1)

b.

Included: may, children, childrens, thousand, generations, continue, enjoy, benefits, conferred, upon, united, country, cause, yet, rejoice, glorious, institutions, bequeathed, Washington, compeers

Excluded: our, and, to, a, the, us, by, have, under, those, his

Why: Based on my understanding, before feed into wordcloud, the token list has been purified to a list of only key contents (get rid of stop words).

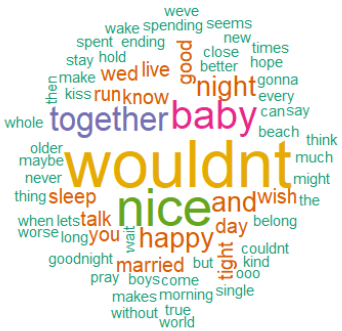
c.

Test list: “Wouldn’t It Be Nice The Beach Boys Wouldn't it be nice if we were older Then we wouldn't have to wait so long? And wouldn't it be nice to live together In the kind of world where we belong?”

Includes: wouldn’t, nice

Excludes: all the other words

My Theory: It seems like the words get excluded are all stop words. However, could be affect by the high frequency of these two words, other words don’t show up on the plot.



d.

Test list: whole lyrics of “Wouldn’t it be nice” from The Beach Boy

While using parameter `colors=brewer.pal(6,"Dark2"), random.order=FALSE` to parsing the whole lyrics, the ‘and’ appears as a component which is unexpected as can be see above left.

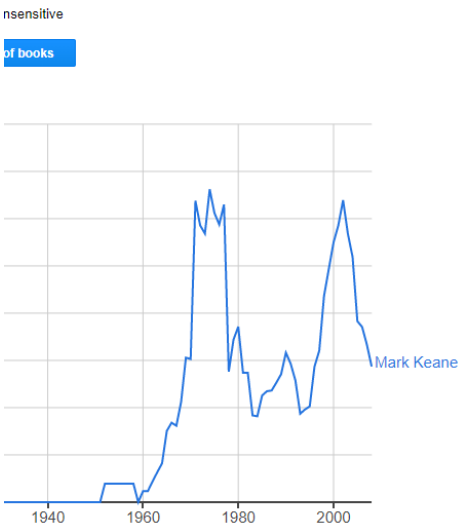
By adding parameter `min.freq=1`, I got the word cloud on the right above, this is a bigger cloud.

I did a little bit research online, in this tutorial, there is a step to remove stop words. So, I guess the automatic removing stop words within wordcloud package could be imperfect.

2)

1

a.



Between 1971 and 1976, Mark Keane had appeared in books most frequently. Fail to find this Mark Keane, but maybe not our professor since he was teenager back then. The second peak is by 2002. This could because of the hurler. The peak around 1990s could be our professor for his contribution on human cognition and artificial intelligence.

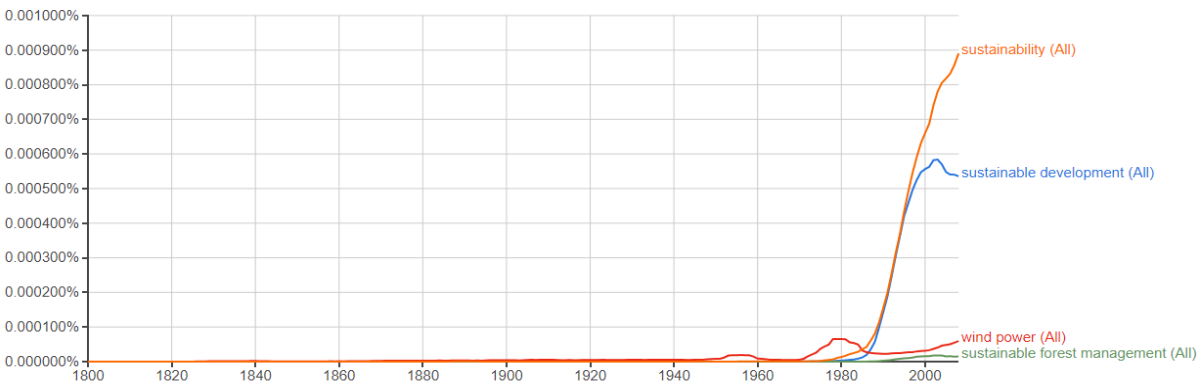
b. When I searched my own name, it returned “No valid ngrams to plot!” which means, the hit hasn’t come yet, but might be in the future. 😊

c. Term: “sustainable development”

In my understanding, this term appeared in the wave of Environmentalism. Start from 1980s. People started to realize pollution destroyed our environment by high speed economic development. Based on this fact, the need to develop economic without depletion of natural resources had appeared more frequently in books.

d. It seems like the ratio has been smoothened by average within a certain period. For example, smoothing of 0, will appear sharp changing each year and smoothing of 3 will change the plot into a smoother curve.

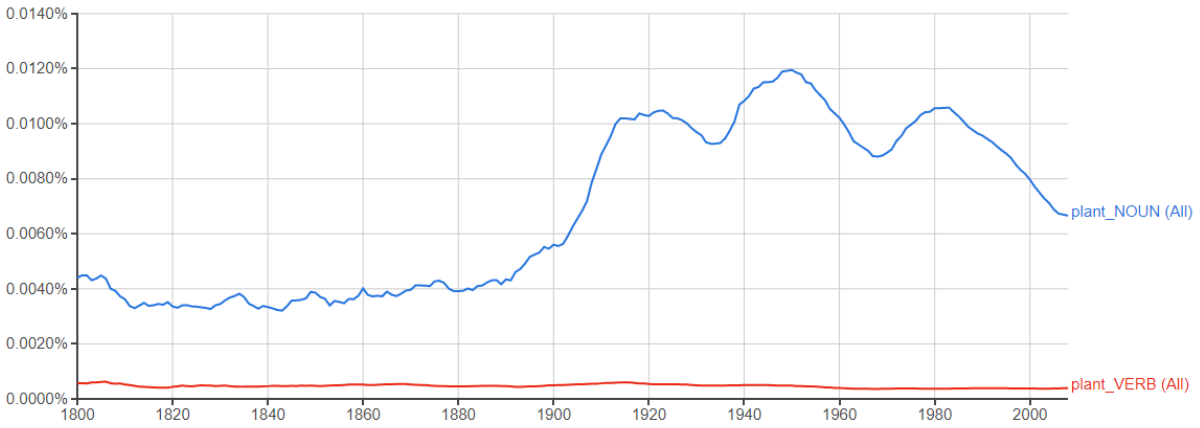
e.



I found the wind power is relatively low in books quite surprising. Wind power is an important part of clean energy, however, obviously it's rarely mentioned in books.

Sustainable development and sustainability are close term, have similar curve while sustainability is more popular. The term sustainable development came from the idea of sustainable forest management. But this term itself never been too popular.

f.



Plant had been used as a noun much more than a verb.

g.

J-pop and K-pop are originally musical genre from Japan and Korea. It started by 1990s and had huge impact to local culture in China. The plot above matched the time and the influence. Obviously, K-pop is less mentioned in books while J-pop is more popular.

3)

a.

words	normalised1			normalised2			difference		
	2010	2011	2012	2010	2011	2012	2010	2011	2012
use	0.02	0.03	0.03	0.054817	0.092633	0.079274	0.69	0.63	0.68
an	0.00	0.03	0.01	0.014554	0.089349	0.021283	0.69	0.63	0.68
excel	0.03	0.03	0.03	0.10956	0.08495	0.092201	0.69	0.63	0.68
spreadsheet	0.00	0.04	0.02	0.005245	0.113018	0.059349	0.69	0.63	0.68
set	0.03	0.04	0.01	0.084072	0.10961	0.024639	0.69	0.63	0.68

up	0.03	0.00	0.04	0.090573	0.010905	0.113341	0.69	0.63	0.68
your	0.04	0.03	0.01	0.142213	0.093252	0.017997	0.69	0.63	0.68
own	0.01	0.01	0.02	0.031841	0.024475	0.075489	0.69	0.63	0.68
list	0.01	0.01	0.02	0.037382	0.03222	0.056992	0.69	0.63	0.68
of	0.01	0.03	0.03	0.039229	0.092633	0.107199	0.69	0.63	0.68
words	0.01	0.03	0.01	0.023493	0.085507	0.03271	0.69	0.63	0.68
and	0.02	0.02	0.03	0.077793	0.057253	0.107985	0.69	0.63	0.68
give	0.01	0.00	0.00	0.036495	0.012144	0.007428	0.69	0.63	0.68
each	0.04	0.02	0.04	0.126773	0.049198	0.112555	0.69	0.63	0.68
a	0.04	0.02	0.03	0.12596	0.052853	0.091558	0.69	0.63	0.68

Based on two ways of normalisation, the normalised word frequencies are different. However, when I'm about to calculate the difference. I wonder what kind of difference can actually reveal the difference. I don't think absolute difference means much since it varies over year and other words' frequency.

Then I think the ratio of change between two methods[1] could make more sense. Surprisingly, the changed percentage is all the same. This make sense. Even though the normalizing base changed from three years to one year, the absolute frequency never changed. Therefore, my conclusion is **these two ways to normalise might have different numeric results, but the relative relationship between words never changed**.

4)

I found the research paper and code in https://static.googleusercontent.com/media/www.google.com/en//googleblogs/pdfs/google_predicting_the_present.pdf .

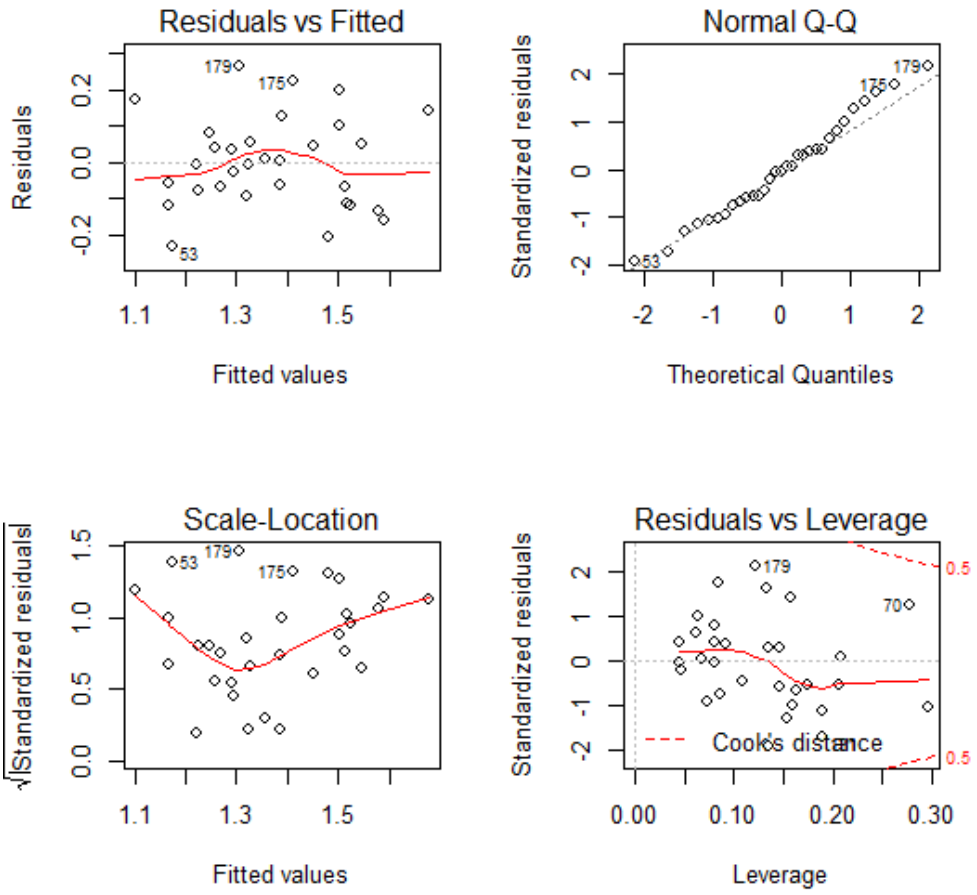
I tried to download the same data from Google Trends and Automotive News to replicate the exact same experiment. However, to download data from Automotive News, subscription is needed. After research, I find data for Ford Fiesta monthly sales from 2015 – 2018 Aug in US[2]. I decide to play around with this dataset to see is the same relationship between Google Trends and current sales exist in sub line of Ford too.

By executing sample code (Figure 1), I got the result as below:

It's pretty similar to original result and I tested 1st week, 2nd week and 3rd week time lag, the one-week lag has the best result.

[1] In my case, I'm using (normalised frequency$_{method2}$ – normalised frequency$_{method\ 1}$) / normalised frequency$_{method\ 2}$.

[2] http://carsalesbase.com/us-car-sales-data/ford/ford-fiesta/



C:\Users\User\Documents\TextAnalytics\Practical\p3\fordFiesta.R - R Editor

```
# Google data
google = read.csv('google.csv');
google$Week = as.Date(google$Week, "%d/%m/%Y");

# CarSalesBase data
dat = read.csv("actual.csv");
# library(zoo);
# dat$Month = as.yearmon(dat$Month, "%Y-%m");
dat$Month = as.Date(paste(dat$Month,"-01",sep=""));

dat = rbind(dat, dat[nrow(dat), ]);
dat[nrow(dat), 'Month'] = as.Date('2018-09-01');
dat[nrow(dat), -1] = rep(NA, ncol(dat)-1);

# Add time lags
dat$sl = c(NA, dat$CarSalesBase[1:(nrow(dat)-1)]);
dat$sl2 = c(rep(NA, 12), dat$CarSalesBase[1:(nrow(dat)-12)]);

# Plot
par(mfrow=c(2,1));
plot(CarSalesBase ~ Month, data= dat, lwd=2, type='l', main='Ford Sales'
plot(trend ~ Week, data= google, lwd=2, type='l', main='Google Trends: F

# Merge data
google$Month = as.Date(paste(substr(google$Week, 1, 7), '01', sep='-'))
```

Figure 1 Snippet of codes