Astronomy & Astrophysics manuscript no. output
April 15, 2020
©ESO 2020

Bayesian Technique for Determining Book Publication Dates

Using Google Ngram Data

T. Dyson[1]

[1]McGill University
e-mail: taj.dyson@mail.mcgill.ca

April 15, 2020

ABSTRACT

Aims. To determine the date of publication for arbitrary collections of words.
Methods. Books are treated as randomly determined collections of words with probability of publication in a given year determined by a multinomial distribution. The coefficients of that distribution are calculated from Google Ngram data. Then, that multinomial likelihood is multiplied by a prior to give the probability distribution of publication as a function of year. Both a uniform prior and an exponential prior (to account for the greater number of publications in recent years) are tried.
Results. For large enough text samples, the prior has little effect, as is hoped. The full texts for two books are analysed. The posteriors have broad peaks, but peaks nonetheless, around the approximate years of publication.
Conclusions. The results are satisfactory for a preliminary exercise. Further refinement can be done in many areas. With better computers, one could use data from many more books, studying many more words. One could also incorporate the study of 2- and 3-grams in a method very similar to this one.

Key words. Bayesian statistics – Text mining

1. Introduction

This project is an exercise in text mining: a marriage of the typically quite separate fields of computer science and literature. In essence, text mining is the act of making a computer read a lot of books for you, not with the goal of learning about the content of the books, but rather about the books themselves.

One such piece of meta-information which can be gleaned through text mining is the date of publication. Given historical information about the usage of words over time, the probability that some input text was written in any given year can be determined. This probability is found in a statistically sound way through the use of Bayes' theorem.

The computational technique discussed here could be used by scholars in possession of some hitherto unknown tome to make a guess at its date of publication.

The python code used in this project can be found in full at: `https://github.com/1sadtrombone/bayesian-1grams`

2. Methods

The meat of this technique is Bayes' theorem, presented below:

$$P(\theta|d) \propto P(d|\theta)P(\theta), \tag{1}$$

where θ is some model parameter, and d is some recorded data. In our case, the parameter θ is year of publication, and d is the words appearing in our input text. The quantity we are after is the posterior $P(\theta|d)$, the probability of publication in some year given the words in the text.

To obtain the posterior, we must calculate the likelihood $P(d|\theta)$, the probability that a text contains a certain collection of words given the year of publication. To calculate this, we first limited our scope only to words appearing the Google Ngram "English 1 Million" dataset from 2009. Next we first assume each book is a "bag of words" each randomly and independently chosen, where the probability of choosing any one word depends on the year. This is a very large assumption which ignores the order of and relations between words, but it will do for a preliminary analysis. We now see books not as sequences of words, but as number of occurrences for each word in the Google dataset. Under these limitations, the probability of some book being written in a given year can be modelled as a multinomial distribution:

$$P(\text{book}|\text{year}) = \frac{n!}{\prod_{i=0}^{n} w_i!} \prod_{i=0}^{n} p_i^{w_i}, \tag{2}$$

where p_i(year) is the probability of choosing the i^{th} word in the database for that year, and w_i is the number of times that word appeared in the book. The p_is must sum to 1. The common analogy the visualise the multinomial distribution is as follows. Say you have a bag full of balls of n different colours, where the probability of picking a ball of the i^{th} colour is p_i. The multinomial distribution gives you the probability that you picked out w_0 balls of the zeroth colour, w_1 balls of the first colour, and so on, after $\sum w_i$ picks, replacing the ball each time. It's exactly that case here, except instead of drawing balls from a bag, the author is putting words on a page, and since they can write the same word as many times as they like, they don't need to put them back. The probability of picking any given word is not due to the number in some bag, but the intangible cultural influences the author is swayed by.

How do we find the probability p_i for some word to be used in a given year? To determine this, we turn to the Google Ngram

"English 1 Million" data mentioned above. The Ngram data has its problems, for instance all books in the database are lumped into one category. So, since there have been a lot more scientific publications recently than erstwhile, non-scientific words, such as "autumn" appear to be on the decline. Since we aren't making any assumptions about the text we're studying, we'll take it as a feature that a text with scientific terms is more likely to be recent. The Ngram data comes in the form of number of occurrences per year for over 3 million (indeed, not 1) 1grams, which for our purposes are whitespace-separated strings of characters – "generalized words" in a sense. To turn these occurrence counts into our p_is, we divide the counts each year by the total counts in that year, leaving us with a p_i for each word for that year. Note that these p_is sum to 1 by definition, and so can be used in the multinomial distribution formula. A side effect of this normalization is that total number of words counted, which is proportional to the number of books published in that year, has no effect on the probability of publication in that year. This lapse in intuition can be rectified via the prior.

The prior $P(\theta)$ is our guess for year of publication knowing nothing about the content of the book. One would expect it to be more likely for a book to be published in a year where more publications occurred. Since the prior should have minimal effect on our results, no great effort was put into researching book publication data, and an arbitrary exponential prior from 1800 to 2008 which looked reasonable was chosen. The upper limit 2008 is imposed by the available data. The lower limit 1800 was chosen to keep the data volume manageable, without sacrificing too much usefulness. Most publications of interest should fall within these bounds. A uniform prior with the same bounds is also used for comparison.

There are some further adjustments that must be made to make the data usable. Since there are so many words, the p_is are tiny, and since books can be quite long, the w_is can get quite large, we have opted to work with the logarithm of the probability. This means that normalization of the probabilities cannot be done in the usual sense of "integrating to one." Instead, probability at each discrete year has been kept between 0 and 1 in the likelihood and prior, so the logarithmic probability is never positive. The multinomial distribution is kept within this range using the combinatoric factor in front, though for large collections of words, such as full novels, the factorials become incalculable and normalization is thrown out the window. Another undesirable, but ultimately necessary consequence of switching to logarithmic probability is we cannot handle a probability of zero. Unfortunately, there are quite a few years where the number of occurrences for some word is zero, so the probability for some book to have that word in it in that year is zero. To get around this inconvenience, I simply add 1 to every count in the Ngram dataset. This way, there are no zero counts. This is not too unreasonable: these are the most common words, so they must have been used at least once a year since 1800. Even "internet" has uses dating back to 1900 before this light tampering.

3. Results

First, we present the results for some "toy" books of our own authorship, for purposes of demonstration. There are two such books. The first goes as follows: "hello internet pogs and the like" and the second goes like this: "hello gentleman farthing and the like". While the publication year of both of these short collections of words is 2020, the first uses words more common in the current era, and will be known from now on as the "new book," while the second uses antiquated terms and will hence-

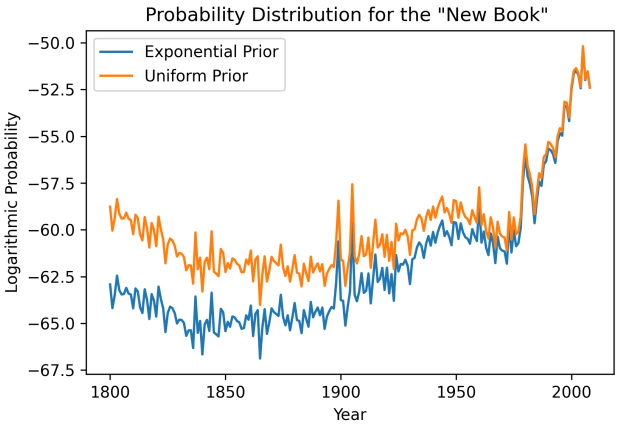

Fig. 1. The probability of publication for the book containing only the words "hello internet pogs and the like," using an exponential and a uniform prior. Note the steady rise as years increase. The spikes near 1900 can be attributed to the word "internet," which saw increased usage in those years.

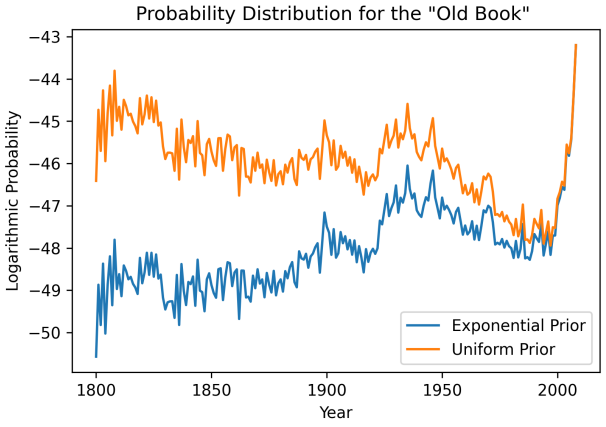

Fig. 2. The probability of publication for the book containing only the words "hello gentleman farthing and the like," using an exponential and a uniform prior. The rapid increase in the last decade or so can be attributed to the resurgence of the word "gentleman" in the dataset studied.

forth be dubbed the "old book." Note that both books include timeless filler words. See Figure 1 for the posterior of the new book, and Figure 2 for the posterior of the old book, both comparing uniform and exponential priors. The change in priors has a large effect on the posteriors for books as short as these.

Now, let's try a real book. The result from loading in the full text of Charles Dickens' "Great Expectations," published in 1861, can be seen in Figure 3. The text was obtained from Project Gutenberg. Note that the prior has little effect on the result for a book with this many words. The posterior peaks quite broadly near the late 1800s, and then spikes again near 2000, for unknown reasons. Percent credible regions could not be found for this logarithmic probability.

The results for another book, "Beyond Light," a pulp sci-fi short story by Nelson S. Bond published in 1940, can be seen in Figure 4. This text was also obtained from Project Gutenberg. Again, the priors have no effect, as they should. The maximum is still broad, though is certainly further in the future than that of "Great Expectations" in Figure 3. The peak after 2000 remains, and is not immediately explainable.

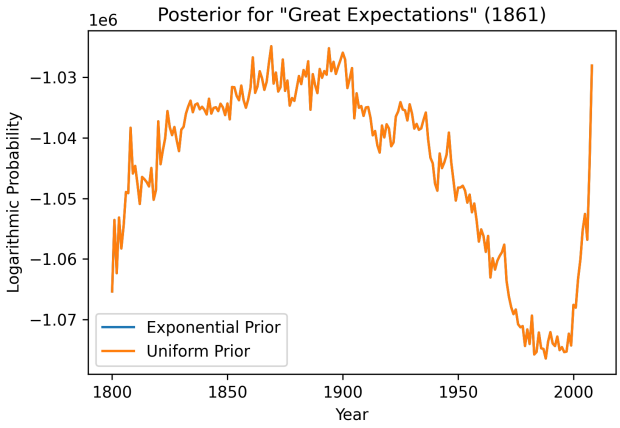

Fig. 3. The probability of publication for "Great Expectations" by Charles Dickens, published in 1861. Note the broad maximum at the late 1800s. The prior has no tangible effect on the posterior for a book with this many words. The spike after 2000 is a mystery.

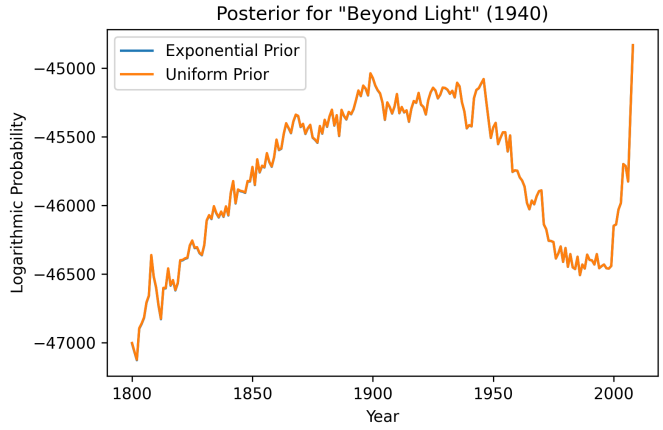

Fig. 4. The probability of publication for "Beyond Light" by Nelson S. Bond, published in 1940. The prior has no tangible effect on the posterior for a book with this many words. The maximum is certainly earlier than that of "Great Expectations" in Figure 3. The spike after 2000 remains a mystery.

4. Conclusions

Using Google Ngram data, under the assumption that books are random collections of words with probabilities dependent on the year, a Bayesian approach can be made to the problem of determining year of publication for an arbitrary text.

The results for the two full books are satisfactory for a preliminary exercise, but the technique can be improved upon in several respects. First, provided access to computers with greater computational power, more words can be considered in the analysis. Additionally, pairs or triplets of words (2- or 3-grams in the parlance) can be incorporated into the analysis. My research team is not going to pursue these paths until at least finals are over.

Acknowledgements. Thank you to Adrian Liu for his helpful correspondences.