in English, in the date range between August 1972 and December 1979, of the three case studies mentioned in the introduction. The case studies were determined based on document titles assigned by {Author X} while conducting research in the archives. Information on language and date for each page in the UNHCR collection was manually added in a metadata database by the historian (see figure 1). For TNA, information on the date was extracted by computer vision methods (see section 3.2). In the latter collection, all pages were in English. This resulted in 4679 records (6962 pages) from UNHCR and 7276 records (17,647 pages) from TNA.

Furthermore, we selected 21,886 records, relevant to the case studies from the Access to Archival Database (hereafter AAD), which is part of the American National Archives and Records Administration. The AAD website, which contains 3,225,364 US Department of State Cables from the Central Foreign Policy Files created between 1 July 1973 and 31 December 1979, is popular among historical researchers interested in digital methods and big data (Allen and Connelly, 2016; Gao *et al*., 2017; Risi *et al*., 2019)[2]. Due to its time frame and content, AAD is valuable for the purposes of the historical research questions. Moreover, as these data are digitally born, they serve as a comparison point to the messier data derived through OCR. After excluding the records that contained no full text (due to restrictions by either Executive Orders or other laws or regulations of the US Department of State) 8646 AAD records were retained. A precise page count for the AAD data is not available based on the metadata in the database.

3 Digitizing documents

In this section, we explain how the paper documents were digitized and how metadata (dates) was extracted from those documents, using optical character recognition (OCR) and computer vision[3].

[2] https://aad.archives.gov/

[3] The following Python packages are used in the study as a whole: colorlover, Fitz, FuzzyWuzzy, gensim, matplotlib, nltk, numpy, opencv, orca, pandas, plotly, pyLDAvis, pyTorch, scipy, sklearn, and SpaCy.

3.1 Optical Character Recognition to digitize documents

The workflow for extracting information from the analogue archived documents (UNHCR and TNA) began with a physical visit to the archives to scan or digitally photograph each page of each document. To convert the images into text, the ABBYY FineReader OCR package was used. The resolution of all the digital images was at least 300 dpi, but image quality was highly contingent on the quality of the paper, as standards for acid-free paper for archival purposes were not established until the 1980s (Kurlansky, 2017, p. 255). The UNHCR data is especially affected by the use of poor-quality paper and ink. For that reason, UNHCR data was manually inserted in a database before OCRing. Metadata about 'document genre' helped to group the pages (common document genres are memorandums, reports, press articles, cables, among others. For a full list please refer to Appendix 1). Because of the high internal consistency per genre and the large disparities between them, each genre was then OCRed separately. Training the OCR software per genre, as was first intended, came to naught, as the process proved too time-consuming for a single researcher, who was also in charge of the content analysis. The failure of the attempt to train OCR software led us to accept the limitations of the source material and to explore the usefulness of NLP methods on imperfect data.

Figure 1 shows an example of a picture in the genre 'incoming cable'. Metadata apart from the genre include the date, the language, the degree of confidentiality, the initials of the secretary, and the place (box, file, page) where the original document can be found in the archive. Several genres follow their own templates with a lot of pro forma text. For instance, each incoming cable had the same heading, which was filled out by the secretary who received the cable. We attempted to filter out this pro forma text during the preprocessing phase as much as possible (section 4.1).

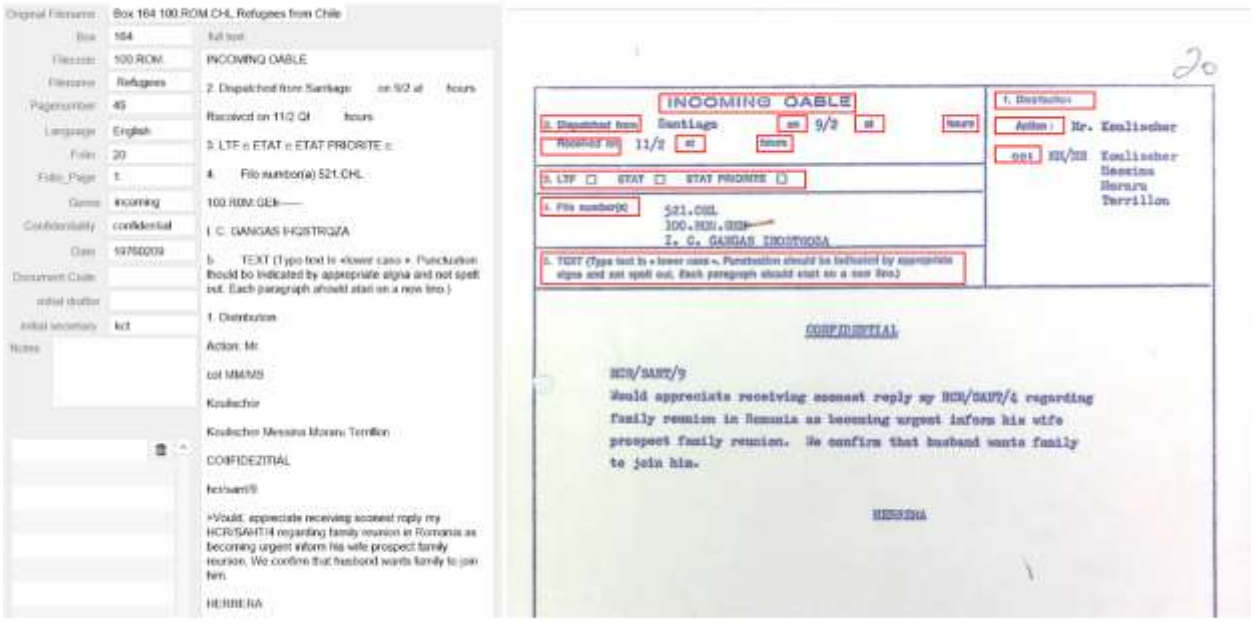

Figure 1. An example of a database entry including the document from the UNHCR archives with the pro forma text marked in red.

When the UNHCR documents were originally scanned, the OCR quality was calculated as an overall percentage per document genre[4]. The genres not only depended on the content, but also on the quality of paper, ink, and the typewriter. The genres with OCR quality above 80% ('memorandum', 'reports', 'letters', 'press articles') comprised 62% of the UNHCR documents. The remaining genres ('drafts', 'cables') had OCR quality lower than or equal to 80%. The genre with the lowest OCR quality was the 'draft agreement' with only 68% OCR quality. The TNA data was collected at different points in time (2014 and 2017 respectively) with two different devices. These differences, however, seem to have had only a limited impact on OCR quality with 88% for the documents photographed in 2014 and 90% for those of 2017 (see Appendix 1).

An acceptable baseline of OCR quality for topic modelling has been proposed at 80% by van Strien *et al*., (2020), who observed that low quality OCR mattered for mathematical measures of topic coherence (though less so for qualitative interpretability of the topics). Since over one third of the UNHCR data fell below the 80% threshold, we investigated the potential impact of

[4] The OCR quality of a document is equal to 1 minus the ratio of low confidence characters in the document. The OCR quality as a measure does not include the characters that are not at all recognized (which are seen by the software as 'images' rather than 'text').

the low quality on the results (Figure 2). For the UNHCR data, we calculated, for each document genre, the average number of lemmas per page that were fed into the topic model after the preprocessing steps which remove data that is likely to be irrelevant or erroneous (section 3.1). This served as a broad measure of the information content of each document type.

Figure 2: Average number of lemmas per page input into the topic model, vs OCR-Q score, per document type in the UNHCR corpus. Vertical dashed line represents the 80% threshold suggested by van Strien *et al*. (2020); horizontal dashed line represents the average lemma count per page across all UNHCR documents included in the topic model. Genre abbreviations are as per Appendix 1.

We observe that the information content for documents in the 65-80% OCR-Q range is similar to that for documents in the 80-90% range. Only above 90% OCR-Q does the lemma count significantly rise. It is important to note, however, that the genres with the highest OCR quality ('external reports', 'memoranda') are also those that contain more text per page. Based on this, we did not attempt to remove the genres with lower OCR quality from the dataset for topic modelling.

3.2 Computer vision to extract date information

While the UNHCR and AAD archives have metadata to indicate to which record a page belongs and a representative date for the topics in the record, the TNA archive lacks this. This information was automatically extracted from the pages using a combination of computer vision methods and the available OCR results.

For the first issue of figuring out which pages belong to which record, we relied on the fact that pages in a record are consecutive and that the first page in a record has a stamp. The stamp can be detected using a Faster R-CNN object detection model (Ren *et al*., 2016). A pre-trained PyTorch model was adapted to perform this task with transfer learning. The presence of consistent textures in the stamp allowed training the model to a reasonable performance (average precision of 0.825 and average recall 0.860 with IoU[5]: [0.5:0.95]) with 600 image annotations for training, testing and validation. Once a stamp was detected, all consecutive following pages were assigned to the same record, until the next page with a stamp was found.

For the second issue of identifying the date of a record, Faster R-CNN training did not lead to a reasonable accuracy using the same amount of data. Instead, the fact that authorship dates for a record are usually visually distinct from the rest of the document and have a standard format as text (Figure 3) was used.

Figure 3: A typical example of a TNA research picture, with an overlying piece of paper, handwriting, and in the red square the rubber stamp with the date that indicates the start of a new record.

[5] Intersection over Union. A measure of object detection accuracy that quantifies how much of the detected stamp overlaps with the actual stamp.

The page image was binarized, dilated, and then watershed segmentation (Kim *et al*., 2014) was applied to identify areas of the image that contained contiguous pieces of text. Next, the text identified by OCR in each of these areas was matched against a regular expression to identify dates. If a record contained multiple pages authored on multiple dates, the median of the detected dates was used to represent the date of the record.

4 Analysing documents

This section describes how the content of the digitized paper documents and the digitally-born documents was analysed using topic modelling and topic clustering.

4.1 Data preparation for topic modelling

For the topic modelling process, we used all selected data including the digitized research pictures from TNA (7276 records, 17647 pages) and UNHCR (4679 records, 6962 pages) and the 8646 digitally-born records from AAD. The text was tokenized and lemmatized using SpaCy's 'en_core_web_sm' model which has a tagging accuracy of 97%. Several preprocessing steps were necessary to mitigate quality issues specific to this data, as well as to distinguish content terms from function words (which reflect grammatical structure, or the mechanics of a communication medium, rather than semantic meaning, and so are not relevant for topic modelling).

First, for the TNA texts, all line breaks were removed before tokenization, since we found that a large number of spurious line breaks had been added by the OCR software, breaking up words incorrectly into small fragments. Second, tokens were excluded if either the token itself, or its lemma, fell within a set of stopwords: (a) NLTK's set of English stopwords; (b) words that are clear OCR artefacts, such as 'wtth' (almost always a mis-scan of 'with'); (c) words that relate to the mechanics of cable distribution, such as 'cable' and 'telecommunications'; and (d) words that are function words, as opposed to content words, but are not included in NLTK's standard list (such as 'although' and 'however'). Apart from the NLTK stopword set, these terms were identified and added to the exclusion list iteratively when they were found to appear in the topic

modelling results. Third, particular multi-word sequences which referred to a single semantic entity were treated as a single lemma (e.g. 'New York', 'Kuala Lumpur', 'United Nations High Commissioner For Refugees'). SpaCy's named entity recognizer (NER) is intended for this purpose, but did not identify many of the relevant cases in this dataset, and so this processing was instead done explicitly in Python code. Fourth, tokens were also removed if they were punctuation, shorter than three characters, or if less than half of the token consisted of ASCII alphabetical characters (words that fail this last criterion in an English document are almost certainly OCR artefacts).

These preprocessing steps certainly did not make the dataset 'perfect' for topic modelling purposes. They simply removed mis-scanned words rather than replacing them with their correct form, and they did not address low-frequency mis-scanned words that were not prominent in the topic modelling results. The level of preprocessing was sufficient, however, to give the useful results we describe. Following the preprocessing, a document–lemma count matrix was constructed using sklearn's standard 'bag of words' count vectorizer. At this stage, lemmas were excluded from consideration if they occurred in fewer than 3 (three), or more than 30%, of the documents in the dataset. Table 1 shows, per archive, the average lemma counts for the documents that formed the input to the topic model itself. The close proximity of the number of lemmas per document between the most clean data (the digitally-born AAD records) and the least well OCRed data (from the UNHCR archives) suggests that all three archives in the data set contributed information broadly equally to the topic model.

Archive	Doc count	Page count	Words per doc	Lemmas per doc, before pre-proc	Lemmas per doc, after pre-proc	Words per page	Lemmas per page, before pre-proc	Lemmas per page, after pre-proc
AAD	8645		305.7	159.3	139.0			
TNA	7207	17568	532.4	265.0	217.1	218.4	108.7	89.0
UNHCR	4530	6250	275.2	153.7	128.0	199.5	111.4	92.8
All	*20382*		*379.1*	*195.4*	*164.2*			
All paper-based	*11737*	*23818*	*433.1*	*222.0*	*182.7*	*213.4*	*109.4*	*90.0*

Table 1: Information content of the set of documents included in the topic modelling. These figures are averages across all the documents and/or pages in each archive. Note that AAD consists of digital data, without pages, so per-page counts are not relevant. 'Words' here means the result of splitting the raw document by spaces; 'Lemmas before pre-proc' refers to the output of SpaCy's standard lemmatizer, and 'Lemmas after pre-proc' here means the result after all the pre-processing steps carried out as described above.

After the preprocessing steps described, some documents were left with no valid lemmas at all (generally because of poor OCR quality), and are not included here. This amounted to 95 documents (112 pages) in total: 0.34% of the pages in the dataset as a whole.

4.2 Topic modelling

From the document–lemma count matrix, the topics can be identified either using Latent Dirichlet Allocation (LDA) or Hierarchical Dirichlet Processing (HDP). As shown in Figure 4, LDA (Blei *et al*., 2003) models each document as a probabilistic mixture of topics, and each topic as a probabilistic mixture of words. Given a corpus, and the desired number of topics as a hyperparameter, the model assigns a weighted combination of topics to each document in the corpus, with each topic being a weighted combination of terms, so as to best fit the text of the corpus. The resulting weights are expressed as a document–topic matrix and a topic–term matrix. HDP (Teh *et al*., 2006) is similar, but, rather than requiring the number of topics as a parameter, attempts to infer the optimal number of topics from the data.

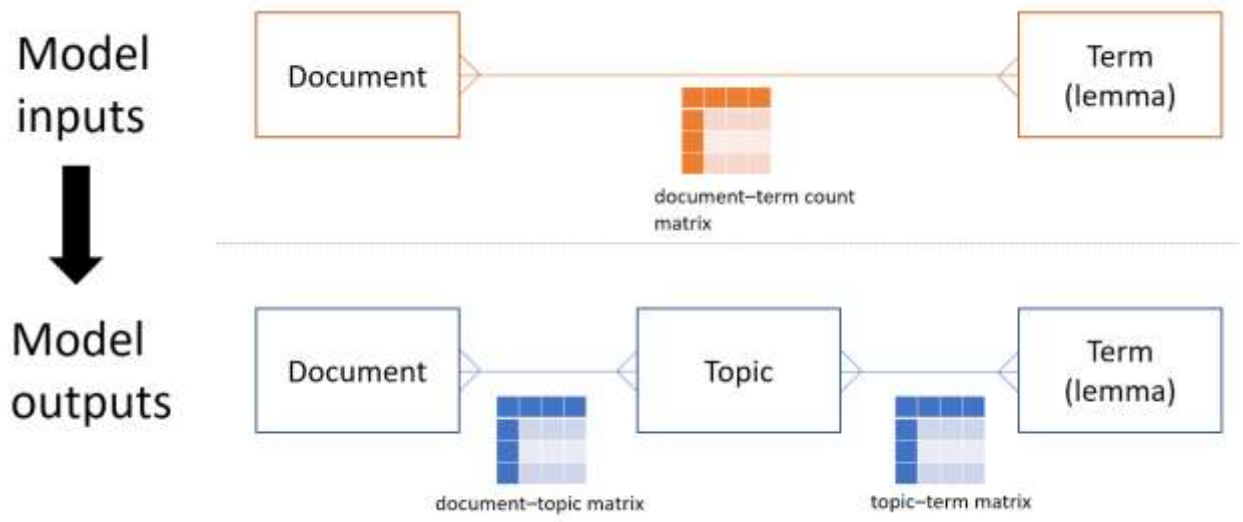

Figure 4: Conceptual view of topic modelling

During this work, although metadata (date, source archive, and the historian's categorization of the documents into the three 'case studies' described in the introduction) was

used to select which documents were in scope, it was not an input to the LDA or HDP model and so the model was 'blind' to this information when determining topics; its input was only each document's text, preprocessed as described in section 4.1.

Since the HDP algorithm automatically selects the optimal number of topics, it was our initial choice. However, it resulted in over 100 topics, which, while a mathematically optimal result according to the algorithm, was not clearly historically interpretable. Given this result, and our aim of using computational modelling to suggest avenues of further exploration for historians (rather than attempting to generate conclusions autonomously), we moved to a 'human-in-the-loop' approach, using the LDA algorithm, experimenting with different numbers of topics, and working iteratively with {Author X} to identify the optimum number of topics (more details in the Preliminary Results section).

For each number of topics tried (8, 12, 16 and 20), first, the PyLDAVis package was used to produce interactive plots which showed the saliency and relevance of each term within each topic, and the level of similarity or difference between topics. Figure 5 shows an example plot.

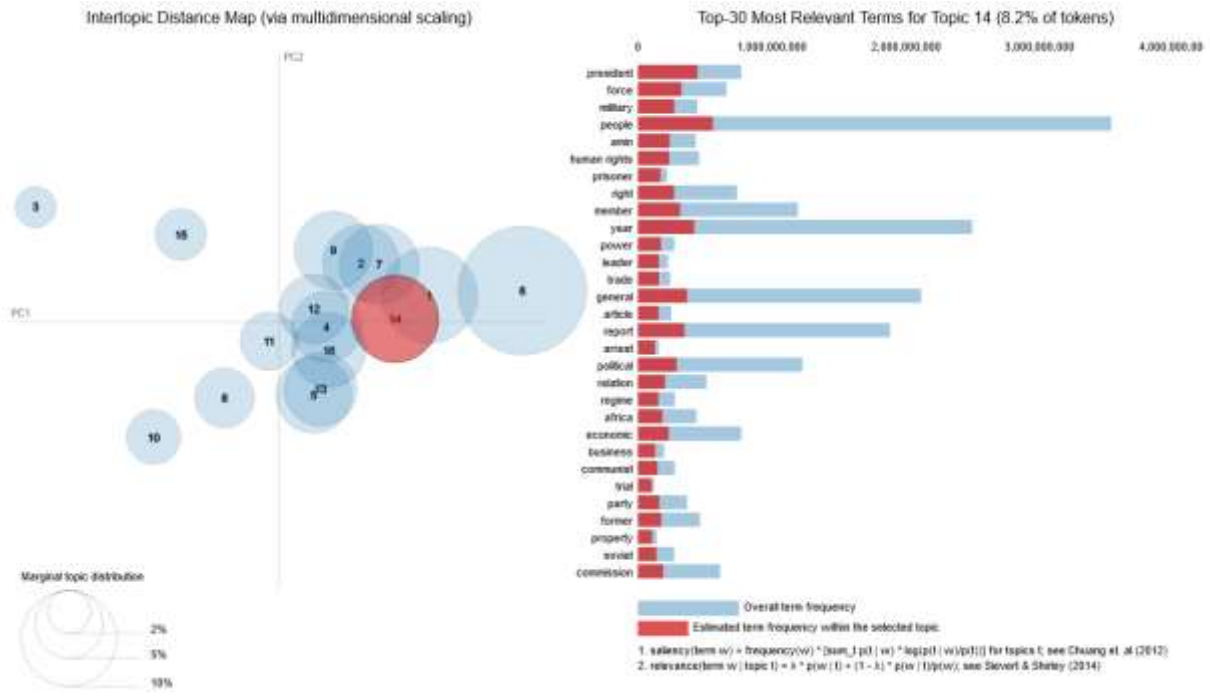

Figure 5: Example of a plot from pyLDAVis. The historian can explore this interactively, selecting different topics to investigate, and adjusting the λ parameter in the formula for topic relevance (which determines the rank order of terms in the right-hand chart).

Second, the topics identified by the model were summarized by their most prominent terms and each assigned by the historian to a case study and a 'theme' (based on the most prevalent terms). In order to be useful, the model should demonstrate (a) that it identifies information available to the historian at the time of the data collection, such as the three different case studies and (b) that it also produces topics that generate new information, such as encompassing themes that occur in several or all topics. Ideally, there should be a number of clearly demarcated 'case study related topics' as well as 'overarching topics'. In the iteration with 8 topics, the case studies did not clearly appear, as lemmas connected with each case study appeared in almost every topic, which inhibited interpretation by the historian. 20 topics produced topics which were clearly demarcated per case study, but did not produce more overarching topics than in the run with 16 topics. In the end, the model with 16 topics proved to be the most interpretable.

The results of the model eventually chosen (LDA with 16 topics) were then presented in human-readable form in several additional ways. First, the topic–term matrix was represented as a table showing the most relevant words for each topic. Second, the document–topic matrix was combined with the document metadata to produce graphs showing (a) the prevalence of each topic in each archive, for documents written in each time window (month, quarter or year) of the period under investigation, and (b) the prevalence of each archive (UNHCR, TNA, and AAD) in each topic, for the period as a whole. Further details about this information are provided in the next section. Third, again based on the document–topic matrix and the document metadata, a 'historian's reading list' was produced, listing (per topic) the documents most strongly associated with that topic.

4.3 Topic clustering

Having identified topics, clustering techniques can provide a further level of insight by finding groups of topics which are in some sense similar. Approaches integrating topic modelling and clustering like this are described in more detail by (Ahmadi *et al*., 2018) and show that topic modelling, followed by clustering of the resulting topics, yields more semantically meaningful clusters of documents than bag-of-words clustering (i.e. clustering based directly on the document–term matrix without the use of a topic modelling algorithm).

Generically, a clustering algorithm attempts to group data points into clusters so that each point is similar to all points in its own cluster, and dissimilar from all points in other clusters. Qualitatively speaking, we regard topics as similar if they contain overlapping vocabulary words, with similar weightings. Quantitatively, 'distances' that measure the dissimilarity between topics

were determined by treating each topic as a vector in a multi-dimensional space, where the dimensions were the words in the vocabulary of the documents. Each row in the topic–term matrix was treated as a vector, and the cosine distance between each pair of rows was calculated. sklearn's AffinityPropagation model was used to find clusters, with the affinity between each pair of points being the negative of the cosine distance described. The algorithm is appropriate for this situation, where the dataset (number of topics) is small, and we do not want to presuppose the number of clusters to be found, or that the clusters should be similar in size. As with the topic modelling, the clustering algorithm did not have access to any of the document metadata: it could not use document date, source or case study as an input to its decisions.

5 Preliminary results

First, we use the topics automatically extracted from the documents to identify the main themes discussed in the documents and compare them with the results of a qualitative analysis by a close-reading historian. Second, we approach the topics and the themes from a diachronic point of view and investigate how the themes are represented across different archives during the 1970s. We use the word 'theme' here when referring to the historian's manual classification of the topics, and 'cluster' when referring to the results of the automatic clustering algorithm.

5.1 Main themes in the documents

From the iterative approach described in section 4.2, the LDA model with 16 topics gave the most interpretable results. This allows a historian to have a quick visualization of the content and classify the most historically significant topics above the smaller topics. Moreover, the model can also list the most relevant document for each topic, which can point the historian directly to candidates for close reading, as shown in Table 2.

		Most prominent lemmas by rank									
Topic	**Theme (manually assigned)**	**1**		**2**		**3**		**4**		**5**	
1	UK Policy	britain	(1.09%)	home	(0.87%)	country	(0.71%)	case	(0.71%)	office	(0.66%)
2	Ugandan Asians	uganda	(2.37%)	britain	(1.64%)	asians	(1.19%)	ugandan	(1.16%)	india	(1.08%)
3	UNHCR	les	(2.34%)	nations	(1.96%)	commissariat	(1.51%)	capital	(1.35%)	letter	(1.29%)
4	Indochina	boat	(1.99%)	singapore	(1.97%)	ship	(1.86%)	vessel	(1.26%)	arrive	(0.96%)
5	USA Policy	usa	(1.85%)	case	(1.82%)	ins	(1.17%)	number	(1.16%)	bangkok	(1.13%)
6	Indochina	country	(1.93%)	government	(1.15%)	problem	(1.02%)	resettlement	(0.96%)	number	(0.68%)
7	Indochina	hong kong	(1.71%)	minister	(1.12%)	britain	(1.02%)	secretary	(1.02%)	government	(1.01%)
8	Indochina	limited	(4.19%)	official	(4.17%)	embassy	(2.17%)	saigon	(1.40%)	reftel	(1.28%)
9	Ugandan Asians	person	(1.41%)	uganda	(1.34%)	high	(1.20%)	office	(0.97%)	family	(0.90%)
10	Indochina	nguyen	(2.54%)	thi	(2.16%)	van	(1.98%)	family	(1.53%)	tran	(1.23%)
11	Chile	chile	(6.38%)	santiago	(3.21%)	embassy	(1.53%)	case	(1.45%)	official	(1.21%)
12	Indochina	hong kong	(7.62%)	office	(1.09%)	ukmis	(1.03%)	number	(0.99%)	government	(0.98%)
13	Indochina	bangkok	(1.72%)	page	(1.57%)	srv	(0.92%)	hanoi	(0.90%)	chinese	(0.79%)
14	Ugandan Asians	people	(0.74%)	president	(0.59%)	year	(0.55%)	government	(0.55%)	general	(0.48%)
15	UNHCR	jaeger	(1.65%)	hour	(1.40%)	indicate	(1.28%)	paragraph	(1.17%)	kelly	(1.15%)
16	ICRC	conference	(1.76%)	icrc	(1.76%)	prg	(1.61%)	usmission	(1.32%)	usun	(1.22%)

Table 2: The five most prominent lemmas in each topic. Lemmas within a topic are ranked 1-5 in descending significance, but the numbering of topics 1-16 has no significance.

A first observation reveals that – even though the model did not have access to information on the three case studies – several, but not all, of the topics identified by the model align well to them. Topics 2 and 9 (and to a lesser extent 14) clearly deal with the Ugandan Asians, topic 11 with Chile, while topics 4, 5, 7, 8, 10, 12 and 13 handle various aspects of the Indochinese crisis across South-East Asia. The remaining topics overarch the case studies (as we have seen in section 4.2). Upon closer inspection, they mainly reveal information about the organizations and governments under scrutiny and their policies with regard to refugee resettlement. Topic 1 concentrates on the British Government and topic 5 (which does not only contain words on Indochina as displayed above, but also on Latin America) on the United States. Topics 3 and 15 are not immediately recognizable for the untrained eye, but they contain words related to processes and actors within UNHCR. Topic 16 is an anomaly: it contains information on the Diplomatic Conference on the Reaffirmation and Development of International Humanitarian Law Applicable in Armed Conflicts, which started in Geneva in 1974 (Kalshoven, 2007), that is unrelated to the refugee policy from the research question. These records were added by mistake and the model has rightly isolated them from all the rest[6]. From this, it appears that the model

[6] Based on the combination of the AAD tags for South Vietnam and the ICRC, which were both related to refugees and the ICRC Conference of 1974 and therefore hard to isolate based on tags only.

generates meaningful results – or in other words, that the OCR input is sufficiently readable to discern the content of the research question.

In order to investigate how closely the identified topics did or did not align to the division of the dataset into its three source archives, we defined a "relative prominence" score for each topic in each archive. High relative prominence for topic T in archive A means that T is highly prevalent in archive A, relative to the other archives. For example, the relative prominence of topic 1 in TNA is defined as:

$$w_1(\text{TNA}) / [w_1(\text{TNA}) + w_1(\text{UNHCR}) + w_1(\text{AAD})]$$

where $w_T(A)$ is the mean weight of topic T over all the documents in archive A. Table 3 shows, for each topic, the relative prominence of the topic in the three archives.

		Relative prominence (%)		
Topic	**Theme (manually assigned)**	**AAD**	**TNA**	**UNHCR**
1	UK Policy	8.2	**70.5**	21.3
2	Ugandan Asians	5.4	**72.5**	22.1
3	UNHCR	2.4	12.5	**85.1**
4	Indochina	32.9	33.3	33.8
5	USA Policy	**86.6**	5.7	7.6
6	Indochina	22.7	**55.4**	21.8
7	Indochina	4.9	**86.0**	9.1
8	Indochina	**88.2**	4.7	7.0
9	Ugandan Asians	2.8	11.7	**85.5**
10	Indochina	**81.5**	7.0	11.5
11	Chile	62.7	15.9	21.4
12	Indochina	6.8	**90.1**	3.1
13	Indochina	**75.4**	12.8	11.9
14	Ugandan Asians	27.3	39.5	33.2
15	UNHCR	1.4	1.8	**96.8**
16	ICRC	**88.3**	6.4	5.2

Table 3. Relative prominence of each archive in each topic. Scores over 50% are shown in bold.

We found that topics 5, 8, 10, 11, 13, and 16 were relatively more prominent in documents from AAD. Topics 3, 9, 15 occurred mostly in UNHCR documents, and topics 1, 2, 6, 7, and 12 mostly in TNA. Topics 4 and 14 did not show a strong association with a single archive.

5.2 Results of clustering analysis

The clustering analysis described in section 4.3 identified three clusters (shown in Table 4). The first cluster A includes topics 1, 2, 6, 7, 9, 11, and 14. The second cluster B only contains topic 3. The third cluster C groups topics 4, 5, 8, 10, 12, 13, 15, and 16. At first sight, the clusters do not exactly match the case studies. Themes related to Uganda and Chile are found together in cluster A, but Indochina is scattered over the clusters. To test whether the clusters follow along the lines of the archives, the relative prominence calculations described in section 4.1 was repeated using clusters in place of topics. The relative prominence of each cluster for each archive is also indicated in Table 4.

			Relative prominence (%)		
Cluster (assigned by model)	**Topics (with manually assigned themes)**		**AAD**	**TNA**	**UNHCR**
A	1	UK Policy	15.1	**49.3**	35.6
	2	Ugandan Asians			
	6	Indochina			
	7	Indochina			
	9	Ugandan Asians			
	11	Chile			
	14	Ugandan Asians			
B	3	UNHCR	2.4	12.5	**85.1**
C	4	Indochina	**53.0**	20.7	26.3
	5	USA Policy			
	8	Indochina			
	10	Indochina			
	12	Indochina			
	13	Indochina			
	15	UNHCR			
	16	ICRC			

Table 4: Assignment of topics to clusters by the model, and relative prominence scores per cluster.

While all clusters extract information from all archives, cluster A derives most information from TNA and cluster C from AAD. Cluster B is heavily dominated by the UNHCR archives, but contains only a single topic, which contains mainly pro forma text. When looking at the word level however, an interesting ordering principle appears. Cluster A concentrates around the following terms: 'government', 'country', 'Britain', 'Uganda', and 'people'. In terms of actors, Britain is the most represented in cluster A. Cluster C focuses on 'Bangkok', 'Hong Kong', 'page', 'official', and 'USA'. In terms of actors, USA is the most represented in cluster C. Cluster A concentrates on government policies of resettlement countries (mostly Britain, but also USA), while cluster C collects the information gathered by the embassies of the USA and (to a lesser extent) Britain in the countries of first asylum (first and foremost Thailand, but also Hong Kong, Singapore and Malaysia). Based on the prominence of the archives and the content of the clusters, we can conclude that cluster A is more concerned with British refugee policy while cluster C focuses on American refugee resettlement strategies.

5.3 The evolution of themes and clusters across archives and across time

The second research question relates to the horizontal and vertical transmission of topics across documents and years, which is visualized in Figure 6. A first finding is that the model is highly accurate for all three archives and across time as far as its outcome corresponds to classical (analogue) historical scholarship. Chronologically, the Ugandan case occurred first. In 1972, President Idi Amin expelled Asians of Indian origin, who were UK passport holders. Figure 6 shows that the event, at the time of its happening, was relatively most important in the UK archives, but soon became mainly the concern of UNHCR. Of a total of 50,000 expellees, 27,000 first found asylum in the UK (Tandon and Raphael 1978: 17). The UK denied further responsibility for the rest of the Asians, who were eventually resettled as refugees in more than 25 countries (Mamdani, 2011). By 1974, most Ugandan Asian refugees were resettled, but UNHCR remained occupied with specific assistance resettled Asians, particularly family reunification, until the end of the decade (Cosemans, 2018). Topic 9, with 'family' as one of its core words (see table 2), demonstrates how this practice developed over time.

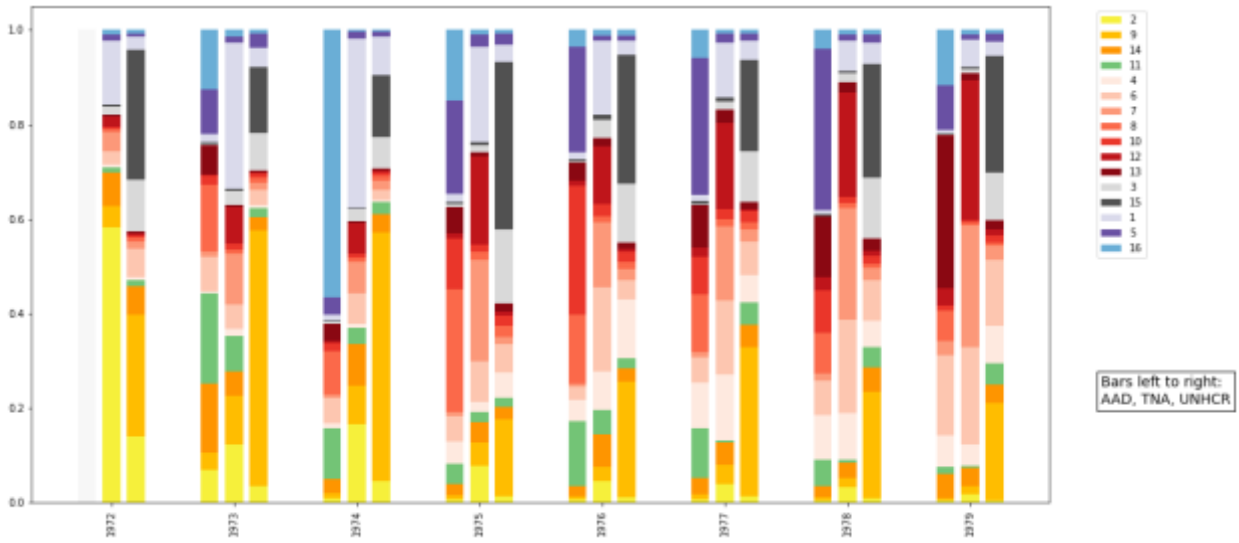

Figure 6. Visualization of the topic variation by themes in UNHCR, TNA, and AAD. The x-axis indicates the time and the organizations. The y-axis refers to the ratio of topics. The main colour palettes refer to the overarching themes identified in the qualitative analysis: *Ugandan Asians* is orange, *Chileans* is green, *Indochina* is red, *UNHCR* is grey, *overarching topics* are purple. Topic 16, relating primarily to documents that were included by mistake, is blue.

The Chilean case appeared significantly for the first time in 1973, at the time of the coup by General Augusto Pinochet. Interestingly, the case was of interest in the American and British archives, while it does not leave a significant mark in the UNHCR archives. The explanation could be that UNHCR in first instance became concerned with the Latin American refugees that were recognised under its mandate in Chile, but who were not Chileans (Rojas Mira and Santoni, 2013; Smith, 2013). UNHCR became responsible for Chileans only at the end of 1976. A specific focus on Chileans (not refugees coming out of Chile) only appears in the UNHCR archives after that year (Vera E., 1986). In 1974, when the issue was discussed by the British Labour Party and subsequently a group was admitted to the UK, the Chilean case gained prominence in TNA (Livingstone, 2018), but overall attention for Chileans was limited. The brief spike in interest in Chilean refugees in the USA in 1976 and 1977 can be explained by the murder on Orlanda Letelier by Pinochet's secret police in Washington DC in 1976 and subsequent congressional debates about raising quota for Chilean refugees (Walker, 2011).

The fall of Saigon in April 1975, is clearly visible in figure 6, as the prominence of the 'red' topics (related to Indochina) increases. In AAD, we see how in 1975 topic 8 about Saigon is the most important, to make room for personal files and family reunion cases in topic 10, and eventually Chinese refugees in topic 13. This is a very consistent picture of what we know about the American involvement in the Indochinese refugee crisis (Robinson, 1999) In TNA, the topics about Hong Kong (12 and 7) only increase over time, which coincides with greater refugee arrives there and rising tension between the British government and its colony (Chan, 1990). In 1979, Margaret Thatcher called for an international conference on the Indochinese refugees, particularly to relieve Hong Kong's burden (Kumin, 2008). UNHCR first occupied itself with the boat people themselves (topic 4) and then with the resettlement of them (topic 6).

While a careful study of the 'case study related topics' can teach us a lot about change over time, the 'overarching topics' show certain continuities between the case studies. Topic 1 is clearly about British refugee policy. Interestingly, it does not only feature in TNA sources, but also in UNHCR ones (not in AAD). Further investigations following from this observation have indicated that Britain indeed had an important influence on the international refugee organization – much more important than American policy (contrary to what is sometimes posited in the literature). Topic 5 on American policy does not feature as much in the UNHCR records. This topic only spans two case studies – the Chileans and the Indochinese – because there are hardly any data on the Ugandan Asians in AAD (due to the fact that the data start a year after the occurrence of the Ugandan Asian crisis) and it shows that only in Indochina did the American policy on refugees really matter. As said previously, topic 3 contains mostly pro forma text from the UNHCR archives, but topic 15 shows something revealing. Among its 5 most prominent words are the names of two important UNHCR staff members (other than the High Commissioner): Gilbert Jaeger and John Kelly. The fact that they remain in function throughout the period and their potential impact on refugee policy merits further research.

The same diachronic analysis was conducted for the three clusters identified by the clustering algorithm (Figure 7). Broadly, cluster A (government policy and human rights, in blue) is more prominent earlier in the decade. In the later 1970s, those topics gradually decreased as cluster C (findings from countries of first asylum and UNCLAS, in green) grew. Notice again how the anomalous topic 16 is overrepresented in 1974, therefore distorting the results.

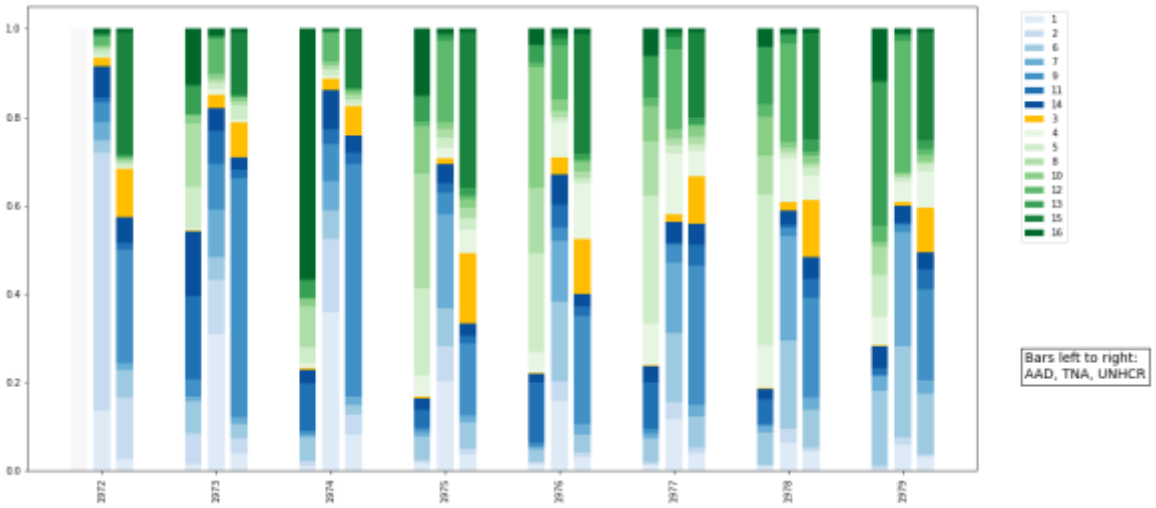

Figure 7. Visualization of the topic variation by automatic clusters in UNHCR, TNA, and AAD. The x-axis indicates the time and the organizations. The y-axis refers to the ratio of topics. The main colour palettes refer to the clusters identified in the quantitative analysis: *Cluster A* is blue, *Cluster B* is yellow, and *Cluster C* is green.

This is intuitive from a historical point of view, as from 1975 onwards, the Indochinese case became much more prominent, which is reflected in the growth of cluster C. It is also interesting to see that while the clusters are shared across archives, topics related to cluster C are more prominent in TNA and topics representing cluster A are mostly found in AAD. This also reflects the involvement of different organizations in different topics.

5.4 Pointing the historian in new directions

Now that we have established that NLP methods can be used to identify what we already know through 'traditional' historical methods, even with imperfect OCRed data, we need to investigate whether it can also generate new results, or at least point the historian in new directions quickly.

Historians come to their research with their own preconceived notions and biases towards history, which can blind them to other results and outcomes than the ones they had in mind. We will highlight how the approach described above shattered these conceptions by presenting a brief case study.

Delving deeper into the divisions which ran through the clusters, we discovered that cluster A contains all topics that include the lemmas 'human rights' (namely topic 14, 6, 11, 1 – in order of prevalence, see Appendix 3), whilst cluster C contains all topics (5, 10, 8 and 16) that include the word UNCLAS (the UN Convention on the Law of the Seas). Having previously ignored all international legal principles other than refugee law, the prevalence of these lemmas within the clusters came as a surprise to our historian. Based on an automatically generated 'reading list' of the 30 most important records per topic (see Figure 8) the issue of human rights was investigated first as it proved to be a particularly productive alley for close reading research.

```
--------------------------------------------------------------------------------
Most relevant docs for topic 11:
AAD 96836: 1976-04-29: CODEL MOFFETT - REPRISALS
AAD 96818: 1976-04-28: CODEL MOFFETT- REPRISALS
AAD 80511: 1974-08-17: POSSIBLE RELEASE OF SOME CHILEAN DETAINEES
AAD 92757: 1976-02-04: PEACE COMMITTEE LAWYERS RELEASED
AAD 92896: 1976-03-03: REFUGEE AND MIGRATION AFFAIRS: PAROLE OF CHILEAN DETAINEES AND  REFUGEES (CORTES DEL CAMPO, AGUSTIN) S-51
AAD 52802: 1975-09-16: PROSPECTIVE CHILEAN PAROLEE: JORGE ALFREDO MONTESI FRAUS.
AAD 93328: 1976-05-05: REFUGEE AND MIGRATION AFFAIRS: PAROLE OF CHILEAN DETAINEES AND  REFUGEES (FRANCISCO HAROLD PARRA GONZALEZ) S-227
AAD 93322: 1976-05-05: OVIP - SECRETARY SIMON VISIT TO CHILE: POLITICAL PRISONERS
AAD 92883: 1976-03-01: REFUGEE AND MIGRATION AFFAIRS: PAROLE OF CHILEAN DETAINEES AND  REFUGEES
AAD 93335: 1976-05-06: HUMAN RIGHTS IN CHILE: ICRC COMMENTS
AAD 96797: 1976-04-27: CODEL MOFFETT AFTERMATH: POSSIBLE EXPULSION OF JOSE FERNANDO ZALAQUETT  DABER
```

Figure 8: Extract of 'historian's reading list' showing the documents most strongly associated with each topic

It soon became apparent that in the early 1970s one particular human right, namely the right to freedom of movement, became an essential feature of refugee resettlement. In the case of the Ugandan Asians, who were British passport holders, their right to move to the UK was removed by British immigration legislation from the late 1960s. At the time of their expulsion, human right advocates focused predominantly on their right to move to 'their own country' (i.e. the country that had issued their passports). Pinochet's Chile has become infamous for its human rights abuses, but historians have always focused on torture (Kelly, 2013). They have overlooked that human rights debates about Chile first focused on liberty of movement - first on the right for the Latin Americans who had been recognized as refugees by the former regime under Salvador Allende to leave the country, which was quickly expanded to Chilean dissidents as well. The developments in Chile marked a shift in thinking about human rights.

By the end of the decade, in 1979, UNHCR negotiated a deal with the government of Vietnam to stem the exodus of Vietnamese on small boats. The provision was that the Vietnamese government would prevent people to leave, in exchange for resettlement places in

the West. UNHCR staff members at the time felt highly uncomfortable about the deal, because it essentially removed the Vietnamese's freedom of movement (Kumin, 2008). However, with thousands at peril on the high seas, other principles prevailed.

The 30 most prevalent records in topic 14, on the Ugandan Asians, provided the first insights. Topic 6, on both resettlement policy and human rights of the Vietnamese, helped to understand the existing literature, which hinted at, but never explicitly referred to, the link between freedom of movement and resettlement. Topic 11 on Chile indicated how in 1973 freedom of movement was still an important human rights principle, but lost its prevalence during the Pinochet dictatorship. Last but not least, topic 1 indicated that human rights considerations were important for UK policy. British anxiety about freedom of movement, from their attempts to divert their responsibility for Ugandan Asians towards the international community to Margaret Thatcher's call for an international conference to control the Vietnamese exodus (particularly to Hong Kong, then still a British colony), informed resettlement practices throughout the 1970s. This insight helps to decentralize dominant narratives about refugee resettlement, that focus on the role of the Cold War and American hegemony (Suhrke, 1998; Bessa, 2009; Hashimoto, 2018)

6 Conclusion

In terms of academic contribution, the three objectives of our study were to (1) identify the main topics in the dataset; (2) investigate the transmission of topics horizontally (between organizations) and vertically (through time); and (3) suggest targeted areas of the document set for further close reading by historians. We consider that the first objective has been completed. We found that topic modelling methods not only detected expected historical events, but also pointed out additional information to deepen the analysis in topics that 'cut across' the case studies, which points *inter alia* to the relevance of human rights debates for refugee resettlement in the 1970s. Likewise, our second objective was also reached, as the extracted topics were tracked in time (within the 1970s) and across the three international organizations involved in this study (AAD, TNA, and UNHCR). The graphs picturing the topics and clusters (figure 6 and 7) clearly show how topics and clusters progressed over time, and which topics retain their prominence throughout the period. Visualization is a powerful tool to enhance interpretation. Our decision to organize the bars per case study also aided analysis. With regard to the third objective, the automatically generated reading list per topic helped to quickly point at the pages most worthy of close-reading research in a corpus of over 20,000 records and over 55,000 pages - a task physically unattainable for a single historian. We therefore consider that our third objective was also attained.

In terms of methodology, we have demonstrated that methods of document digitization combined with NLP models can be of great support in historical research. We used already implemented tools that in most cases were not specifically built for this task (ABBYY for general purpose OCR, object detection for metadata extraction, SpaCy for text parsing and lemmatization, LDA for topic modelling, and affinity propagation for clustering), combined with interdisciplinary communication and iteration. Instead of dealing with approximately 55,000 pages of text, the historian was directed towards the most important records to investigate in order to inform the answer for a particular question. The insight gained from this small sample of pages was then extrapolated: it helped to formulate new search terms for full text search. Interpretation remained a human job, but searching for what deserved human attention was supported by the computer. During this process, it is of the utmost importance to highlight that NLP approaches such as used in the current study are only considered as a supplement and support in hypothesis-testing and/or data exploration. The backbone of the research is (and should always be) a solid theoretical underpinning and a thorough understanding of the data and its socio-historical-cultural context.

Finally, we would like to point out a few limitations in the current study and point towards future research. First, we only considered a few specific tools for digitization of documents. The parameters of these tools were not further tuned either. More state-of-the-art OCR methods that deal better with historical documents (Lu and Dooms, 2019) could be tested to assess if OCR performance could be improved. Second, we needed to work iteratively to explicitly exclude frequent words which were not relevant to our analysis but were not present in NLTK's list of stopwords, as well as words that were frequent artefacts of OCR. Whilst, after having done this, we do not expect that any remaining such words had a significant impact on the analysis, it would be useful to explore less *ad hoc* ways of doing this, including the assessment of different NLP tools. Third, the optimum number of topics in the LDA model was ultimately determined by a person with deep contextual knowledge reviewing their content. While we do not envisage a process that does not have a 'human in the loop', it would be interesting to explore how quantitative and automated assessment of topic quality could be brought into the process. One approach to this was our experiment with using the HDP algorithm, which does not require the number of topics as an input. This gave a large number of topics that were not clearly interpretable: a future experiment, however, could be performing clustering analysis on those results, to review whether the clusters (as opposed to individual topics) are interpretable. Another approach is to use mathematical measures of topic quality. Such measures have been proposed, but it is not fully clear how closely their results correspond to what actually needs to be measured – i.e. the usefulness of the topics to scholarship. For example, van Strien *et al*. (2019) found that calculated topic coherence was lower when the OCR quality of input text was lower, but that there was no clearly correlated decline in human interpretability of topics. Last but not least, a major objective for further research would be to conceive an ideal process and toolkit for conducting research from archive materials to interpretable output. As an example,

the methods we used could be encapsulated in a more user-friendly interface that does not require practical knowledge of programming languages.

References

Ahmadi, P., Gholampour, I., Tabandeh, M., 2018. Cluster-based sparse topical coding for topic mining and document clustering. Adv Data Anal Classif 12, 537–558. https://doi.org/10.1007/s11634-017-0280-3

Allen, D., Connelly, M., 2016. Diplomatic history after the big bang: using computational methods to explore the infinite archive, in: Costigliola, F., Hogan, M.J. (Eds.), Explaining the History of American Foreign Relations. Cambridge University Press, Cambridge, MA, pp. 74–101.

Baierer, K., Dong, R., Neudecker, C., 2019. okralact - a multi-engine Open Source OCR training system, in: Proceedings of the 5th International Workshop on Historical Document Imaging and Processing - HIP '19. Presented at the the 5th International Workshop, ACM Press, Sydney, NSW, Australia, pp. 25–30. https://doi.org/10.1145/3352631.3352638

Barron, A.T.J., Huang, J., Spang, R.L., DeDeo, S., 2018. Individuals, institutions, and innovation in the debates of the French Revolution. Proceedings of the National Academy of Sciences 115, 4607–4612. https://doi.org/10.1073/pnas.1717729115

Bessa, T., 2009. From Political Instrument to Protection Tool? Resettlement of Refugees and North-South Relations. Refuge: Canada's Journal on Refugees 26, 91–100.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022.

Chan, K.B., 1990. Hong Kong's Response to the Vietnamese Refugees: A Study in Humanitarianism, Ambivalence and Hostility. Southeast Asian Journal of Social Science; Singapore 18, 94–110.

Chandelier, M., Steuckardt, A., Mathevet, R., Diwersy, S., Gimenez, O., 2018. Content analysis of newspaper coverage of wolf recolonization in France using structural topic modeling. Biological Conservation 220, 254–261. https://doi.org/10.1016/j.biocon.2018.01.029

Cosemans, S., 2018. The Politics of Dispersal: Turning Ugandan colonial subjects into postcolonial refugees (1967–76). Migration Studies 6, 99–119. https://doi.org/10.1093/migration/mnx024

Gao, Y., Goetz, J., Mazumder, R., Connelly, M., 2017. Mining Events with Declassified Diplomatic Documents. arXiv:1712.07319 [stat].

Hashimoto, N., 2018. Refugee Resettlement as an Alternative to Asylum. Refugee Survey Quarterly 37, 162–186. https://doi.org/10.1093/rsq/hdy004

Kalshoven, F., 2007. Reflections on the law of war: Collected essays, in: The Diplomatic Conference on Reaffirmation and Development of International Humanitarian Law Applicable in Armed Conflicts, Geneva, 1974 - 1977. Brill Nijhoff, Leiden; Boston.

Kelly, P.W., 2013. The 1973 Chilean coup and the origins of transnational human rights activism. Journal of Global History 8, 165–186. https://doi.org/10.1017/S1740022813000090

Kim, S.-H., Na, I.S., Kim, GwangBok, 2014. Text Line Segmentation using AHTC and Watershed Algorithm for Handwritten Document Images. International Journal of Contents 10, 35–40. https://doi.org/10.5392/IJOC.2014.10.3.035

Kumin, J., 2008. Orderly Departure from Vietnam: Cold War Anomaly or Humanitarian Innovation? Refugee Survey Quarterly 27, 104–117. https://doi.org/10.1093/rsq/hdn009

Kurlansky, M., 2017. Paper: paging through history. W. W. Norton & Company, New York.

Livingstone, G., 2018. Britain and the Dictatorships of Argentina and Chile, 1973–82, 1st ed. ed, Security, Conflict and Cooperation in the Contemporary World. Springer International Publishing;Palgrave Macmillan.

Lu, T., Dooms, A., 2019. A Deep Transfer Learning Approach to Document Image Quality Assessment, in: 2019 International Conference on Document Analysis and Recognition (ICDAR). Presented at the 2019 International Conference on Document Analysis and Recognition (ICDAR), IEEE, Sydney, Australia, pp. 1372–1377. https://doi.org/10.1109/ICDAR.2019.00221

Mamdani, M., 2011. From Citizen to Refugee: Uganda Asians Come to Britain, Second Edition. ed. Pambazuka Press, Oxford; Nairobi; Dakar; Cape Town.

Noll, G., van Selm, J., 2003. Rediscovering resettlement (No. Vol. 3), MPI Insight. Migration Policy Institute, Washington, D.C.

Ren, S., He, K., Girshick, R., Sun, J., 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv:1506.01497 [cs].

Riedl, M., Padó, S., 2018. A Named Entity Recognition Shootout for German, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Presented at the Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, pp. 120–125. https://doi.org/10.18653/v1/P18-2020

Risi, J., Sharma, A., Shah, R., Connelly, M., Watts, D.J., 2019. Predicting history. Nat Hum Behav 1–7. https://doi.org/10.1038/s41562-019-0620-8

Robinson, W.C., 1999. Terms of Refuge: The Indochinese Exodus and the International Response. Zed Books, London ; New York : New York.

Roe, G., Gladstone, C., Morrissey, R., 2016. Discourses and Disciplines in the Enlightenment: Topic Modeling the French Encyclopédie. Front. Digit. Humanit. 2. https://doi.org/10.3389/fdigh.2015.00008

Rojas Mira, C., Santoni, A., 2013. Geografía política del exilio chileno: los diferentes rostros de la solidaridad. Perfiles latinoamericanos 21, 123–142.

Romein, C.A., Veldhoen, S., de Gruijter, M., 2020. Dataset: Entangled Histories: Ordinances of the Low Countries.

Smith, Y.E., 2013. Una perspectiva institucional del proceso de asilo para los refugiados y perseguidos políticos en Chile después del Golpe de Estado. Museo de la Memoria y de los Derechos Humanos, Santiago de Chile.

Suhrke, A., 1998. Burden-sharing during Refugee Emergencies: The Logic of Collective versus National Action. J Refug Stud 11, 396–415. https://doi.org/10.1093/jrs/11.4.396

Tangherlini, T.R., Leonard, P., 2013. Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research. Poetics, Topic Models and the Cultural Sciences 41, 725–749. https://doi.org/10.1016/j.poetic.2013.08.002

Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M., 2006. Hierarchical Dirichlet Processes. Journal of the American Statistical Association 101, 1566–1581. https://doi.org/10.1198/016214506000000302

van Strien, D., Beelen, K., Ardanuy, M., Hosseini, K., McGillivray, B., Colavizza, G., 2020. Assessing the Impact of OCR Quality on Downstream NLP Tasks:, in: Proceedings of the 12th International Conference on Agents and Artificial Intelligence. Presented at the Special Session on Artificial Intelligence and Digital Heritage: Challenges and Opportunities, SCITEPRESS - Science and Technology Publications, Valletta, Malta, pp. 484–496. https://doi.org/10.5220/0009169004840496

Vera E., M., 1986. Exilio y Repatriación. Asistencia Social y Laboral, in: FASIC (Ed.), Exilio 1986-1978. Amerinda Ediciones, Santiago de Chile, pp. 81–121.

Walker, V., 2011. At the End of Influence: The Letelier Assassination, Human Rights, and Rethinking Intervention in US-Latin American Relations. Journal of Contemporary History 46, 109–135.

Appendix 1: OCR Quality per Archive and Genre in the Physical Archives[7]

Archive	Genres (manually assigned)	Abbreviation	Pages	Low confidence characters(%)
UNHCR	UNdocuments_english	UR	16	3
UNHCR	statement_english	STA	80	3
UNHCR	meetingnotes_english	MN	147	3
UNHCR	cabletokyo_english	CT	20	5
UNHCR	externalreport_english	ER	444	6
UNHCR	memorandum_english	MEM	893	6
UNHCR	noteforthefile_english	NFTF	244	9
UNHCR	internalreport_english	IR	445	9
UNHCR	english_rest	ER	1571	9
UNHCR	telegram_english	TEL	79	11
UNHCR	pressarticle_english	PA	359	12
TNA	collected_in_2014	TNA1	5114	12
TNA	collected_in_2017	TNA2	16624	10
UNHCR	withthecomplimentof_all	WTC	185	15
UNHCR	internalnote_english	IN	800	18
UNHCR	table_english	TAB	109	19
UNHCR	internalletter_english	IL	181	19

[7] Here we show the values as generated in the ABBYY Finereader 14 Enterprise Hot Folder function. As the OCR-process took place before the sample selection for this paper (based on date, language, and case study), the overall number of pages in this table is higher than the amount used for the NLP analysis.

UNHCR	draftletter_english	DL	41	20
UNHCR	draftmemorandum_english	DM	172	25
UNHCR	incomingcable_english	IC	1913	25
UNHCR	outgoingcable_english	OC	1225	28
UNHCR	draftagreement_english	DA	3	32

Appendix 2: numerical data for Figures 6 and 7

		Topic prominence in documents															
Archive	Year	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
AAD	1973	0.0154	0.0686	0.0017	0.0046	0.0955	0.0742	0.0098	0.1421	0.0381	0.0192	0.1904	0.0015	0.0634	0.1449	0.0050	0.1257
AAD	1974	0.0118	0.0093	0.0030	0.0109	0.0366	0.0541	0.0064	0.0924	0.0099	0.0176	0.1058	0.0029	0.0378	0.0316	0.0032	0.5667
AAD	1975	0.0176	0.0067	0.0053	0.0475	0.1974	0.0546	0.0094	0.2594	0.0089	0.1070	0.0418	0.0113	0.0556	0.0238	0.0040	0.1498
AAD	1976	0.0148	0.0062	0.0022	0.0456	0.2247	0.0273	0.0065	0.1473	0.0082	0.2727	0.1366	0.0096	0.0390	0.0207	0.0033	0.0352
AAD	1977	0.0134	0.0070	0.0022	0.0955	0.2896	0.0534	0.0111	0.1218	0.0106	0.0803	0.1055	0.0186	0.0920	0.0347	0.0048	0.0595
AAD	1978	0.0083	0.0051	0.0018	0.0959	0.3422	0.0722	0.0151	0.0866	0.0055	0.0883	0.0545	0.0279	0.1288	0.0250	0.0030	0.0398
AAD	1979	0.0059	0.0060	0.0007	0.0658	0.0941	0.1696	0.0308	0.0639	0.0041	0.0115	0.0165	0.0362	0.3243	0.0496	0.0035	0.1175
TNA	1972	0.1359	0.5832	0.0186	0.0060	0.0115	0.0281	0.0412	0.0068	0.0445	0.0039	0.0122	0.0201	0.0046	0.0694	0.0035	0.0105
TNA	1973	0.3087	0.1229	0.0293	0.0138	0.0133	0.0527	0.1063	0.0091	0.1028	0.0117	0.0765	0.0785	0.0047	0.0514	0.0046	0.0137
TNA	1974	0.3591	0.1650	0.0255	0.0076	0.0150	0.0655	0.0657	0.0078	0.0829	0.0096	0.0343	0.0640	0.0046	0.0881	0.0025	0.0027
TNA	1975	0.2025	0.0789	0.0141	0.0222	0.0246	0.0858	0.2130	0.0188	0.0492	0.0152	0.0222	0.1851	0.0096	0.0415	0.0066	0.0107
TNA	1976	0.1579	0.0454	0.0384	0.0804	0.0109	0.1797	0.1365	0.0133	0.0303	0.0254	0.0526	0.1224	0.0170	0.0680	0.0108	0.0110
TNA	1977	0.1160	0.0385	0.0165	0.1381	0.0185	0.1570	0.1582	0.0153	0.0416	0.0211	0.0045	0.1808	0.0288	0.0473	0.0093	0.0087
TNA	1978	0.0629	0.0324	0.0200	0.0974	0.0140	0.1986	0.2358	0.0100	0.0199	0.0119	0.0069	0.2231	0.0196	0.0325	0.0054	0.0095
TNA	1979	0.0590	0.0168	0.0093	0.0457	0.0104	0.2055	0.2571	0.0064	0.0182	0.0048	0.0048	0.2963	0.0150	0.0377	0.0034	0.0098
UNHCR	1972	0.0262	0.1403	0.1091	0.0049	0.0070	0.0609	0.0158	0.0074	0.2584	0.0044	0.0138	0.0039	0.0049	0.0591	0.2759	0.0081
UNHCR	1973	0.0398	0.0345	0.0782	0.0075	0.0307	0.0305	0.0168	0.0096	0.5402	0.0079	0.0189	0.0030	0.0049	0.0290	0.1403	0.0081

UNHCR	1974	0.0826	0.0448	0.0668	0.0048	0.0101	0.0207	0.0202	0.0105	0.5258	0.0068	0.0253	0.0053	0.0031	0.0393	0.1303	0.0038
UNHCR	1975	0.0365	0.0134	0.1583	0.0549	0.0254	0.0593	0.0159	0.0227	0.1630	0.0214	0.0180	0.0081	0.0170	0.0268	0.3524	0.0070
UNHCR	1976	0.0299	0.0116	0.1253	0.1248	0.0126	0.0408	0.0231	0.0139	0.2446	0.0213	0.0224	0.0062	0.0129	0.0275	0.2718	0.0112
UNHCR	1977	0.0392	0.0126	0.1073	0.0572	0.0155	0.0705	0.0255	0.0161	0.3154	0.0234	0.0474	0.0050	0.0149	0.0489	0.1929	0.0082
UNHCR	1978	0.0444	0.0087	0.1275	0.0574	0.0178	0.0846	0.0273	0.0080	0.2265	0.0177	0.0427	0.0101	0.0264	0.0503	0.2408	0.0100
UNHCR	1979	0.0311	0.0050	0.1025	0.0807	0.0169	0.1383	0.0299	0.0067	0.2063	0.0147	0.0454	0.0129	0.0190	0.0375	0.2454	0.0078

Appendix 3: Interactive PyLDAVis plot highlighting topics with the lemma 'human rights'