AI Interlude 1

Iteration as Method in Digital History Research

After learning some foundational text-mining techniques, we can begin experimenting with AI-assisted analytical workflows while maintaining critical methodological control. Writing and revising code often involves an iterative process of refining prompts, testing and critically evaluating AI-generated results, and revisiting corpus context through close reading to refine our interpretation. In this sense, iteration emphasizes the dynamic interplay between computation and historical interpretation. In this mini-chapter, we will integrate historical theory, code, and artificial intelligence (AI) in a hybrid method of iterative analysis. Such an approach is co-creative: it brings attention to how data processing techniques shape historical analysis, while also recognizing that interpretive insights can lead us to revise our computational approach.

We will use AI mini-chapters like this one to critically examine how AI is used in historical research. Key questions we will explore are whether current AI chatbots can be used to generate meaningful historical interpretations of 19th-century Parliamentary debates, either with a human analyst or by itself.

While AI offers powerful tools for exploratory analysis, we will also examine the limitations of using AI chatbots as an approach to computational text analysis, including the conditions under which they may produce inaccurate or misleading interpretations of history.

We started *Text Mining for Historical Analysis* by teaching readers how to write code without AI because working through the process of writing one's own code step-by-step provides insight into the underlying concepts and steps behind working with text as data. Understanding the premise behind syntax, functions, and how they tie back to data processing enables us to actualize our analytic goals.

With our basic knowledge of data processing we can begin to use AI chatbots in a meaningful and critical way, and we can develop the insight to critique or revise the responses generated by AI chatbots. In this mini-chapter, we demonstrate how to iterate on our analyses by practicing with the AI chatbot, ChatGPT.

Using AI Chatbots for Exploratory Data Analysis

AI chatbots can support iterative research and Exploratory Data Analysis (EDA) by supporting an interactive brainstorming process. Analysts can ask questions, review AI generated responses, and adjust their queries, engaging in a cycle of exploration and refinement. When used this way, an AI chatbot is guided by the analyst to generate focused, relevant responses. However, the analyst must also critique the chatbot's output to determine whether it is accurate or useful. This cyclical method allows analysts to use AI chatbots as a tool for inquiry while maintaining control over the research process.

Here is one example for how we could use a chatbot to support brainstorming: In the previous chapter we removed stop words (e.g., "the" or "and") from our dataset, which significantly improved our results. We can now use ChatGPT to generate additional stop words, specifically attuned to the linguistic characteristics of 19th-century Hansard.

Analyst Prompt

Act like an R Programmer and generate a list of stopwords to be removed for an analysis of the 19th-century Hansard corpus.

ChatGPT-4o Response

```
custom_stopwords <- c("lord", "lords", "hon", "mr", "sir", "gentleman", "gentlemen",
                      "speaker", "right", "honourable", "house", "parliament",
                      "minister", "bill", "member", "question", "motion", "committee",
                      "government", "debate", "order", "division", "clause")
```

Whatever the list of words the chatbot returns, save them now as a list called `custom_stop_words`, using the format for a list in R, where the list is enclosed in parentheses, `c()`, with quotation marks around each word and commas separating the words.

Before we use this list to remove stopwords, however, we should add a layer of human review. What is it that we want to study? Is it really appropriate to remove all the words? If we are interested in the history of the state, we might think twice before removing the word "government" from our search because it may be central to the political concepts, institutional debates, and historical developments we are trying to understand, and removing the word "government" could obscure important patterns in how governance was discussed over time. Each stage in the cleaning and search process requires the analyst to think about what they are doing. Involving AI in the search process does not absolve the analyst of this critical function.

We can also use AI during brainstorming to help us decide which keywords we may want to analyze.

User Prompt

I want to text mine and analyze heated topics from within the 1850 Hansard Parliamentary debates. Suggest several keywords that might be interesting.

ChatGPT-4o Response

When analyzing the 1850 Hansard Parliamentary debates for heated topics, you'll want to focus on issues that were contentious or widely discussed at the time. Here are some keyword suggestions:
Corn Laws – Although the Corn Laws were repealed in 1846, the economic and social impacts continued to be a hot topic of debate.
Irish Famine – The Great Irish Famine (1845–1852) was a significant issue, particularly regarding the British government's response.
Chartism – Though the Chartist movement had declined by 1850, discussions about labor rights, suffrage, and political reform were still prevalent.
Poor Law – The administration and reform of the Poor Laws were highly debated, particularly in relation to poverty and public welfare.
Colonial Governance – Issues around the governance and management of British colonies, particularly in places like India and Ireland.
Free Trade – Discussions around free trade policies, especially in the aftermath of the Corn Laws repeal.
Slavery – Although the British Empire abolished slavery in 1833, discussions about slavery and the slave trade (particularly in relation to other nations) continued.
Catholic Emancipation – Ongoing debates about the rights of Catholics in Britain and Ireland.
Industrialization – Issues related to the rapid industrialization of Britain, including worker rights, factory conditions, and urbanization.
Public Health – Particularly in the context of the cholera outbreaks and the general health of the urban poor.

Whatever the list of keywords the chatbot returns, save them now as a list called `custom_keywords`, using the format for a list in R.

These keywords should help generate a more focused set of results and may assist in identifying debates that were especially contentious or historically significant. You can further refine your analysis by focusing on specific phrases or related terms within these broader topics.

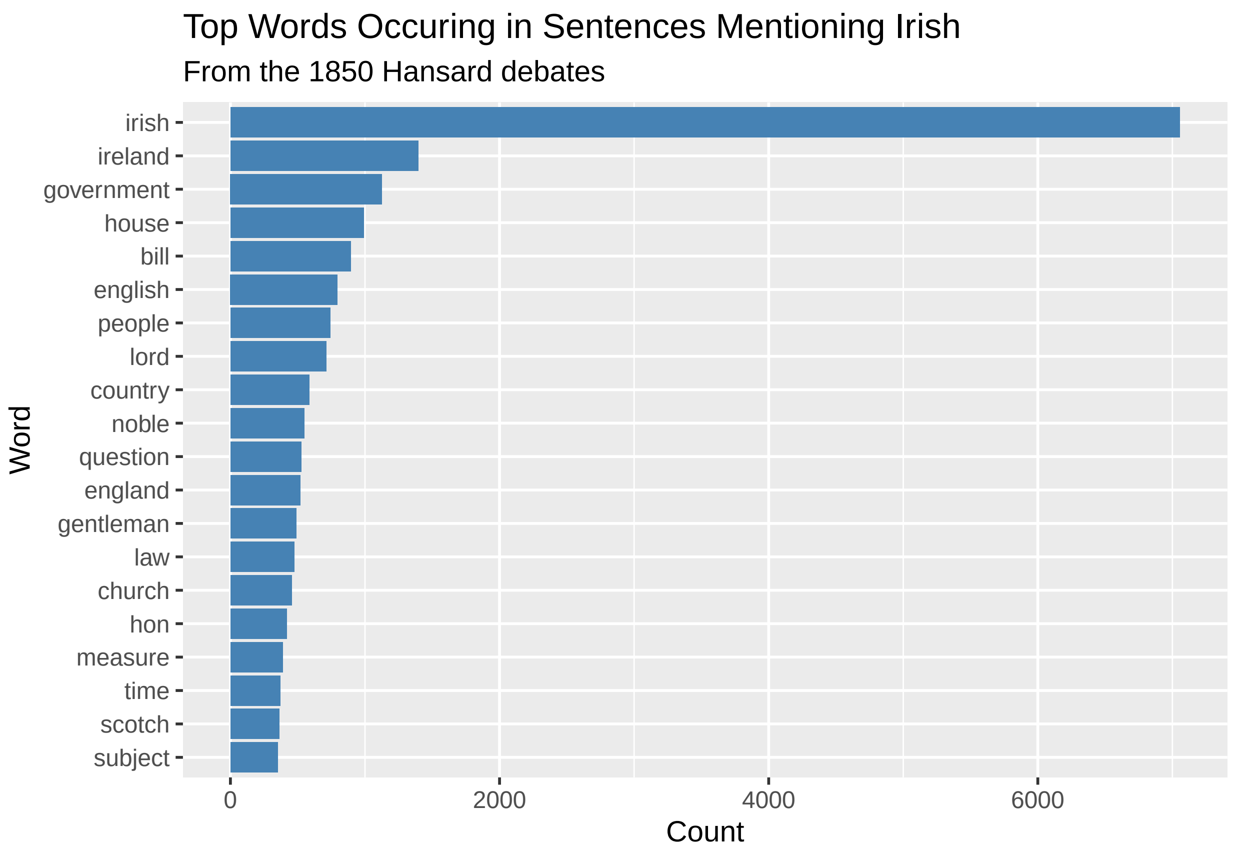
Treating these debate topics as keywords can lead to additional, meaningful analyses. To demonstrate this, we will use the code from Chapter 1 to identify the context for the word "Irish."

```r
library(tidyverse)
library(tidytext)
library(hansardr)
library(kableExtra)

irish_context <- hansard_1850 %>%
  filter(str_detect(text, "\\bIrish\\b")) %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  filter(!str_detect(word, "[:digit:]")) %>%
  count(word) %>%
  arrange(desc(n))
```

```r
top_irish <- irish_context  %>%
  top_n(20)

ggplot(data = top_irish) +
  geom_col(aes(x = reorder(word, n),
               y = n),
           fill = "steel blue") +
  coord_flip() +
  labs(title = "Top Words Occuring in Sentences Mentioning Irish",
       subtitle = "From the 1850 Hansard debates",
       x = "Word",
       y = "Count")
```



Now, alter the code above so that the computer searches for items in the list `custom_keywords`, eliminating instances of words in the list `custom_stop_words`.

Next, we will use the AI chatbot to refine the list of events we searched for in the previous chapter. Here, we will ask the chatbot to generate additional events that might be of interest to someone studying language relating to war in the 1860 Hansard debates. We will guide the chatbot's response by giving it an example of the types of events that interest us taken from the `eventslist` dataset.

Including a template in the prompt (e.g., "return your answer in this format . . .") guides the chatbot to produce a response in the specified structure.

Analyst Prompt

I am providing you with a list of events followed by their year. The purpose of this list is to analyze key events that were mentioned in the 1860 Hansard debates. Name some different events that might be relevant based on the list I provide you. Don't mention the same as provided by the list. Also provide a reason for why the events are relevant.
Return your answer in this format:
Event:
Reason:
These are the events from the list: (copy/paste contents of eventslist)

ChatGPT-4o Response

Event: Abolition of the Slave Trade (1807)
Reason: The abolition of the slave trade would be highly relevant in discussions about moral and economic reforms during the 1860 Hansard debates, particularly in the context of human rights and British colonial policies.
Event: Reform Act (1832)
Reason: This act, which expanded suffrage and restructured parliamentary representation, would likely influence discussions on electoral reform and representation in 1860.
Event: Crimean War (1853–1856)
Reason: The Crimean War had significant impacts on British foreign policy, military reforms, and public health improvements, all of which could be relevant topics of debate in 1860.
Event: Irish Potato Famine (1845–1852)
Reason: The famine's aftermath would remain a pressing issue, influencing debates about Irish governance, economic policies, and relief efforts.
Event: Great Exhibition (1851)
Reason: As a symbol of industrial progress and imperial power, the Great Exhibition could be discussed in terms of economic innovation and Britain's global role in the mid-19th century.
Event: Indian Rebellion (1857)
Reason: The rebellion was a turning point in British imperial governance, leading to significant debates about military strategy, colonial administration, and the justification of imperial rule.
Event: Public Health Act (1848)
Reason: This act, which addressed urban sanitation and public health, would be relevant to discussions about improving living conditions and public health policies in 1860.
Event: Establishment of the Penny Post (1840)
Reason: The reform in postal services, increasing communication efficiency and accessibility, could relate to debates on economic and infrastructural development.

Our initial prompt to the AI chatbot successfully guided it to identify and return events similar to the query. However, when we provided the chatbot with the entire contents of eventslist, the results lacked precision and specificity. This outcome suggests that more targeted guidance is necessary to focus on the specific characteristics of interest. To tailor the results to a specific research question, we will refine the prompt further by narrowing in on the events we are most interested in from `eventslist` and instructing the chatbot to return like-events, in this case, other names of meetings that occurred in the 1860s.

Analyst Prompt

I am providing you with a list of events followed by their year. The purpose of this list is to analyze key events that were mentioned in the 1860 Hansard debates. Name some different events that might be relevant based on the list I provide you. Don't mention the same events as provided by the list. Also provide a reason for why the events are relevant based on the list I provide.
Return your answer like this:
Event:
Date:
Reason:
These are the events from the list: "greek war", "napoleonic wars."

ChatGPT-4o Response

Here are some different events that might be relevant based on your list:

Event: Chartist Movements
AI Assigned Date: 1838–1850
Reason: The Chartist movements, with their organized public gatherings for political reform, directly relate to the mentioned events like "public meeting" and "recent meeting," illustrating how assemblies have historically been used to drive societal change.

Event: Anti-Corn Law League Meetings
AI Assigned Date: 1839–1846
Reason: These meetings share a thematic connection to "public meeting" by demonstrating how public gatherings have been pivotal in shaping discourse around economic reform, similar to those discussed in the 1860 debates.

Event: Hyde Park Demonstrations
AI Assigned Date: 1855–1860
Reason: Hyde Park demonstrations resonate with "public meeting" and "recent meeting," showing how such gatherings functioned as platforms for political and social advocacy, reflecting the period's reliance on public discourse for progress.

Event: Meetings on the Reform Act of 1832
AI Assigned Date: 1831–1832
Reason: These meetings are akin to "yorkshire meeting" and "clontarf meeting," showcasing historical instances where regional or national public assemblies catalyzed significant legislative reforms, serving as a possible reference point in 1860.

Event: Fenian Meetings
AI Assigned Date: 1858–1860
Reason: While slightly later, Fenian meetings parallel "clontarf meeting" by highlighting Irish political concerns and illustrating how public gatherings were used to address issues of representation and governance.

Now that we have a more specific list of events that aligns with our interests, we can use them in our analysis. To use our new events, we will create a data frame for them. We can prompt the chatbot to return the needed code. Given that we are very familiar with creating data frames at this point, we can also proof the chatbot's results and ensure they match our desired output.

Analyst Prompt

Return this text as a data frame in R with two columns: a column for the event, named event_name and a column for the date, named AI Assigned Date. Write the events in lower case.

ChatGPT-4o Response

Here is the R code to create the requested data frame:

```
# Create the data frame

events_df <- data.frame(
  event_name = c(
    "crimean war",
    "russo-turkish war",
    "congress of vienna",
    "siege of acre",
    "battle of navarino",
    "schleswig-holstein wars"
  ),
  AI_Assigned_Date = c(
    "1853-1856",
    "1828-1829",
    "1814-1815",
    "1799",
    "1827",
    "1848-1852"
  )
)

# Print the data frame
print(events_data)
```

This will create a data frame in R with two columns: event_name (containing the event names in lowercase) and AI Assigned Date.

We can now test the chatbot's responses by running the code.

```r
# Create the data frame
events_df <- data.frame(
  event_name = c(
    "crimean war",
    "russo-turkish war",
    "congress of vienna",
    "siege of acre",
    "battle of navarino",
    "schleswig-holstein wars"
  ),
  AI_Assigned_Date = c(
    "1853-1856",
    "1828-1829",
    "1814-1815",
    "1799",
    "1827",
    "1848-1852"
  )
)

# Print the data frame
```

Table 1: Excerpt from Hansard 1850

sentence_id	text	speechdate	debate	matched_event
S3V0156P0_3555	As to the noble Earl's argument against the abandonment of a prohibitive duty on coal in the event of war, I have only to say that at the time of the Crimean war we did issue a proclamation prohibiting the export of contraband of war to countries north of Dantzic, but that we did not extend it to coal.	1860-02-20	QUESTION.	Crimean war
S3V0156P0_8154	The Crimean war, while it created an enormous extra demand for ships, curtailed in a very small degree the usual channels of commerce, and the consequence was that a larger number of ships was called into requisition than was necessary for the usual trade of the country.	1860-01-31	COMMITTEE MOVED FOR.	Crimean war
S3V0156P0_8400	Member for Sunderland himself, who, in a work which he had recently published, made mention of the adversity under which the shipping interest had laboured over since the date of the Crimean war, and declared the prediction that that adversity was only a mere passing cloud to have been completely falsified.	1860-01-31	COMMITTEE MOVED FOR.	Crimean war
S3V0156P0_10788	With regard to the claims of officers of the Land Transport Corps, a Committee of that House had sat, before which was brought the case of certain officers who had entered that corps under very peculiar circumstances, and who at the conclusion of the Crimean War were dismissed from that service.	1860-02-03	OBSERVATIONS.	Crimean War
S3V0156P0_14522	He did not think it possible to conceive waste like that which had taken place in disposing of the surplus stores after the Crimean war.	1860-02-10	COMMITTEE.	Crimean war

```r
print(events_df)
```

```
##                 event_name AI_Assigned_Date
## 1             crimean war        1853-1856
## 2       russo-turkish war        1828-1829
## 3      congress of vienna        1814-1815
## 4           siege of acre             1799
## 5      battle of navarino             1827
## 6 schleswig-holstein wars        1848-1852
```

```r
data("hansard_1860")
data("debate_metadata_1860")

hansard_1860 <- hansard_1860 %>%
  left_join(debate_metadata_1860)
```

```
## Joining with ‘by = join_by(sentence_id)‘
```

The following code filters the `hansard_1860` dataset to include rows where the `text` column matches any event name in `events_df$event_name`, using case-insensitive matching. It then adds a new column, `matched_event`, which contains the specific matched event name from the `text`. The result is a filtered data frame with an additional column identifying the matched event for each row.

```r
# Match with each event and
# add a column to capture the matched event
new_matched_events <- hansard_1860 %>%
  filter(str_detect(text,
                    regex(paste0("\\b(", paste(events_df$event_name, collapse = "|"), ")\\b"),
                          ignore_case = TRUE))) %>%
  mutate(matched_event = str_extract(text,
                                     regex(paste0("\\b(", paste(events_df$event_name, collapse = "|"),
                                           ignore_case = TRUE)))
```

The results show that not all of the events were mentioned word-for-word in the Hansard corpus; however, we successfully matched several, including the Crimean War, Congress of Vienna, and Battle of Navarino.

One important feature of working with AI is that the AI may not be a reliable substitute for working with text mining. As of the writing of this book, AI cannot reliably perform corpus-scale counting or quantitative validation, making it useful for writing code but less useful for performing the analysis itself. It can generalize probabilistically, allowing it to find related words and possible stopwords, but it can also return false positives and false negatives. Checking the match of AI wordlists to the actual data is therefore an important step. The checking step allows the scholar to ask: is this a tool that I can trust for this specific purpose with a dataset of this size, right now?

While an AI chatbot can suggest a wider range of related terms to search for–and these terms might spark our curiosity–those terms may not actually appear in the corpus. For experts 19th-century history, it might be expected that some phrases would not be found, but for others, this process can offer a useful way to explore which events were and were not mentioned in the data. This kind of exploration is part of an ongoing, iterative process—where analysts refine their searches by adding or adjusting keywords, either manually or with AI.

It is also important to note that the AI chatbot did not automatically fix formatting issues in the text, such as inconsistent capitalization. To obtain more reliable and visually meaningful results, the analyst may want to take extra processing steps to clean the data—such as making all text lowercase and fixing other formatting problems—before analysis.

```r
event_counts <- new_matched_events %>%
  count(matched_event, name = "count") %>%
  arrange(desc(count))

print(event_counts)
```

```
##        matched_event count
## 1        Crimean war   360
## 2        Crimean War   143
## 3 Congress of Vienna    42
## 4 battle of Navarino     5
```