1 Exercise 1

This is the answer to question 1.

1. I have ran these commands in R and produced a plot.

Figure 1: Question 1 Part 1

2. **Inclusions** = ['thousand', 'institutions', 'enjoy', 'conferred', 'cause', 'benefits', 'children', 'bequeathed', 'childrens', 'continue', 'generations', 'glorious', 'united', 'yet', 'may', 'country', 'upon', 'rejoice', 'Washington', 'compeers'] .
Exclusions = ['under', 'his', 'us', 'to', 'those', 'and', 'our', 'the', 'a', 'have', 'by']

   Wordcloud removes stopwords which is why some words are included and others are excluded. This is because we passed a string and left the frequency parameter unspecified which results in stopword removal.

3. **Sentence**: "Operational Incident affecting application hosted on MuleSoft Anypoint Platform occurs. Organization operational support resources perform necessary initial incident assessment and triage. Decision: is assistance required from MuleSoft support to assist with incident diagnosis or resolution? Gather the minimum information required based on the profile of the incident. Please see: section below for minimum information that is required when opening a support a ticket."
Plot:

Figure 2: Question 1 Part 3

   Inclusions: Required, Incident, Support
   Exclusions: All words are excluded except the three above.
   While my hypothesis that stopwords are excluded is validated, we see that fewer words are returned than we would expect if only stopwords were removed. Checking the wordcloud documentation here we see that while stopwords are removed if the text is passed as a string and frequency is not set, we can also see a default minimum frequency of three is set (it's not clear to me why this is not impacted in the first part of this question).

4. Repeating some words multiple times with the new sentence:
   "Operational Operational Operational Incident affecting affecting affecting application hosted on MuleSoft Anypoint Platform occurs. Organization operational support support resources perform perform perform perform necessary necessary necessary initial incident assessment and triage. Decision: is assistance required from MuleSoft MuleSoft MuleSoft MuleSoft MuleSoft MuleSoft support to assist with incident diagnosis or resolution? Gather the minimum information required based on the profile of the incident. Please see: section below for minimum information that is required when opening a support a ticket."
   Plot:

Figure 3: Question 1 Part 4

We see our repeated words that appear three or more times are now included. While it's still not clear to me as to why words which appeared less than three times appear in the wordcloud in the original text provided, we see by passing in the parameter min.freq=1 that we can alter this behaviour to produce the below:

Figure 4: Question 1 Part 4 ii

2 Exercise 2

This is the answer to question 2.

1. The key peaks we see are approximately in 1820, 1955, 1970, 1975, and 2001. For the 19th century peak, we see "Mark Keane" is featured in "Reports from Committees" for the British House of Commons in this time period. In the 1954-1956, we see Mark Keane (here) was the name of a village manager in Oak Park Illinois resulting in this appearing in public service bulletins. By the 1970s peak we see Mark Keane was an executive director of the International City Management Association resulting in a greater volume of publications referring to him through government documents. The 2001 peak primarily stems from UCD lecturer Mark Keane's publications in psychology, cognitive science, and computer science.

2. Adam Ryan has peaks in 1812, the 1940s, and 2008-2010. The hits in 1812 are predominantly driven by false hits from co-authors with a first name of Adam and a first name (or surname) of Ryan but not a single individual. A putting quotes around the name in this time period reveals no hits. In the 1940s Adam Ryan peaks due to a publication in the Proceedings of the National Tax Association however, again, this is an incorrect citation from one attendant named "Adam Rumoshovsky" and "Frank J Ryan" which, due to their formatting, is captured as "Adam Ryan" in NGram. An Adam Ryan from Philadelphia set a Ring Record resulting in his publication in "Nat Fleischer's All-time Ring Record Book" in 1943. The peak in 2008-2010 does not stem from a singular Adam Ryan, but is present due to an increasing usage of the name both in fiction, a Benedictine monk named Adam Ryan resulting in publications in religious texts, and in society at large resulting in references to numerous students/children/workers called Adam Ryan.

3. The word I chose is "Tweet" which is used to refer in its modern form to the act of posting on Twitter. We see that, in fact, it is actually a relatively recent word in its usage. While The earliest usages of it pre-Twitter stem from onomatopoeic references to birds, or characters named tweet, it's explosion in usage coincides with the introduction of social media and the altering of meaning of the word.

4. I've looked at the term "Reusable Energy" as the other graphs were relatively uninteresting with smoothing. The results of various smoothing is shown n figures 6, 7, 8. Smoothing these searches smoothes some of the blips which occur in the charts by averaging values over the number of years chosen (per here)to make it easier to analyse the trend

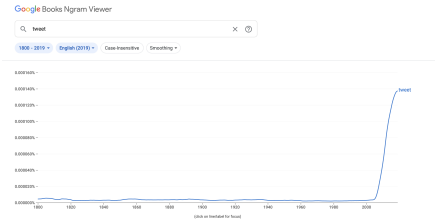

Figure 5: Question 2 Part 3

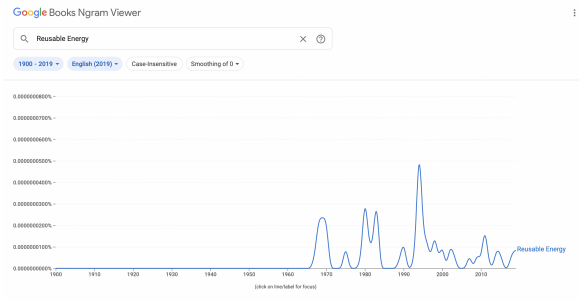

Figure 6: Q2P4 Smoothing=0

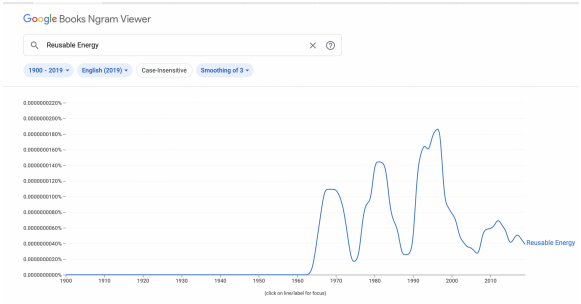

Figure 7: Q2P4 Smoothing=3

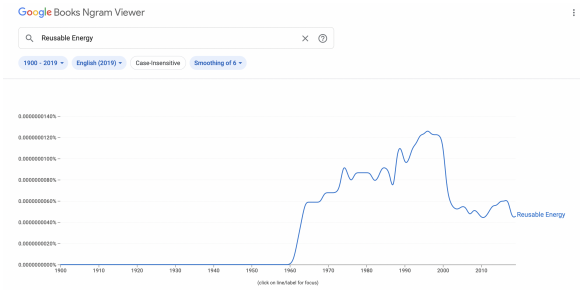

Figure 8: Q2P4 Smoothing=6

over time. Smoothing reusable energy over three years, we more easily see the peaks, while over six years we see it was a prominent term in literature from 1960 to 2000

5. The terms I chose to plot were "Lannister, Greyjoy, Targaryen, Baratheon, Westeros". Each of these terms are surnames, with the exception of Westeros which is the setting, in the series "A Song of Ice and Fire" which was first published in 1996 and brought to prominence by the TV series in 2011. By observing figure 9 we see that prior to the television show, the 'Lannister' characters are the most referenced from publication to air date due to the high volume of 'Lannister' characters. Following airing, we see during the initial year of publication, this term significantly rose in prominence while the other surnames experienced an influx in prevalence. Interestingly we see dips in each term between seasons, and as the relevance of certian characters on the plot became increasingly prominent (Targaryen) and as the world of the show expanded via other novels ("The World of Ice and Fire") references to these characters increased in prominence while references to Lannisters (previously highly central characters but less prominent as the show progressed) fell to

become less prominent than references to the show's overall universe of Westeros. What is particularly interesting to me is how the references to the surnames split in 2012 after previously being relatively equal (with the exception of the Lannisters)

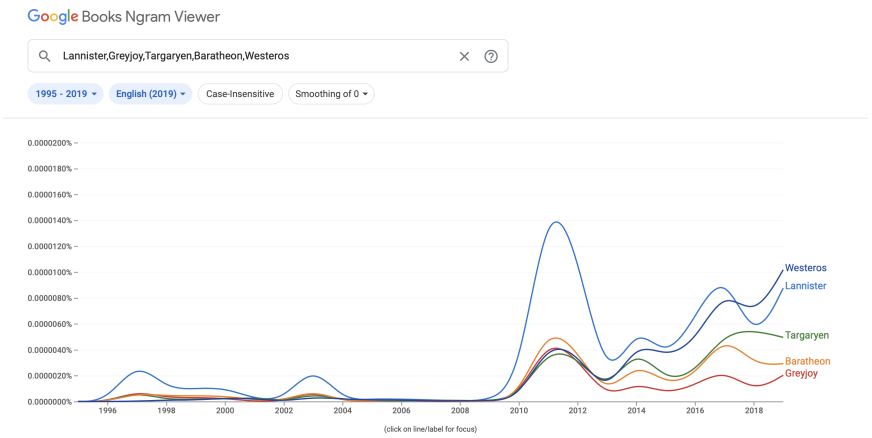

Figure 9: Game of Thrones Characters

6. To use POS tags here for details. I plot the noun 'Tweet' against the verb 'Tweet'. We see that the noun 'tweet', referring to messages on the socia media platform Twitter or the noise of a bird, outpaced the growth of references to the action of tweeting (to post on the platform, or, for a bird, to chirp) as the platform grew, with the rate of growth of the noun rising faster than the verb which flattened around 2014.

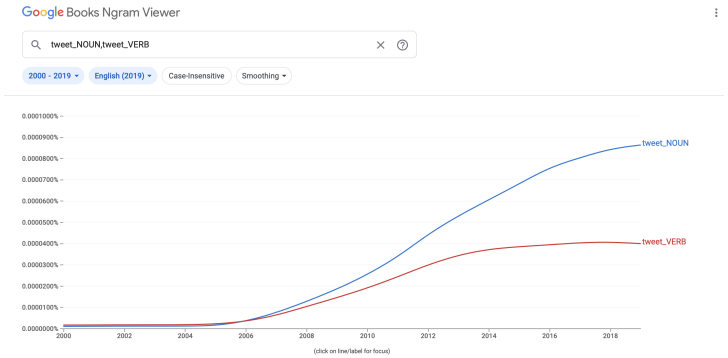

Figure 10: Tweet_Noun vs Tweet_Verb

7. One of the key events which happened over the last 500 years is the Cold War which emerged at the end of the second world war, and a key cultural effect this had was in the creation of terror surrounding communism and socialism in the United States. We see through Google NGram that references to socialist and communist dramatically increase following the end of the second world war, and decline following the end of the Cold War, while references to 'nuclear' similarly spike at a time when the fear of nuclear war between the Soviet Union and US was a pervasive cultural fear; this term in particular reaches its apex with the Chernobyl disaster in the mid-1980s, however, we can clearly see this had spiked during the height of the Cold War. The Space Race was a key component of the Cold War, and we see references to space peak during this time (unsurprisingly with the Moon Landing). Finally, following the fall of the Soviet Union, we see a dramatic decrease in the usage of the word soviet. The key cultural impact of this event (the Cold War) was the impact on the villification of communism and socialism in the United States, combined with the impact this had on fear of nuclear war, and both of these aspects can be clearly seen by their relative emergence towards the end of the second World War, and the subsequent pronounced spikes during the middle of the Cold War.

3 Exercise 3

I generate a dataset using the RANDBETWEEN function in Excel and applying this to each of the 15 words across all years. I then normalise. The resulting table is given in figure 12

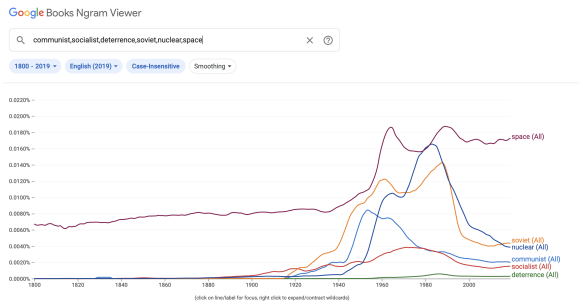

Figure 11: Cold War references

	Raw Data				Formula:	=B3/SUM(B$3:B$17)	=C3/SUM(C$3:C$17)	=D3/SUM(D$3:D$17)	=E1/SUM(E$3:E$17)	=F1/SUM(F$3:F$17)	=SUM(B3)/SUM($B$3:$F$17)	=SUM(C3)/SUM($B$3:$F$17)	=SUM(D3)/SUM($B$3:$F$17)	=SUM(E3)/SUM($B$3:$F$17)	=SUM(F3)/SUM($B$3:$F$17)
Word	**2010**	**2011**	**2012**	**2013**	**2014**	**2010_YearNorm**	**2011_YearNorm**	**2012_YearNorm**	**2013_YearNorm**	**2014_YearNorm**	**2010_OverallNorm**	**2011_OverallNorm**	**2012_OverallNorm**	**2013_OverallNorm**	**2014_OverallNorm**
balloon	958	261	272	1764	703	5%	2%	2%	10%	5%	1%	0%	0%	2%	1%
shark	1680	562	388	1770	356	9%	3%	2%	10%	2%	2%	1%	1%	2%	0%
potato	1205	1030	1204	378	1541	7%	6%	7%	2%	9%	2%	1%	2%	0%	2%
lecturer	351	1085	917	1763	436	2%	6%	5%	10%	2%	0%	1%	1%	2%	1%
crisis	1736	307	1784	726	1442	10%	2%	10%	4%	8%	2%	0%	2%	1%	2%
plane	1250	1371	1086	1395	1103	7%	8%	6%	8%	6%	2%	2%	1%	2%	1%
finance	490	267	740	1972	225	3%	2%	4%	11%	1%	1%	0%	1%	3%	0%
stock	1989	1516	540	960	1475	11%	9%	3%	5%	8%	3%	2%	1%	1%	2%
CRM	804	1835	24	1713	943	5%	10%	0%	10%	5%	1%	2%	0%	2%	1%
crisis	1664	262	1648	468	744	9%	1%	9%	3%	4%	2%	0%	2%	1%	1%
loyalty	1031	1693	1010	1703	721	6%	10%	6%	10%	4%	1%	2%	1%	2%	1%
piano	1558	1564	1419	160	946	9%	9%	8%	1%	5%	2%	2%	2%	0%	1%
dour	43	377	1151	618	1017	0%	2%	6%	3%	6%	0%	0%	1%	1%	1%
good	1750	1397	8	1847	914	10%	8%	0%	10%	5%	2%	2%	0%	2%	1%
sad	1247	1655	1000	952	566	7%	9%	6%	5%	3%	2%	2%	1%	1%	1%

Figure 12: Raw Data And Normalised Data

1. For ByYear Normalisation I divide each entry in the year by the total words in that year (see formula in table). I set these to a percentage for legibility.

2. For Overall Normalisation, I divide each entry in the year by the total sum of words (all raw data entries). I set these to a percentage for legibility.

3. We see in 13 that there are pronounced differences in the figures obtained. Looking at yearly data, and in observing the change by year, By Year Normalisation is a better view of the frequency on a per-year basis if you are interested in examining individual years and comparing year vs year. If you're interested instead in the dataset as a whole time period (2010 to 2014) rather than specific year-by-year relative change (i.e. you're interested in the prevalence of that word, in the year, out of the entire five year time period), using the overall normalisation may be more appropriate.

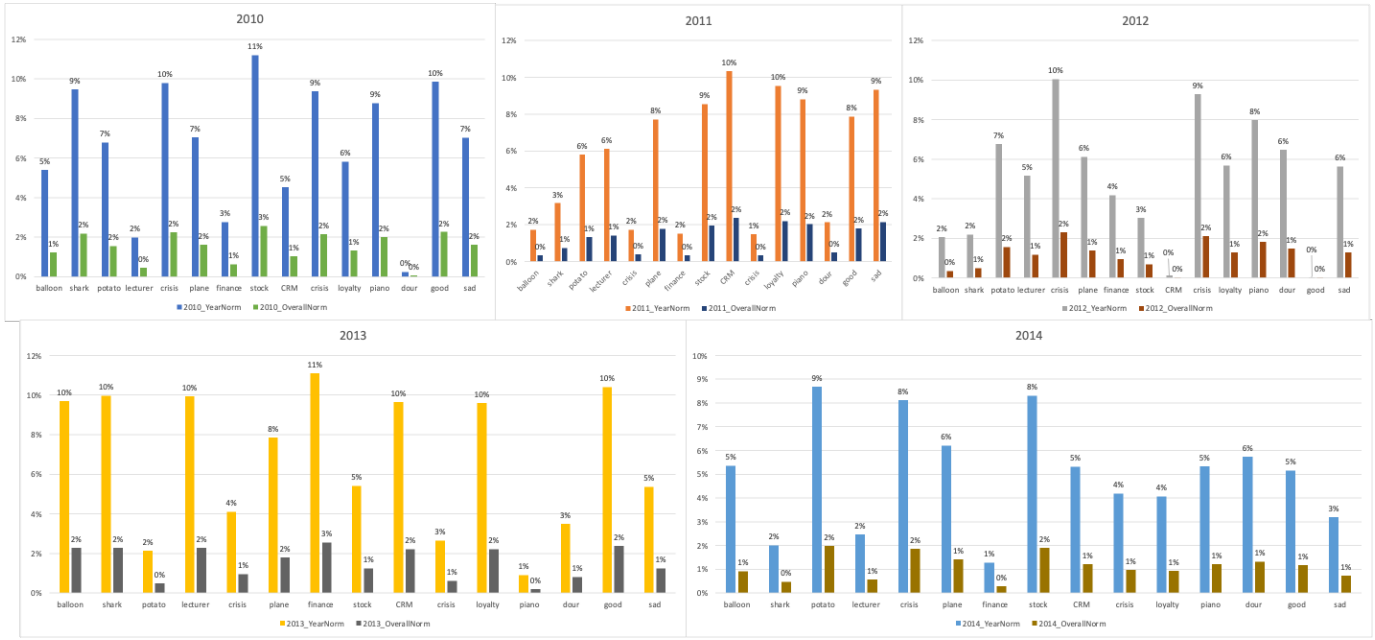

Figure 13: Normalisation By Year