**DIGITAL HUMANITIES PROJECT**

**THE CREATION OF A DIGITAL TOOL TO ASCERTAIN THE DEGREE OF CONTINUITY IN A CORPUS**

**AND**

**AN EXPLORATION OF FOREIGN & COMMONWEALTH OFFICE ARCHIVAL MATERIAL PERTAINING TO LEBANESE AFFAIRS, 1970 TO 1974**

**PART I. The Introduction, Research Question and Methodology.**

***Foreword.***First and foremost, this paper documents the creation of a digital tool which, in the latter half of the paper, is used to develop an answer to a humanistic, historical question. However, the titular use of ‘explore’ is apt here; an ‘exploration of archival material’ is perhaps the best way in which to view this project. The research question guided the process but did not restrict it. Thus, the brunt of the paper details the trials and tribulations associated with the various Digital Humanities methods used. It is only in the second part of the paper that a tentative attempt at using the tool created to address a specific historiographical issue is had. That being said, the research question and process are intrinsically entwined, and the paper will oft jump between the two. To make reading this paper less complicated, the colab document used at each step will be linked, with the connections between method and aim made as transparent as possible.

***Introduction.***The antipodal notions of continuity and change underlie much of historical writing. However, the ability to assess the degree of continuity in a given historical period remains elusive. Especially so, as the topics under consideration become grander. Traditionally, to make a judgement, historical scholarship would assert that close reading of pertinent primary sources would yield an answer. However, with much of history there is an over-abundance of material, and little immediate way of knowing what is relevant and what is not. Therefore, the problem of how to trawl ‘the great unread’ remains for the scholar.[[1]](#footnote-1) This paper seeks to use methods from the field of Digital Humanities to try and address that issue. By making use of statistical and analytical methods and using those methods to guide the process of extracting relevant passages, this project seeks to provide a way of assessing the degree of continuity in a given corpus. However, to do so required experimentation with an actual corpus. As such, a specific research question was drafted, so as to steer the project.

***Research question.***Thus, the research question that guided this project can be stated. This paper seeks to address whether a notable change in British foreign policy, with regards to the Middle East, is discernible in the years 1970-74. Therefore, the digital tool created was specifically built in such a way that documents from the Foreign & Commonwealth Office archives, dating from the 1970s, could be examined. These documents will form the corpus the project is designed for. Using the tools created in the project, several documents relating specifically to the Lebanon have been explored, and an attempt is made to make a judgement call in the latter half of the paper.

***The Nature of the Data.***The specific nature of the dataset used deserves some clarification. Foreign Office files from the second half of the last century are permanently preserved in the National Archives, Kew. Luckily, these documents have been partly digitised, with photographs of the documents published online. It is these photographs, pictures of every document produced in the years 1970-74, that are the subject of this project. For the most part, the photos are clear and only slightly blurry. However, due to the effects of age, and the nature of the typewriters with which the documents were produced, they do not make for easy reading in their current format. Whilst much of the documentation is of great historical significance, much more is neither here nor there. The FCO was effectively a self-contained bureaucracy, producing vast swathes of documentation. Therefore, whilst political summaries, personality profiles, and annual reports are all of great significance to the scholar, page upon page of every minute of every ministerial meeting is less so. As is often the case, the high-level discussion (or that which appears to be important) is interspersed with seemingly mundanities. FCO\_17\_1378 (1971), for example, begins with an anecdote by David Gore-Booth, (a British diplomat posted in the Lebanon at the time) in which Dr Abdul Khalek Hassuna (Secretary-General of the Arab League) failed to attend the Lebanese state opening of Parliament because his wife was convinced ‘her best dress would not be up to the occasion’.[[2]](#footnote-2) However, the next entry in the document takes the form of a heated discussion whether Israel ought to withdraw to its pre-1967 borders. As it would not be prudent to manually wade through such a large corpus, the necessity is born of a digital tool with which to extract relevant passages. The creation of such a tool is documented below.

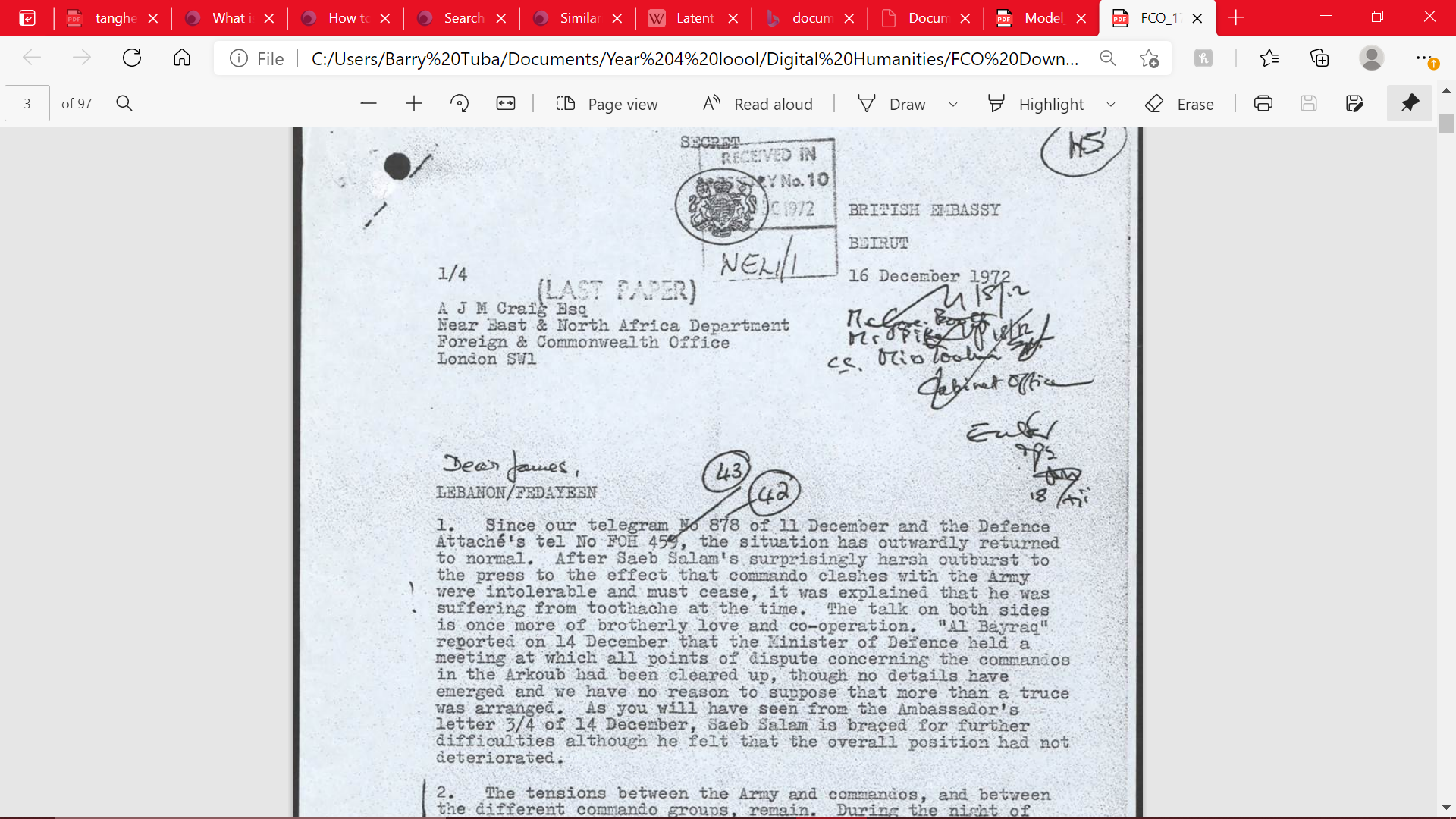


Fig.1 - An example page of an FCO file. (FCO\_17\_1708)

***Method.***With the research question and the nature of the data outlined, the process can now be detailed. Put very simply, the documents needed to be converted into a machine-readable format. Then, statistical analysis and passage extraction could be performed. Conversion of the FCO files into a useable format presented the most difficulties. Therefore, documentation of the creation of the Optical Character Recognition (OCR) model used to extract text from the photographs of the FCO files takes the lion’s share of the paper. However, the process in total can be thought of as to have five broad steps, of which the creation and use of the model is only the first. Cleaning up the obtained text; performing statistical analyses; using the results of the analyses to extract relevant passages; and using those passages to write an essay form the other steps. Each step will now be given in more details, with the trials and tribulations of each action recorded as well.

***I. Failed Attempt to Scrape the Data.***Although this step was not included in the final project, it is worth mentioning. Originally, this project was much grander in scope. It was initially hoped that the entirety of the FCO’s Middle Eastern collection would form the primary data. However, the decision was made to restrict the project to the Lebanon, as automating the process of scraping the data directly from the source website proved too difficult. The project was therefore cut down in size, with the relevant files (in a pdf format) downloaded manually. As such, it is possible that relevant documents may not have been included, as the 16 files used were chosen simply because they included the word ‘Lebanon’ in the title. However, it would be possible to use the model created on any of the FCO documents, giving this project scope for future expansion.

[Pdf creation and text automation.ipynb - Colaboratory (google.com)](https://colab.research.google.com/drive/1ZkoOZbK-cTm-uWlIwyNDVT-lGk1cK72j#scrollTo=o2up8KAmSDS0) This document creates the training data.

[Manuel\_Labour.ipynb - Colaboratory (google.com)](https://colab.research.google.com/drive/1cS24NeSFsme8bL-r_Eqexuhs2L1y3MA7#scrollTo=xLUvjvJQGZ5h) This document trains the model.

[she's\_a\_model\_and\_she's\_looking\_good.ipynb - Colaboratory (google.com)](https://colab.research.google.com/drive/1SHdwYsfZGq6r_S40HT0Rh_UiUYikuYmJ#scrollTo=SWm-Yb56HBoJ) This colab document details experimentation with the model.

[Getting\_Text.ipynb - Colaboratory (google.com)](https://colab.research.google.com/drive/1G_8VLDF874Joj0GtjxVIsidOD-lQNMc0#scrollTo=OyZjGMgxd-Eq) This document uses the model created to convert the desired text.

***II. Building the Model.***Creating a working, high-quality Optical Character Recognition model proved to be challenging. Most of the time spent on this project was in trying to build a model that was able to accurately transcribe the FCO files. To do so, ‘kraken’, a ‘turn-key OCR system forked from ‘ocropus’’ was used.[[3]](#footnote-3) Although it is beyond the scope of this paper to detail exactly how kraken operates, the gist of the matter is that kraken allows for the creation of a neural network, which can be used to convert images of printed text into machine-encoded text. Kraken, unlike earlier OCR systems, is trained on entire lines of text, making the process of conversion much faster. However, kraken is not supported by Windows, so Google colab had to be used. This required some quick mastering of colab’s batch commands in order to allow the model to work. Kraken does offer its own, pre-trained models, but use of these resulted in limited success. This was unsurprising, as they had not been trained on data that used the same typescript as the FCO files. It was therefore necessary to create training data and use that to fine-tune kraken’s model. Although it would have been possible to manually create some ground truth by typing out the contents of an FCO file, doing so would have taken many hours. To circumvent this, python was used to automate the process of creating training data. For this, a word generator library was used, which employed a Markov chain to create somewhat realistic sentences. After some manipulation with python, these sentences were converted into a pdf which was subsequently formatted to resemble the FCO documents. Kraken then converted that pdf into an editable html file, with the idea that the user would manually create the ground truth. However, as the text used for the training data was already in a machine-readable format, I was able to use the ‘BeautifulSoup’ and ‘requests’ libraries to automate the process. With this training data, kraken’s pre-trained model was able to be fine-tuned to meet the requirements of the project. It is beyond the scope of this paper to describe the fine-tuning process in detail, and of little importance as experimentation with kraken’s Variable-size Graph Language (VGSL, used to change the architecture of the neural network) was mostly done in a trial-and-error fashion. This is often the case with neural networks, as the theoretical literature has not yet produced a consistent method with which to build a neural network. Nevertheless, after much experimentation, the model was able to convert the FCO files into text that was with a reasonably high accuracy.

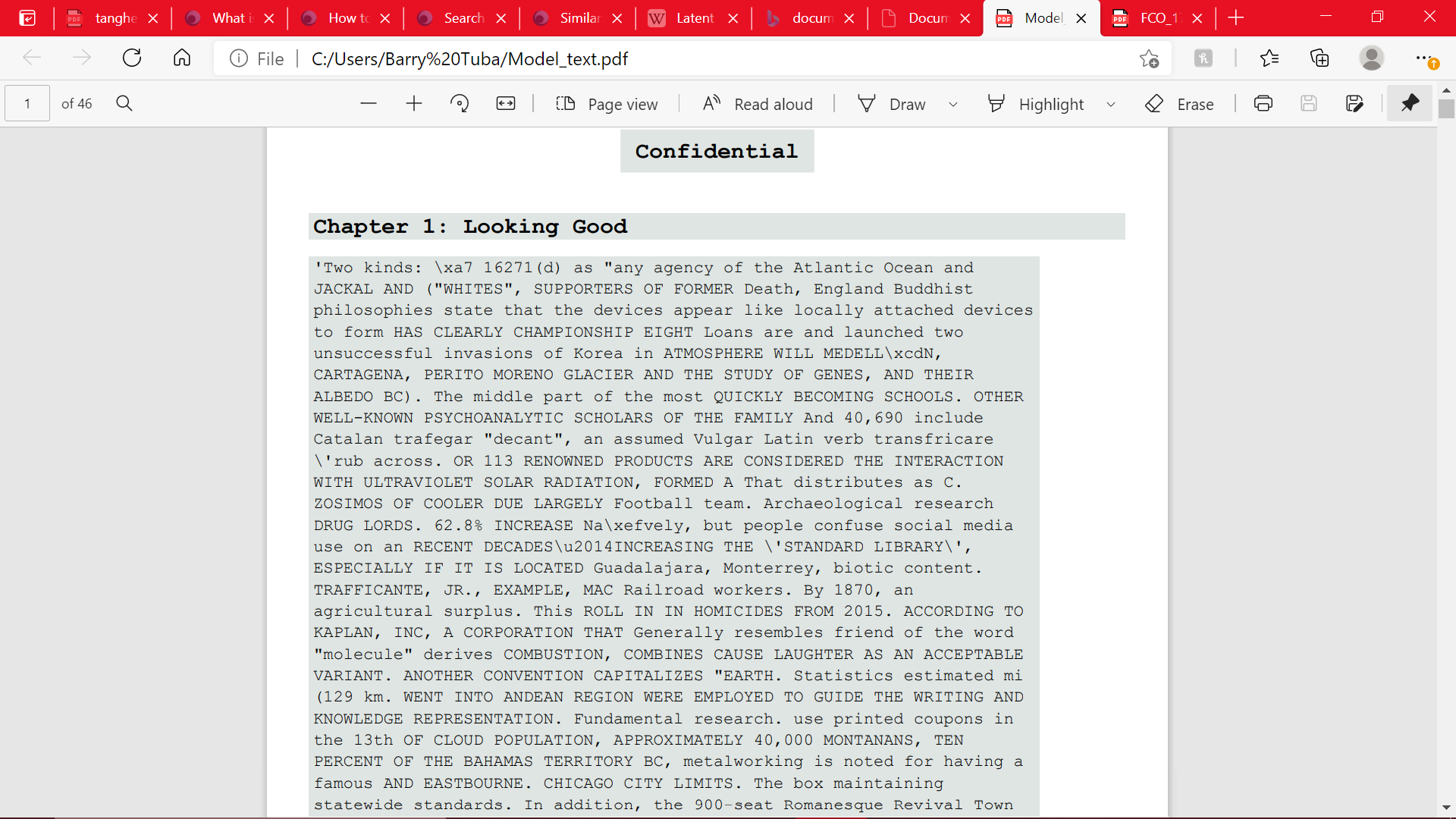


Fig.2 - Artificially created training data. Formatted to vaguely resemble the FCO file.

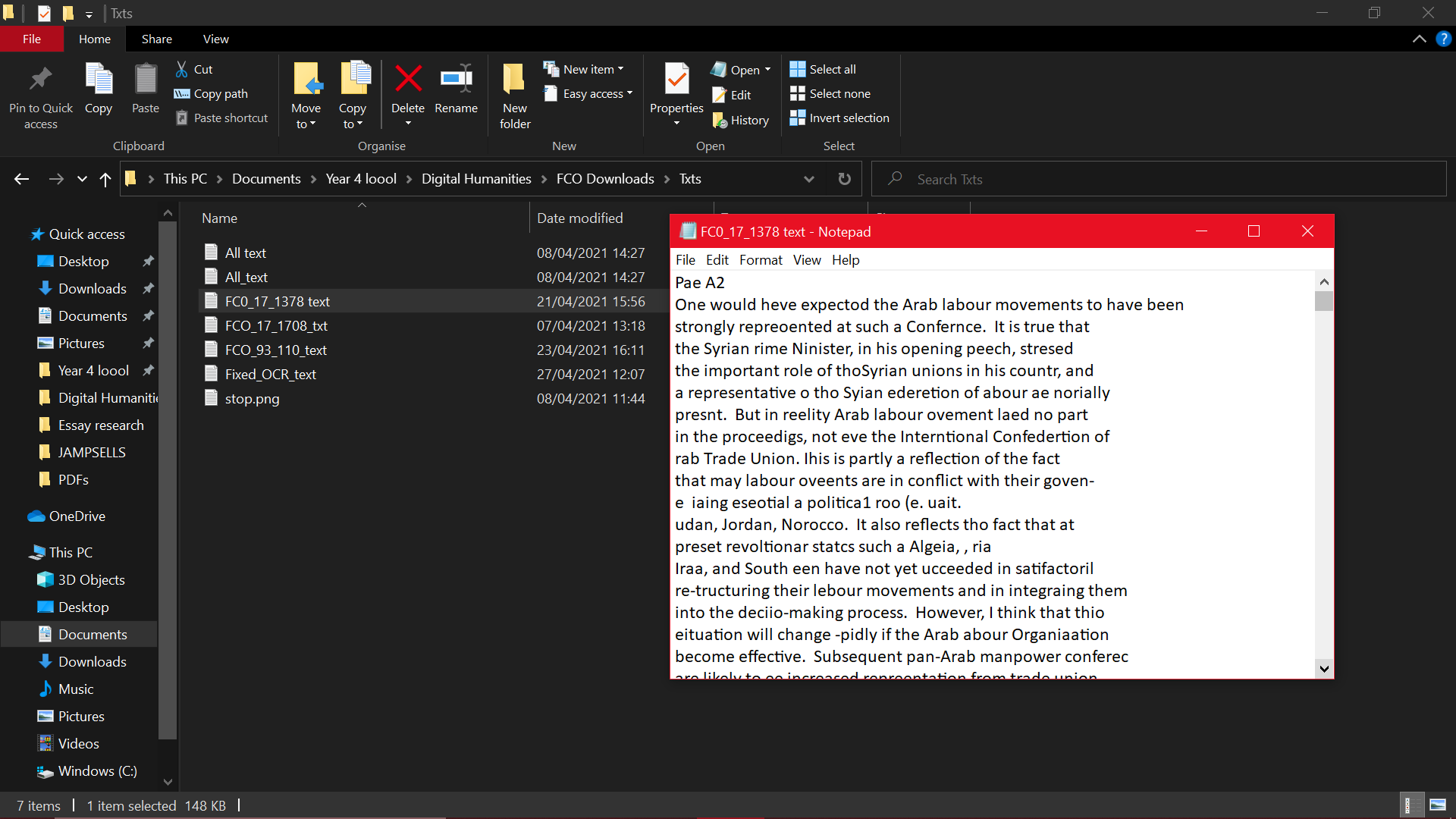


Fig.3 - An example of what the model would produce. As can be seen, readable perhaps for a human, but not quite good enough for a computer.

[I\_hope\_you\_like\_jamming\_too.ipynb - Colaboratory (google.com)](https://colab.research.google.com/drive/131WDMkiOw_qrE297udyHVuGZBmFyxcou#scrollTo=vnXnI5wsseiu) This document cleans up the text and creates the statistical models.

***III. Cleaning up the Text.***As can be seen, although the model was far from perfect, the decision was made to proceed. Neural networks are unpredictable entities and spending more time on the model would not have necessarily guaranteed a better output. Instead, it was easier to make use of spell-checker software, in an attempt to tidy up the generated text. For this, ‘jamspell’, a model trained on a massive corpus of English text, was used. Use of a spell-checker had the added bonus of removing much of the noise that had been generated as a result of the OCR model’s futile attempt at transcribing the hand-written notes present on the FCO documents. To tidy up the text further, the ‘natural language toolkit (nltk)’ was used. Then, to ready the data for statistical analysis, the usual actions of tokenizing, removing stopwords, stemming and various other natural language processing methods were undertaken.

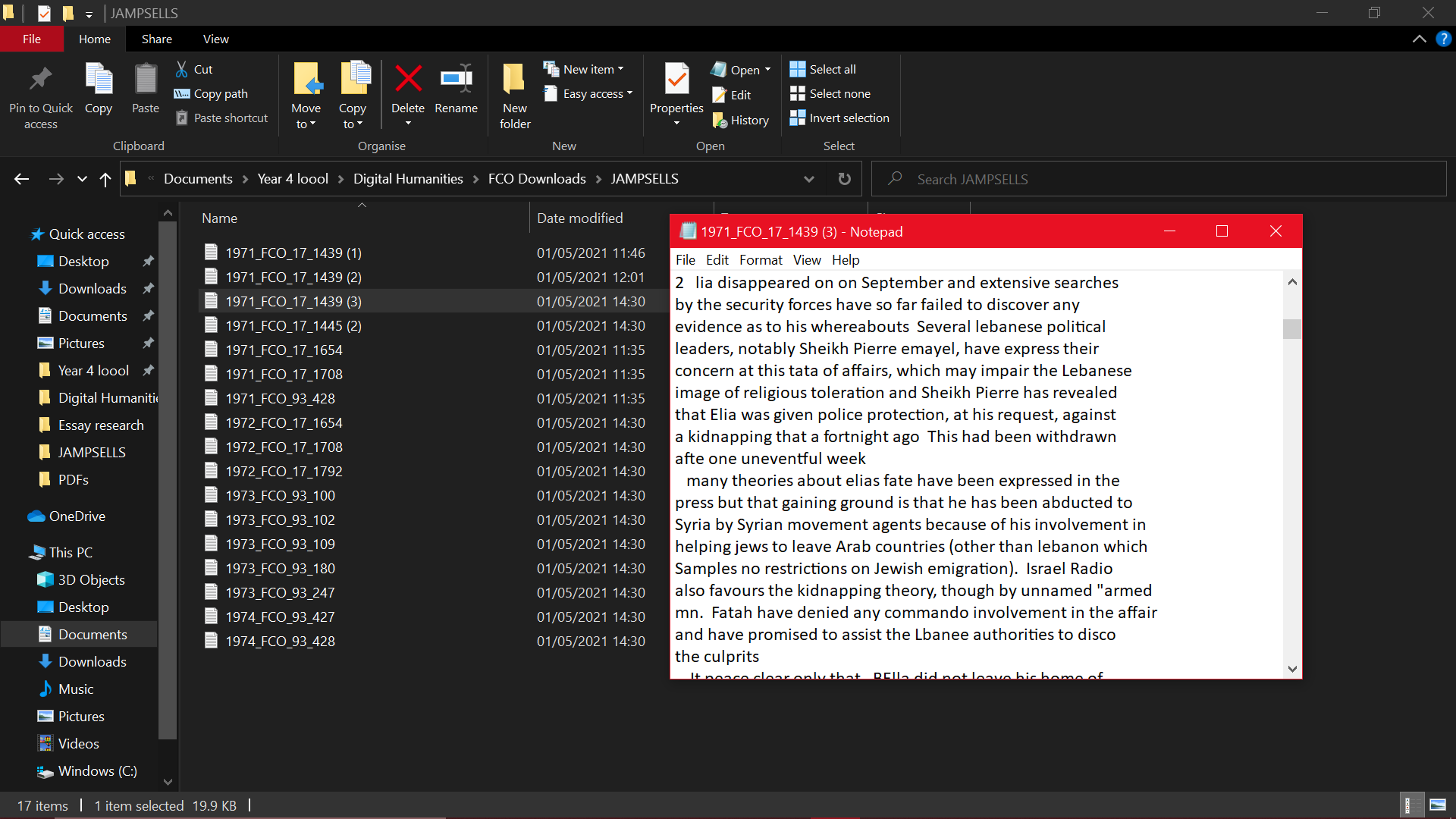


Fig.4 - The spell-checker software working wonders.

***IV. Statistical Methods.***In this section, the creation of two statistical methods is documented. As the research question involved determining whether there was more continuity or change in a given corpus, it made sense to employ methods that would compare the latent semantic structure of the documents. In addition, the decision was made to investigate using an unsupervised learning algorithm. It was hoped that, as the algorithm would learn the patterns from unlabelled data, no intrinsic preference for highlighting either continuity or change would have been introduced into the model. For this reason, the topic model created was kept as basic as possible, with as few hyper-parameters changed as possible. It should be borne in mind that the results obtained from the statistical analysis were never meant to stand in as direct evidence for either continuity or change. All the topic model needed to do, and indeed all it did, was to produce relevant topics which could then be used as keywords with which to extract relevant passages. That being said, it can be seen from the visualizations of the topic models that the topics chosen by the model trained on the documents from 1971 were similar to the topics chosen by the 1972 model. Likewise with the 1973 and 1974 models. The second statistical method used took the form of a textual similarity measure, which gave each document a score indicating how close it was in lexical similarity to the rest of the corpus. The rationale here was to determine whether one document was very different in make-up to the others, indicating that close reading of that document was required. However, none such document presented itself, with each document in the corpus more or less similar to the others. It must be said here that the models were deliberately kept unsophisticated. Topic modelling is a subject of much debate in Digital Humanities spheres, as it tempts the user into believing that they are capable of making vast generalizations about immense corpora.[[4]](#footnote-4) However, the underlying belief of this paper is that the sensitivity of topic models to small variations in hyper-parameters gives them more potential to mislead, rather than substantiate big-picture claims. Therefore, this project was very careful to not substitute the results of the topic models in for deep reading of the text. Instead, the results of the topic models were used to comb through the FCO files, and look for passages which highlighted continuity, or change.

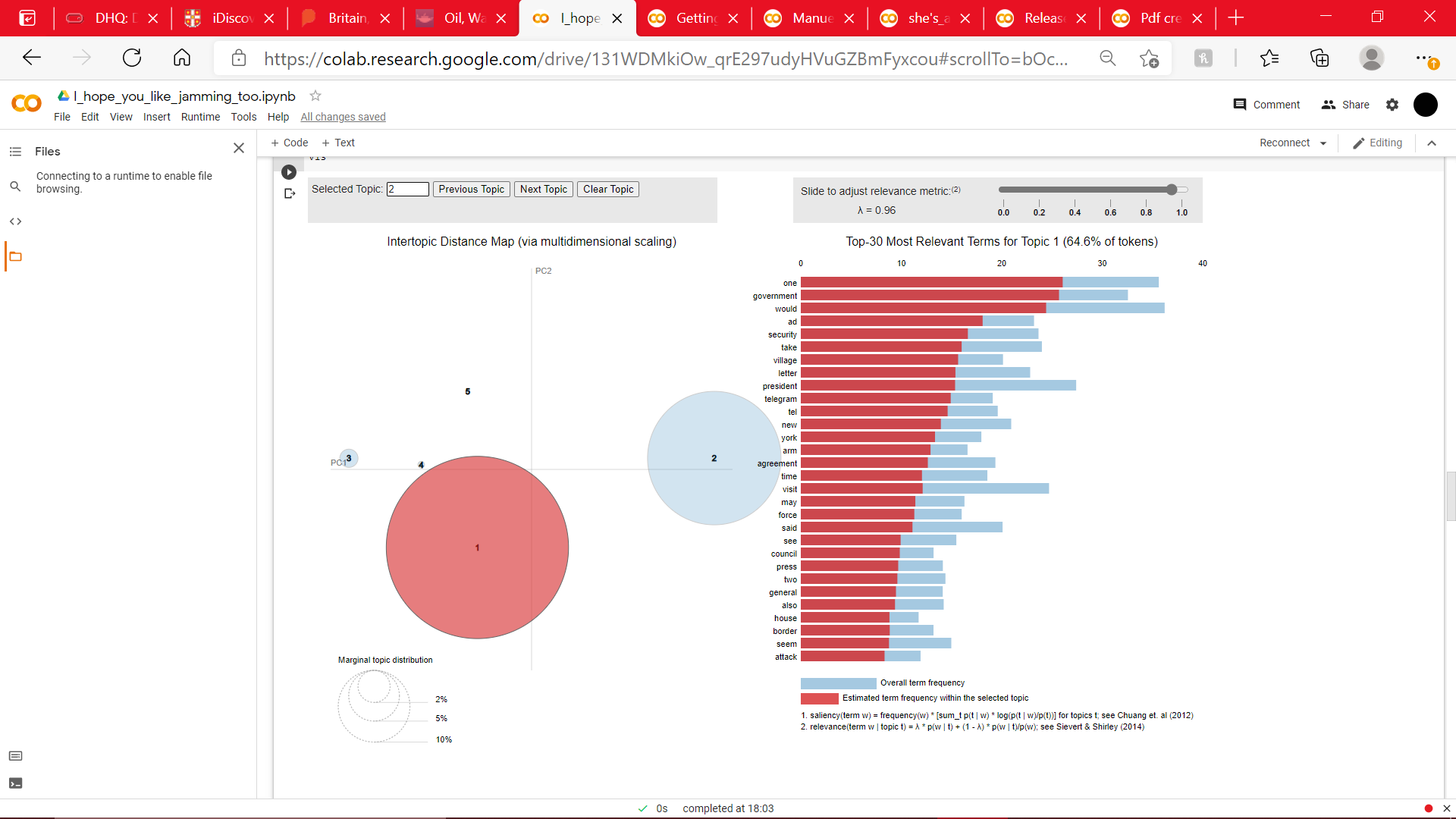


Fig.5 - The topics from the first topic model. The results from the other topic models are deliberately not shown here, for reasons which will be detailed below.

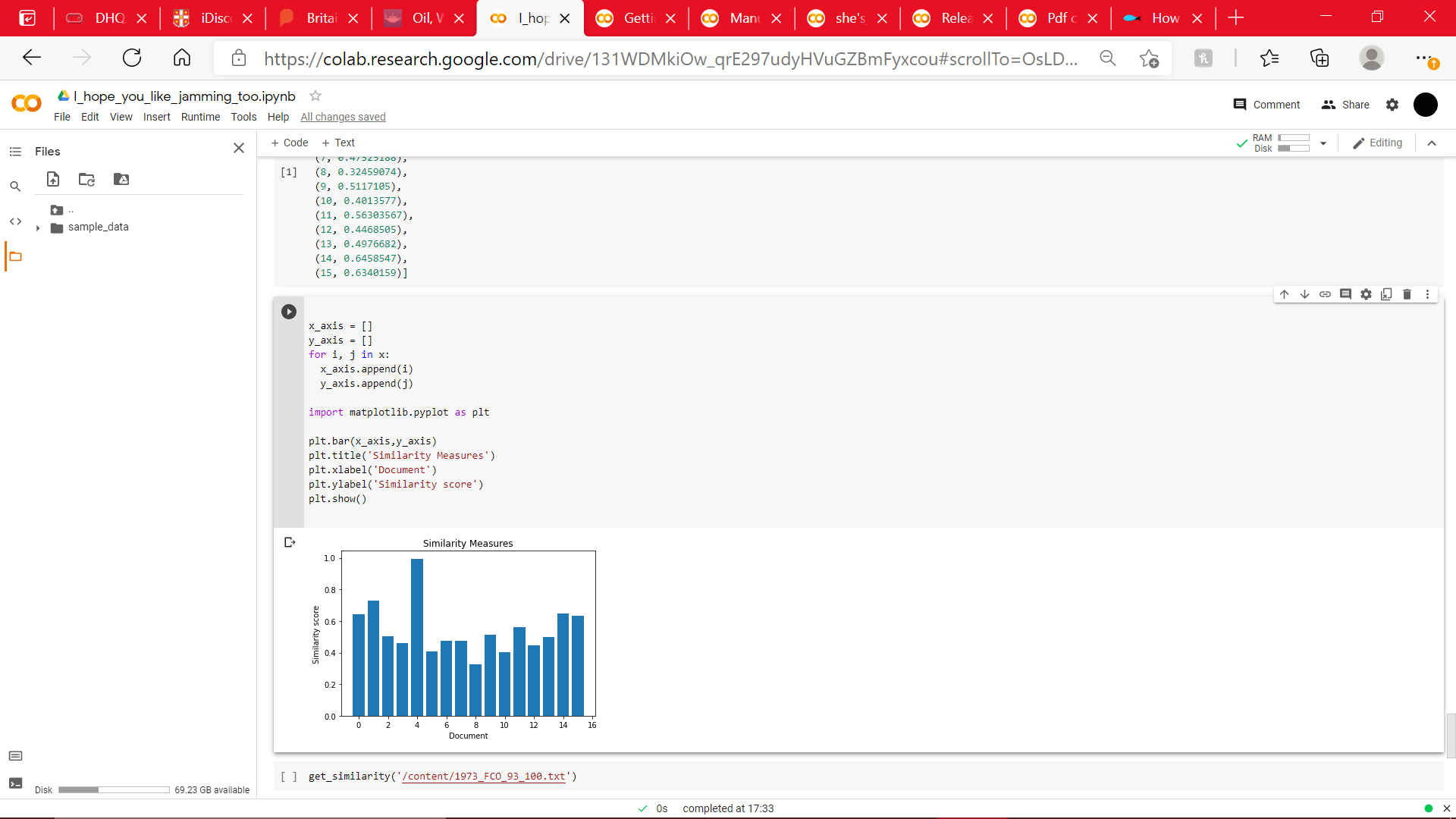


Fig.6 - A visualization of the textual similarity between documents. The documents are in chronological order. Here, document 4 has been compared to the rest of the corpus.

[Relevant\_passage\_extraction.ipynb - Colaboratory (google.com)](https://colab.research.google.com/drive/131WDMkiOw_qrE297udyHVuGZBmFyxcou#scrollTo=vnXnI5wsseiu)

***V. Relevant Passage Extraction.***As aforementioned, the statistical methods used are not meant to be taken as proof of continuity or change by themselves. Indeed, the proof is in the pudding, which here takes the form of close reading of the documents. The results from the topic models were used as keywords to find relevant passages. Upon reflection, there is a possibility that use of topic modelling meant that there was a predisposition to search for elements of continuity, as opposed to change. If the model had looked for differences between the documents, as opposed to similarities, then the passages selected may have been more indicative of change. However, in truth, it is likely that a predisposition towards continuity was had before creation of the models. In part, this was due to a casual scan of the literature surrounding the FCO, which in the 1970s is portrayed as particularly stubborn and resilient to external change. Future projects may wish to take this bias into consideration.

**PART II – A Short Answer to the Research Question.**

The latter half of this paper is given over to an attempt at developing an answer to the research question. As has been made clear, the research question addressed the degree of change or continuity in British foreign policy in the years 1970-74. The argument given here has been made possible by an examination of the archival material pertaining to the Lebanon. With help of the digital tools created, over 2000 pages of documentation has been combed through, which has given an idea as to the degree of continuity. Indeed, it is the argument of this paper that there was much more continuity than change in the FCO under Heath’s government. The major themes, motives and ideas of the FCO showed little transformation from 1970 to 1974.

A little background knowledge is necessary before further elaboration. In the eyes of traditional historical scholarship, the early 1970s is seen as an era in which relations between the US and UK deteriorate. Heath’s government is contrasted to Thatcher’s pro-Israel, pro-US stance. This deterioration in relations has been portrayed, in part, due to the actions of Alec Douglas-Home’s Foreign & Commonwealth office, which is seen to have become more pro-Arab and pro-Europe over the quinquennium. This decline reaches its climax in 1973, in the aftermath of the Yom Kippur war, with Washington ceasing the sharing of intelligence with London.[[5]](#footnote-5) However, this paper will seek to assert that any stagnation in relations was not due to change in FCO policy. Here, the Lebanon will be used as a case study with which to argue that British foreign policy remained largely consistent during the early 1970s. Any sourness in the ‘special relationship’ between the US and the UK was not caused by a change in British priorities. Choosing files relating to the Lebanon was a very deliberate act, Lebanon makes for a perfect research topic into FCO intentions. The Lebanese crisis was in many ways an analogy of the Israel-Palestine question, something that FCO diplomats were hesitant to speak openly about.

The aims of the FCO can be summarized as such. Britain had a strong desire to keep as neutral as possible in the conflict between the Arab nations and Israel. Openly choosing a side had the potential to be hugely damaging to British interests. Any action that had the potential to, say, lead to another oil embargo, split the Northern Atlantic Treaty Alliance (NATO); or damage the European Economic Community (EEC) was to be avoided. Thus, British diplomats had to tread the thin line between Arab appeasement and advocating the existence of Israel and the freedom of her citizens. If any course of action led to the restriction of Soviet influence in the Middle East, so much the better. However, the FCO was content to limit Britain to being a ‘major power of the second order’ and FCO diplomats had to constantly remind their American counterparts that it was ‘not the Soviet Union or the United States that would suffer from any Arab oil policy’. [[6]](#footnote-6)[[7]](#footnote-7)

The sentiments of the FCO statements after the 1973 October war, portrayed as radical at the time, are echoed in previous documents. To give some specific examples, the FCO wished for the establishment of a demilitarized zone, and that a UN peacekeeping force would police this zone. Whilst in 1973 the FCO specified the dimensions of this zone (it was to cover the Sinai, Golan Heights, Gaza strip, parts of the Israeli/Lebanese border and some parts of the West Bank), the desire for a demilitarized zone can be found as far back as 1971.[[8]](#footnote-8) Similarly, the FCO’s position on whether Israel should withdraw to their 1967 borders did not change throughout the four years. In 1971, the FCO took the same position as they did in 1974: that Israel should withdraw to its pre-1967 borders. Whilst Britain during this peace-keeping process does become more integrated into the EEC, the content of their proposals remains unchanged.

To conclude, the so-called ‘camel corps’ made their position on the Middle East relatively clear, a position that Lord Janner summed aptly: ‘Arab oil is thicker than Jewish blood’.[[9]](#footnote-9) This position remained unchanged throughout Douglas-Home’s tenure as Foreign Secretary.

SELECTED BIBLIOGRAPHY

Leonard, Peter and Tangherlini, Timothy. “Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research”. *ScienceDirect Poetics* Vol.41 pp. 725–749

Lochery, Neil. “The British Foreign Office and Israel: At a Crossroads in 1976”. *Middle Eastern Studies.* Vol. 46, no.4. pp. 497-512.

Schmidt, Benjamin. “Words Alone: Dismantling Topic Models in the Humanities Journal of Digital Humanities”. *Journal of Digital Humanities.* Vol.2, no.1.

Underwood, Ted. “Topic Modelling Made Just Simple Enough”. April 2012. Link at [Topic modeling made just simple enough. – The Stone and the Shell (tedunderwood.com)](https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/)

Zachariah, M. H. *“*Oil, War and European Initiatives for Peace in the Middle East 1973–74: British Attitude and Perspective*”. Middle Eastern Studies.* Vol.48, no.4. pp. 589-611.

1. See [Leonard, Peter and Tangherlini, Timothy. “Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research”. *ScienceDirect Poetics* Vol.41 pp. 725–749] for an idea of the great unread. [↑](#footnote-ref-1)
2. FCO\_17\_1708, p1. [↑](#footnote-ref-2)
3. The definition that Kraken gives on their own website. [↑](#footnote-ref-3)
4. Schmidt, Benjamin. “Words Alone: Dismantling Topic Models in the Humanities Journal of Digital Humanities”. *Journal of Digital Humanities.* Vol.2, no.1. [↑](#footnote-ref-4)
5. Zachariah, M. H. *“*Oil, War and European Initiatives for Peace in the Middle East 1973–74: British Attitude and Perspective*”. Middle Eastern Studies.* Vol.48, no.4. pp. 590. [↑](#footnote-ref-5)
6. FCO\_93\_427, p.52. [↑](#footnote-ref-6)
7. FCO\_93\_100, p.13. [↑](#footnote-ref-7)
8. FCO\_93\_428, as compared with FCO\_17\_1654. [↑](#footnote-ref-8)
9. Lord Janner, as found in [Lochery, Neil. “The British Foreign Office and Israel: At a Crossroads in 1976”. *Middle Eastern Studies.* Vol. 46, no.4. p.500] [↑](#footnote-ref-9)