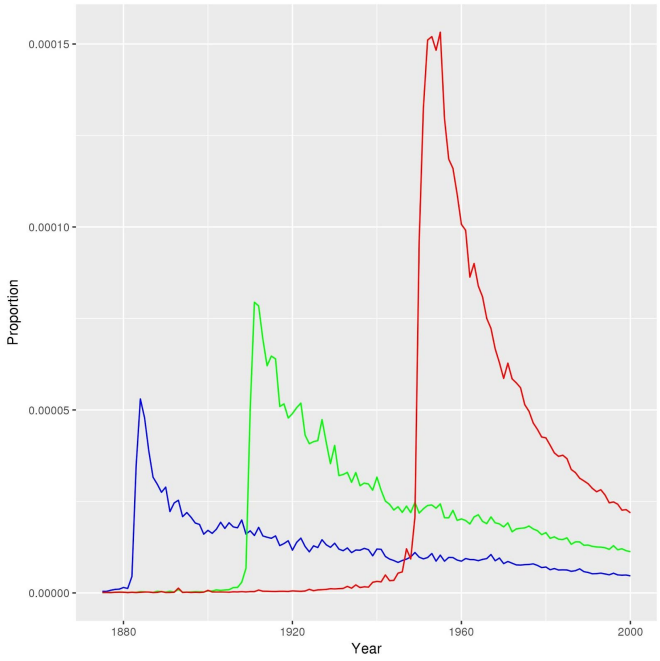
SODA 501 First Exercise

Claire Kelling, Sara Francisco, Rosemary Pang, Fangcao Xu

A. Get the raw data from Google Books NGram Viewer

Done!

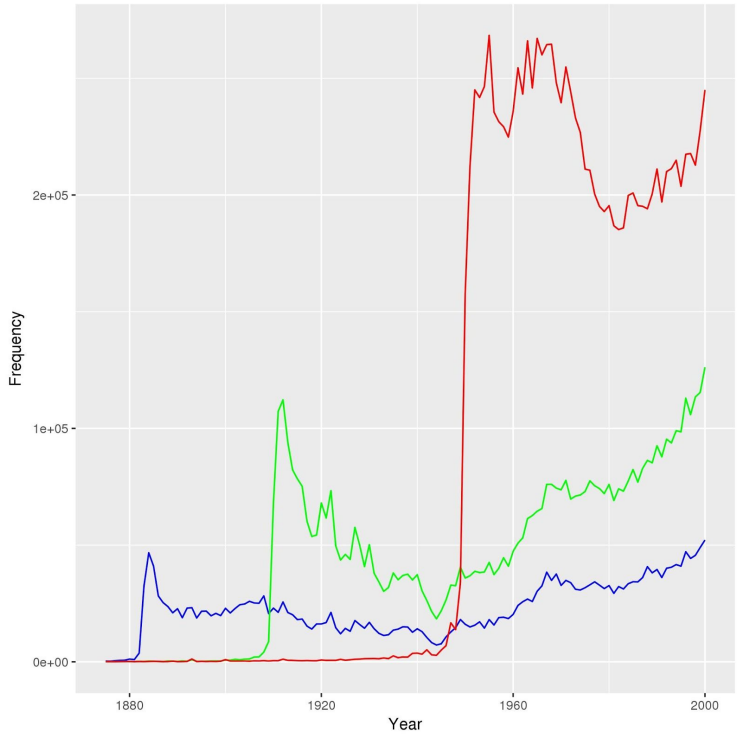
B. Recreate the main part of figure 3a



Usage frequency is computed by dividing the number of instances of the «-gram in a given year by the total number of words in the corpus in that year. It highlights two central factors that contribute to trends. Cultural change guides the concepts we discuss. Linguistic change, affects the words we use for those concepts ("the Great War" versus 'World War).

C. Now check your graph against the graph created by the Ngram viewer.

D. Recreate figure 3a (main figure), but change the y-axis to be the raw mention count.

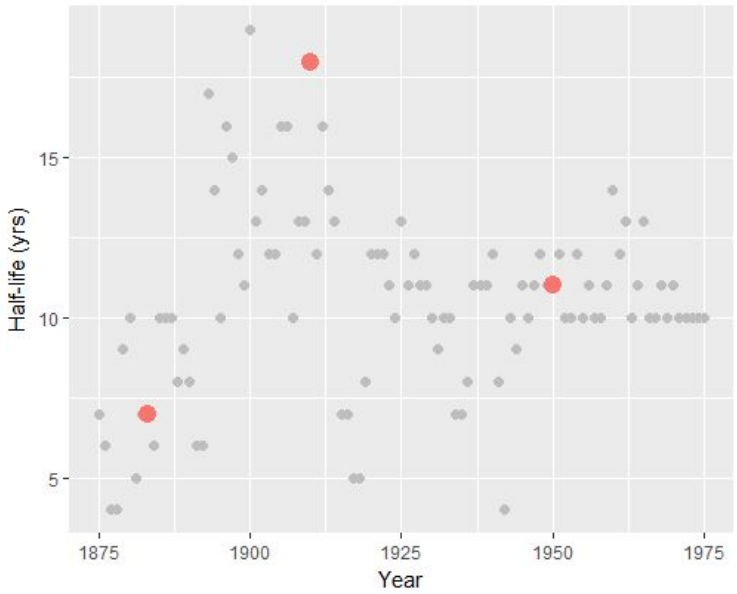


E. Does the difference between (b) and (d) lead you to reevaluate any of the results of Michel et al. 2011). Why or why not?

The difference between (b) and (d) does lead us to reevaluate some of the results form Michel et al. 2011. Looking at (d), while the number of raw counts does decrease for each example, the trend is not consistent. After 20-50 years, the raw count increases steadily, suggesting that these “years” are not being “forgotten faster”.

F. Calculate the half-life of each of the years. Does version 2 of the NGram data produce similar results of those presented in Michel et al. (2011), that are based on version 1 data?

Looking at the graphs created using data from version 2 and version 1, they do not appear to produce similar results. The graph created from version 1 shows a decreasing trend, whereas the graph produced using data from version 2 does not have a consistent trend. The differences between both graphs is probably due to the data and the differences between both versions.



G. Were there any years that were outliers, such as years that were forgotten particularly quickly or slowly?

While I did not statistically calculate outliers, it seems that there was a year around 1940 that was forgotten particularly quickly, as the half-life is very low. In other words, it decayed very quickly. However, it is very similar to the years at the start of the range. Not surprisingly, 1900 has a high half life and took awhile to forget. There don't appear to be other extreme outliers.