Optimisation as a tool for natural language processing

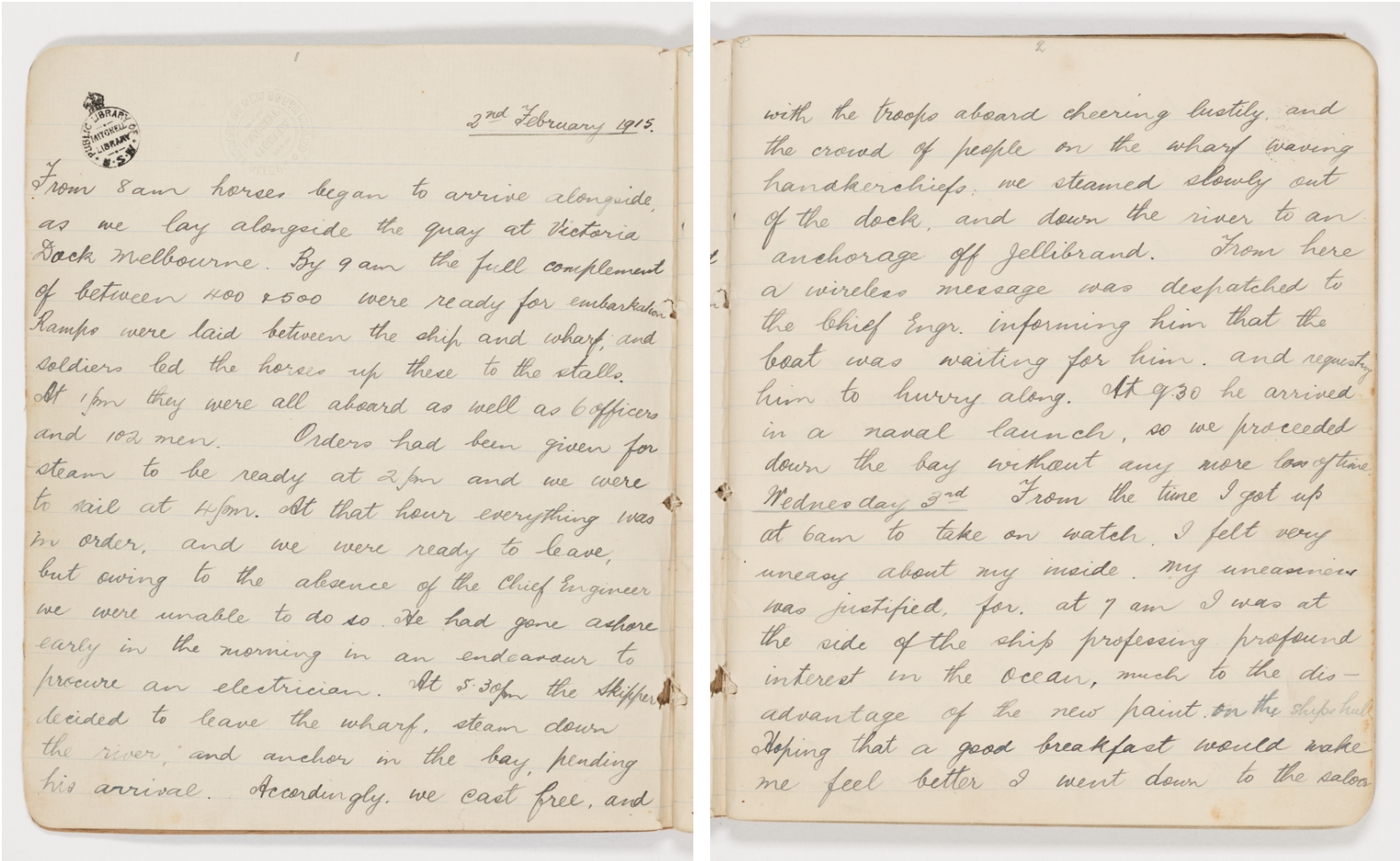
Ashley Dennis-Henderson

ashley.dennis-henderson@adelaide.edu.au
ashley-dennis-henderson
@AshleyGrace63



THE UNIVERSITY of ADELAIDE

Motivation: SLNSW WW1 Collection



2nd February 1915.

From 8 am horses began to arrive alongside as we lay alongside the quay at Victoria Dock Melbourne. By 9 am the full complement of between 400 & 500 were ready for embarkation. Ramps were laid between the ship and wharf, and soldiers led the horses up these to the stalls. At 1 pm they were all aboard as well as 6 officers and 102 men. Orders had been given for steam to be ready at 2 pm and we were to sail at 4 pm. At that hour everything was in order, and we were ready to leave, but owing to the absence of the Chief Engineer we were unable to do so. He had gone ashore early in the morning in an endeavour to procure an electrician. At 5.30 pm the Skipper decided to leave the wharf, steam down the river, and anchor in the bay, pending his arrival. Accordingly, we cast free, and

with the troops aboard cheering lustily, and the crowd of people on the wharf waving handkerchiefs, we steamed slowly out of the dock, and down the river to an anchorage off Jellibrand. From here a wireless message was despatched to the Chief Engr. informing him that the boat was waiting for him, and requesting him to hurry along. At 9.30 he arrived in a naval launch, so we proceeded down the bay without any more loss of time.

Wednesday 3rd From the time I got up at 6 am to take on watch I felt very uneasy about my inside. My uneasiness was justified, for, at 7 am I was at the side of the ship professing profound interest in the ocean, much to the disadvantage of the new paint on the ship's hull. Hoping that a good breakfast would make me feel better I went down to the saloon

Figure: Gilroy, Norman. Norman Thomas Gilroy war diary, Volume 1: 2 February-29 April 1915. Mitchell Library, State Library of New South Wales.

Life still goes on:
Analysing Australian WW1 Diaries through Distant Reading

Ashley Dennis-Henderson, Matthew Roughan, Lewis Mitchell, Jonathan Tuke
ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS)
School of Mathematical Sciences, The University of Adelaide
{ashley.dennis-henderson,matthew.roughan,lewis.mitchell,simon.tuke}@adelaide.edu.au

Abstract

An increasing amount of historic data is now available in digital (text) formats. This gives quantitative researchers an opportunity to use distant reading techniques, as opposed to traditional close reading, in order to analyse larger quantities of historic data. Distant reading allows researchers to view overall patterns within the data and reduce researcher bias. One such data set that has recently been transcribed is a collection of over 500 Australian World War I (WW1) diaries held by the State Library of New South Wales. Here we apply distant reading techniques to this corpus to understand what soldiers wrote about and how they felt over the course of the war. Extracting dates accurately is important as it allows us to perform our analysis over time, however, it is very challenging due to the variety of date formats and abbreviations diarists use. But with that data, topic modelling and sentiment analysis can then be applied to show trends, for instance, that despite the horrors of war, Australians in WW1 primarily wrote about their everyday routines and experiences. Our results detail some of the challenges likely to be encountered by quantitative researchers intending to analyse historical texts, and provide some approaches to these issues.

PhD Project Aims

- To develop an optimisation technique which accurately extracts dates from diaries
- To investigate and improve the accuracy of current sentiment analysis techniques on historic text
- To use natural language processing techniques to investigate Australian World War I diaries

Date Extraction: Aim

Date	Entry
02/02/1915	from 8am horses began to arrive alongside, ...
08/02/1915	was so ill during this period that i was unable ...
28/03/1915	many large french and british transports ...
05/04/1915	at 5.30a we were towed out from the wharf ...
25/07/1915	arose about 6.30 and went down to church ...
27/07/1915	at 5am we were shifted to a berth further down ...

Date Extraction: Challenges

"Tuesday 9th February 1915. At 6am woke to find we were "

"Wed Jan 30th Left line tonight being relieved by the 54th BN."

"Wednesday 24th Feby. Very disappointed when we received a"

"Tuesday May 8.5.17. Stand to 5.30 A.M. cool morning fed up"

"5th. Last night I had a party of 20 men repairing the blown in"

"17: Pte. Sellars fell down hold 50 ft., dying four hours later,"

"8th Field Ambulance. Breakfast had been previously arranged"

Date Extraction

date	month	year
2	2	1915
	⋮	
2	4	0
3	3	0
4	3	0
5	4	0
	⋮	
25	7	0
5	5	0
26	7	0

Date Extraction

date	month	year	
2	2	1915	
	⋮		
2	4	0	
3	3	0	← mistaken dates
4	3	0	
5	4	0	
	⋮		
25	7	0	
5	5	0	← non-entry date
26	7	0	

Date Extraction

$$\min_{d_i,\, m_i,\, y_i} \sum_{i=1}^{n} \Big(\alpha I_i^d |d_i - \hat{d}_i| + 31\beta (I_i^m |m_i - \hat{m}_i|) + 372\gamma (I_i^y |y_i - \hat{y}_i|) \Big)$$

$$+ \sum_{j=1}^{n-1} \theta (x_{j+1} - x_j) + \omega (x_1 - s_d)$$

$$\begin{aligned}
\text{such that} \quad s_d &\leq x_i, \quad \text{for } i = 1, \ldots, n \\
x_i &\leq e_d, \quad \text{for } i = 1, \ldots, n \\
x_i &\leq x_{i+1}, \quad \text{for } i = 1, \ldots, n-1 \\
x_i &= 372(y_i - s_y) + 31(m_i - 1) + d_i - 1
\end{aligned}$$

Date Extraction

Example, 25th April 1915

$$\begin{aligned} x_i &= 372 * (1915 - 1914) + 31 * (4 - 1) + 25 - 1 \\ &= 489. \end{aligned}$$

Date Extraction

date	month	year
2	2	1915
	⋮	
2	4	0
3	3	0
4	3	0
5	4	0
	⋮	
25	7	0
5	5	0
26	7	0

$\Longrightarrow$

date	month	year
2	2	1915
	⋮	
2	4	1915
3	4	1915
4	4	1915
5	4	1915
	⋮	
25	7	1915
25	7	1915
26	7	1915

Date Extraction

date	month	year
2	2	1915
	⋮	
2	4	0
3	3	0
4	3	0
5	4	0
	⋮	
25	7	0
5	5	0
26	7	0

$\Longrightarrow$

date	month	year
2	2	1915
	⋮	
2	4	1915
3	4	1915
4	4	1915
5	4	1915
	⋮	
25	7	1915
25	7	1915
26	7	1915

Fine Tuning Parameters

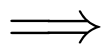
$$\min_{d_i,\, m_i,\, y_i} \sum_{i=1}^{n} \left(\alpha I_i^d |d_i - \hat{d}_i| + 31\beta (I_i^m |m_i - \hat{m}_i|) + 372\gamma (I_i^y |y_i - \hat{y}_i|) \right)$$

$$+ \sum_{j=1}^{n-1} \theta (x_{j+1} - x_j) + \omega (x_1 - s_d)$$

Most accurate parameter set: $(\alpha, \beta, \gamma, \theta, \omega) = (25,\ 1,\ 1,\ 1,\ 1)$ or $(50,\ 1,\ 1,\ 1,\ 1)$

Example

	A	B	C
1	Date	Month	Year
35	1	4	0
36	2	4	0
37	3	3	0
38	4	3	0
39	5	4	0
40	6	4	0
41	7	4	0
42	8	4	0
43	8	0	0
44	9	4	0
45	18	0	0
46	19	4	0
47	20	4	0
48	21	4	0
49	21	4	1915
50	22	4	0
51	23	4	0
52	24	4	0
53	25	4	1915
54	26	4	0
55	27	4	0
56	28	4	0
57	29	4	0



	A	B	C
1	Date	Month	Year
34	1	4	1915
35	2	4	1915
36	3	4	1915
37	4	4	1915
38	5	4	1915
39	6	4	1915
40	7	4	1915
41	8	4	1915
42	9	4	1915
43	18	4	1915
44	19	4	1915
45	20	4	1915
46	21	4	1915
47	22	4	1915
48	23	4	1915
49	24	4	1915
50	25	4	1915
51	26	4	1915
52	27	4	1915
53	28	4	1915
54	29	4	1915

Example

	A	B	C
1	Date	Month	Year
50	9	7	1915
51	10	7	1915
52	4	0	0
53	11	7	1915
54	12	7	1915
55	13	7	0
56	14	7	1915
57	15	7	1915
58	16	7	1915
59	17	7	1915
60	4	0	0
61	18	7	1915
62	18	7	1915
63	19	7	1915
64	2	0	0
65	20	7	1915
66	21	7	1915
67	22	7	1915
68	23	7	1915
69	24	7	1915
70	25	7	1915
71	5	5	0
72	26	7	1915
73	12	0	0
74	1	3	0

	A	B	C
1	Date	Month	Year
75	27	7	1915
76	28	7	1915
77	29	7	1915
78	30	7	1915
79	1	8	1915
80	2	8	1915
81	3	8	1915
82	4	8	0
83	5	8	0
84	6	8	1915
85	7	8	1915
86	8	8	1915
87	4	0	0
88	9	8	1915
89	10	8	1915
90	11	8	1915
91	12	8	1915
92	13	8	1915
93	14	8	1915
94	15	8	1915
95	16	8	1915
96	17	8	1915
97	18	8	1915
98	1	7	1915
99	2	7	0

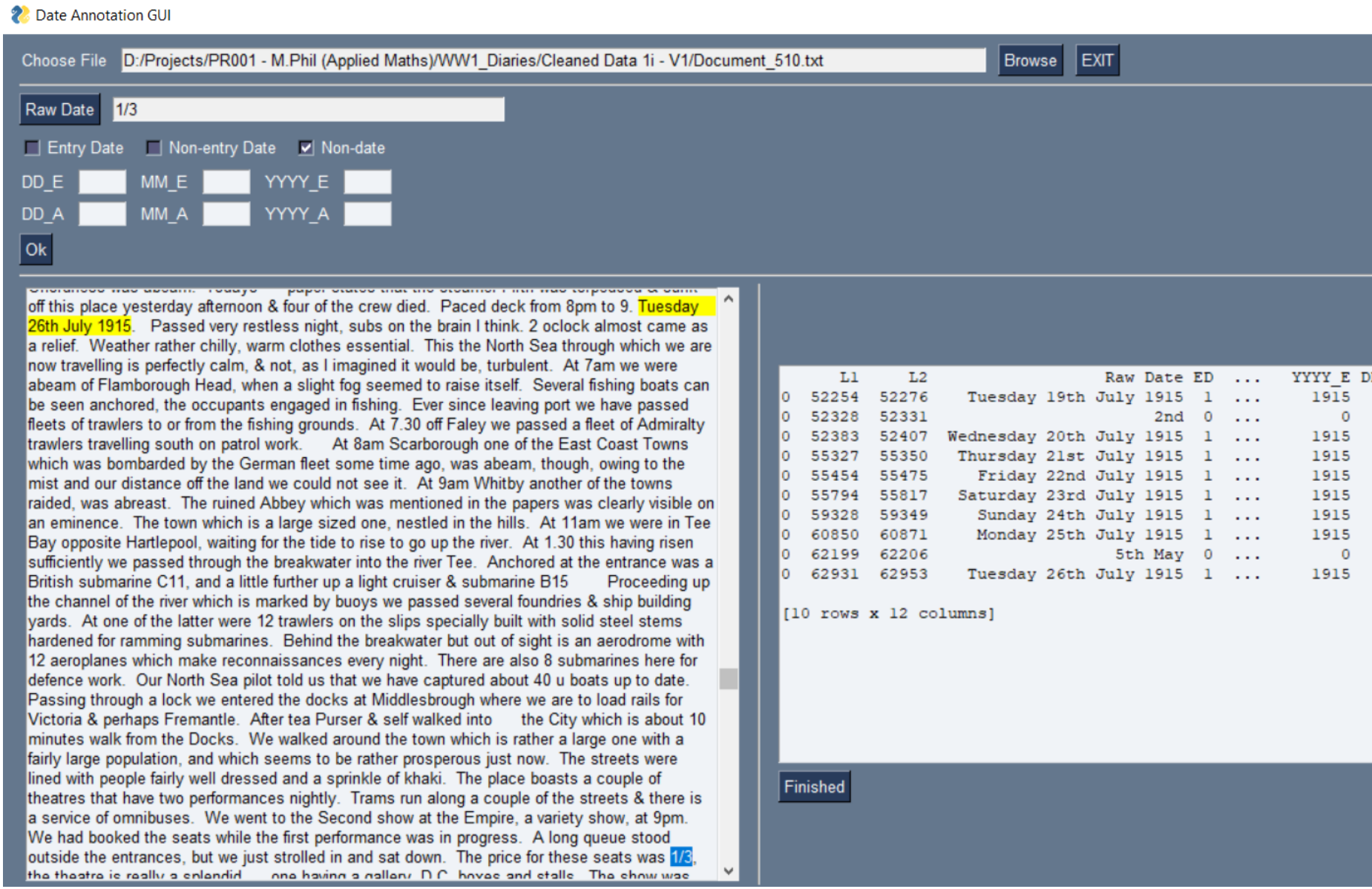
$\Longrightarrow$

	A	B	C
1	Date	Month	Year
40	1	7	1915
41	2	7	1915
42	3	7	1915
43	4	7	1915
44	5	7	1915
45	6	7	1915
46	7	7	1915
47	8	7	1915
48	9	7	1915
49	10	7	1915
50	11	7	1915
51	12	7	1915
52	13	7	1915
53	14	7	1915
54	15	7	1915
55	16	7	1915

Future Work

- Manually extract dates from subset of diaries to determine accuracy
- Research and test other methods of extracting dates (NER)
- Investigate difference between including/excluding end date
- Include and indicator for out-of-order diaries
- Create a method for determining the optimal parameters of the program

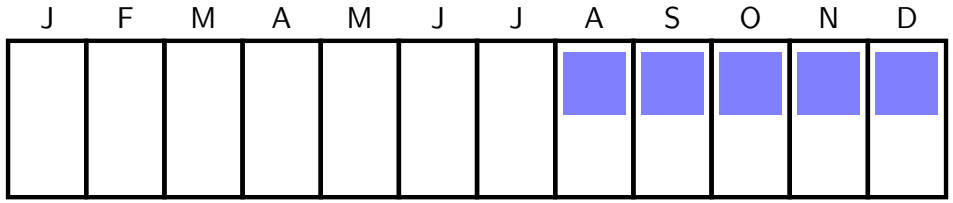
Future Work: Testing Accuracy



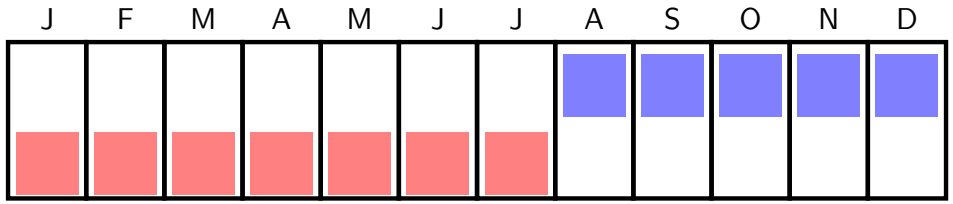
Future Work

- Manually extract dates from a subset of diaries to determine accuracy
- Research and test other methods of extracting dates (NER)
- Investigate the difference between including/excluding end date
- Include an indicator for out of order diaries
- Create a method for determining the optimal parameters of the program

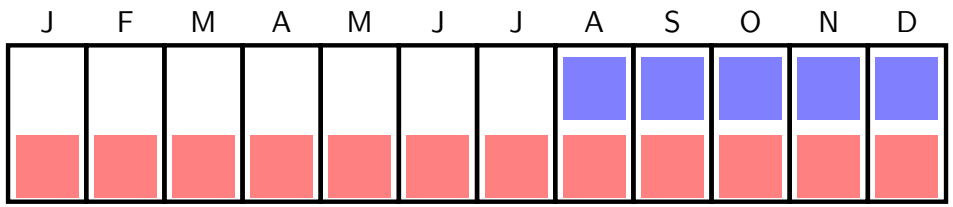
Out of Order Diaries



The diarist begins by writing the entries for all the 1916 dates after purchasing the diary.



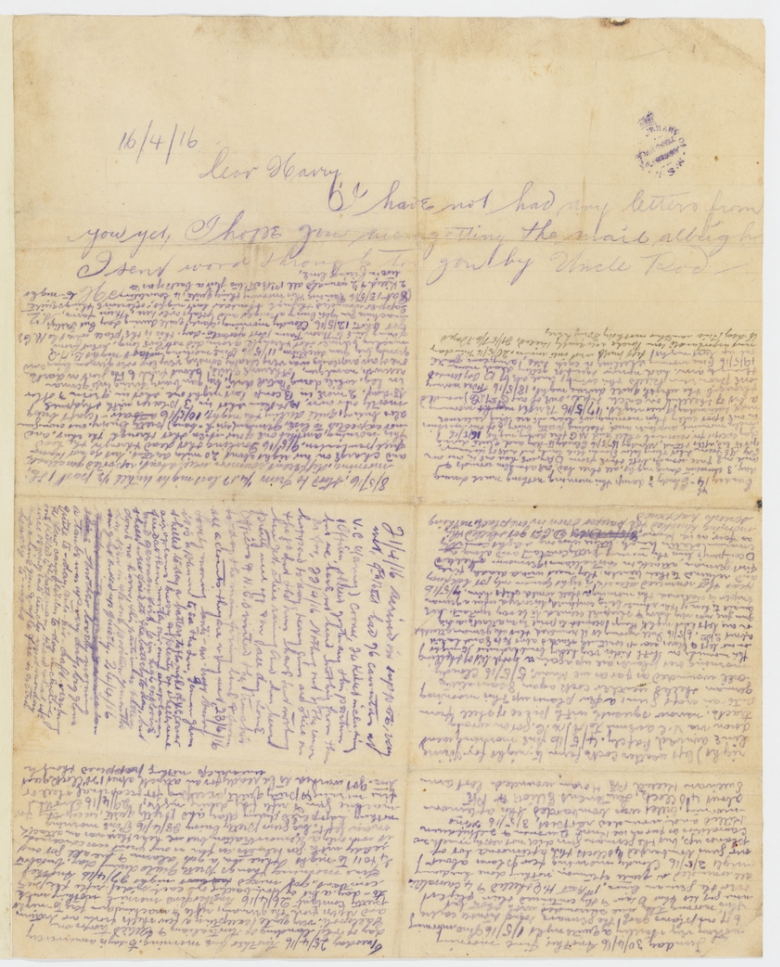
The diarist then writes their 1917 entries from the beginning of the diary.



Some diarists then continued writing their 1917 entries underneath the 1916 entries until the year finished.

Legend: 1916 entries, 1917 entries

Out of Order Diaries



Letter: 16/04/16 **1** **1**	
08/05/16 **2** to 13/05/16 **6**	14/05/16 **3** to 21/05/16 **7**
21/04/16 **4** to 24/04/16 **2**	04/05/16 **5** to 07/05/16 **5**
25/04/16 **6** to 29/04/16 **3**	30/04/16 **7** to 03/05/16 **4**

Future Work

- Manually extract dates from a subset of diaries to determine accuracy
- Research and test other methods of extracting dates (NER)
- Investigate the difference between including/excluding end date
- Include an indicator for out of order diaries
- Create a method for determining the optimal parameters of the program

Questions?