Final Project – Fail Log

- I thought I'd be able to knock out a few of the war-diary .jpgs at once, but I realized how straining it is on the eyes almost immediately. While I have no problem staring at my computer all day, actually having to heavily focus made it a lot harder
- A lot of the text blended together and I found myself going back to correct errors because I would input the date or location of the person below or above them, and some letters were just ink blots.
- While I do think the OCR text I did through RStudio is significantly worse than my own transcription, I can see the need for OCR technology, because there's a huge chance of human error with this
- However, we can't accept the quality of work that RStudio produced

- I ended up doing manual translation for the last ten files in the war-diary .txt, and it was compiled with Lauren Rollit and Alexei Tipenko's transcriptions as well, which were all merged into Lauren's GitHub, which is open for anyone in the class to download
- Personally, transcribing is one of my favourite things to do, although perhaps I enjoy more tedious work
- The most difficult aspect of it for me was ensuring that all the words were spelled the way it was shown on the original .jpg because my eyes would read the word, and even if there was a spelling error, my brain would overlook it and put in the correct spelling when I transcribed it, so I had to be very careful of that

- I first put the .txt files through the Topic Modeling Tool, and set the topics to 20 as I wasn't too sure where to go first
- I noticed the first time I did this, the list of topics were a little weird. I got results such as: 12 tng ravenscroft transferred res sick quarters officers infectious deaths billets epidemic ward held open bigger day cases brighton prevalent
- I'm not entirely sure if that's how they're supposed to look however I can pick out the theme from that topic as it has to do with more severe injuries and illnesses, and the location of Ravenscroft and Brighton
- I'm not sure if there's another way to make it cleaner?
- I went back to look at Module 4's Topic Modeling exercise where I used the Colonial Newspaper Database, and while the words are still very random, they make much more sense; all the words are complete words
- This may be due to the war-diary .jpgs quality as some parts had ink blots or crossed out information, or just half a word
- I reran the war-diary .txt files and used 10 topics instead, and the results were much clearer (you can see in my GitHub the first and then second time I used the Topic Modeling tool to compare
- I'm trying to upload the .csv file that was created from the Topic Modeling Tool to Overview, however I'm getting an error saying that: the first line of the file must list the column names. One column must be named "text", "snippet", or "contents" and I'm not sure what that means as the .csv file has headings

- I went back to compare the CND.csv to see what I could be missing but I'm not sure
- I changed the heading to 'Text' instead of 'Topic ..' and it worked!
- Ran my files through AntConc to see the frequency of specific words, and then eliminated single letters, function, and content words such as 'and' or 'the' to clean it up
- I found AntConc a little easier to figure out the frequency of the words as Overview gave me a word cloud, and the way that worked was they showed me which documents that word was in; both very useful but I think AntConc makes more sense for this purpose
- I then put the .txt files into Voyant Tools, I'm trying to decide which method to use to see what works best for my analysis (or multiple tools)
- I think I prefer Voyant Tools because it shows a lot more information, and the overall look is much cleaner for me to understand and pull data from
- The TermsBerry was a great tool to use because I could see when I hovered over one word, which other words were used frequently with it
- For example, 'nursing' and 'sister' were very frequently used together as in the texts, they were commonly referred to as Nursing Sister