Project 2

Peter Liu

2023-10-02

I choose Lovecraft, H. P. (Howard Phillips) who is a famous horror novelist. The five works are “The Shunned House”, “The Dunwich Horror”, “The colour out of space”, “The call of Cthulhu”, “He”.

```
## Determining mirror for Project Gutenberg from https://www.gutenberg.org/robot/harvest
```

```
## Using mirror http://aleph.gutenberg.org
```

The sentiment by page is shown in Figure 1. To be honest, I am not surprised and would applaud the author that did a great job in horror. We can observe that the novels are not too long, but the sentiment is almost always negative. We perform the cumulative sentiment analysis in Figure 2. As we observed in Figure 1, expect for the novel ‘He’ (a very short one, early work of Lovecraft and originally written in Spanish), the rest of the novels shows rather great horror as story progresses, and the cummulative sentiment seems to be monetone decreasing.

Figure 1: We can observe that the novels are not too long, but the sentiment is almost always negative. This implies the monetone cumulate sentiment in Figure 2.

For the other authors, I choose the first one as “United States” with works “The United States Constitution”, “1995 United States Congressional Address Book”, “Copyright Law of the United States of America in Title 17 of the United States Code”, “Copyright Law of the United States of America and Related Laws Contained

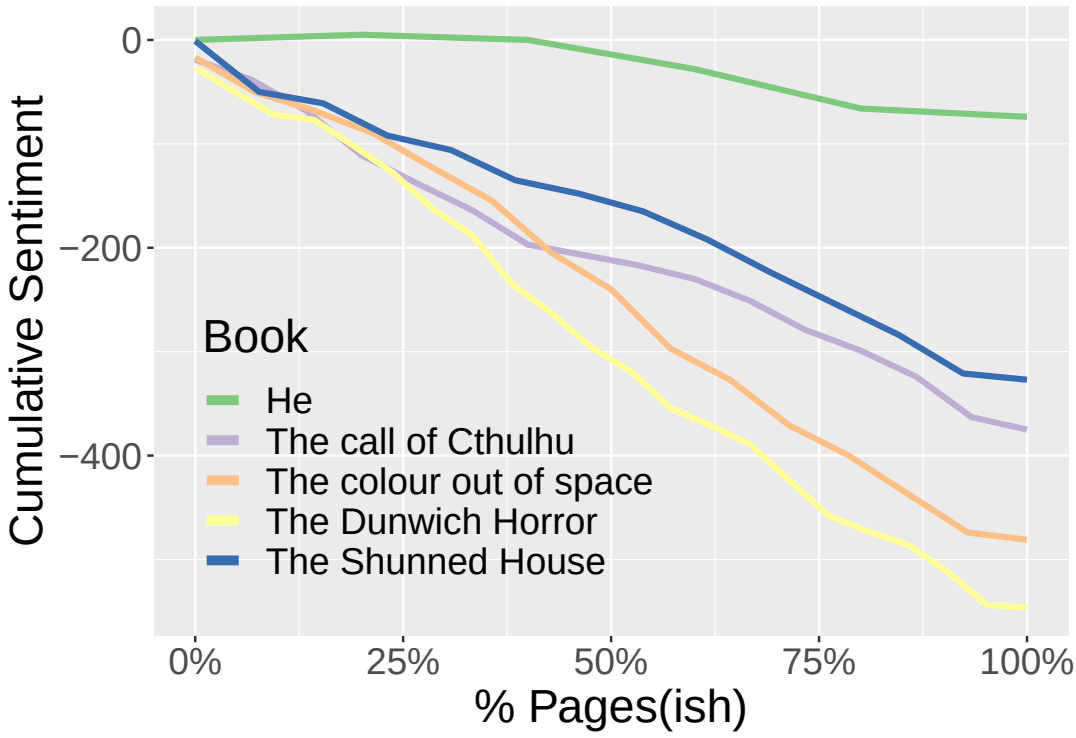

Figure 2: Expect for the novel 'He' (a very short one, early work of Lovecraft and originally written in Spanish), the rest of the novels shows rather great horror as story progresses.

in Title 17 of the United States Code, Circular 92", "Amendments to the United States Constitution". For the second one I choose Shakespeare with works "History of King Henry the Sixth, Second Part", "The History of King Henry the Sixth, Third Part", "The Tragedy of King Richard III", "The Comedy of Errors", "The Rape of Lucrece". For simplicity, I will abbreviate Lovecraft with "LH", United States as "US", abd Shakespeare as "SH".

```
## Warning: ! Could not download a book at
##   http://aleph.gutenberg.org/1/9/5/8/19581/19581.zip.
## i The book may have been archived.
## i Alternatively, You may need to select a different mirror.
## > See https://www.gutenberg.org/MIRRORS.ALL for options.
```

```
## Warning: Unknown or uninitialised column: `text`.
```

We give the highest frequency words by author by book in the three tables in the Appendix. From top to bottom, we have LH, US, and SH.

We perform LDA on the combined data set. We first use three topics. A first thing I would admit is that I observed a lot of numbers in the US works. This is not surprising, as the works are law and regulation related. Yet too many numbers cause problem, i.e. there are too many section indexes as "1", "2", which really hinders the topic analysis and make the topics almost identical. I did an (almost) post-selection to get rid of the section numbers, but I kept some of the law/act/bill numbers , as in the analysis if there is a topic related to law and politics, these numbers can serve as indicator of the underlying political sentiment. Second, I notice that since SH wrote in ancient English, "thee","thy", etc. are not removed in stopping word and I have to remove it mannually. After the cleaning above, I first calculate how each author corresponds to each topic. I calculate the cummualtive topic beta values for each word from the author (counting repeatedly by occurence). Also, to allow comparison I also normalize the score with the length of the author's text. I plot the results in Figure 3. As we can observe there is a distinct pattern for each of the author/topic, with

LH on 1st topic, SH on 3rd topic, and US ont 2nd topic.

We then look into how exactly the topic differs and what they corresponds to. To this end I plot the word beta values by each topic. Not surprising, topic 1 corresponds to words mostly appears in LH, such as "strange", "night", "horror", etc. I will denote the 1st topic as horror. Topic 2 corresponds to words mostly appears in US, such as "section", "copyright", "act", etc. I will denote the 2nd topic as law. Topic 3 corresponds to words mostly appears in SH, such as "king", "lord", "queen", etc. I will denote the 3rd topic as royal.

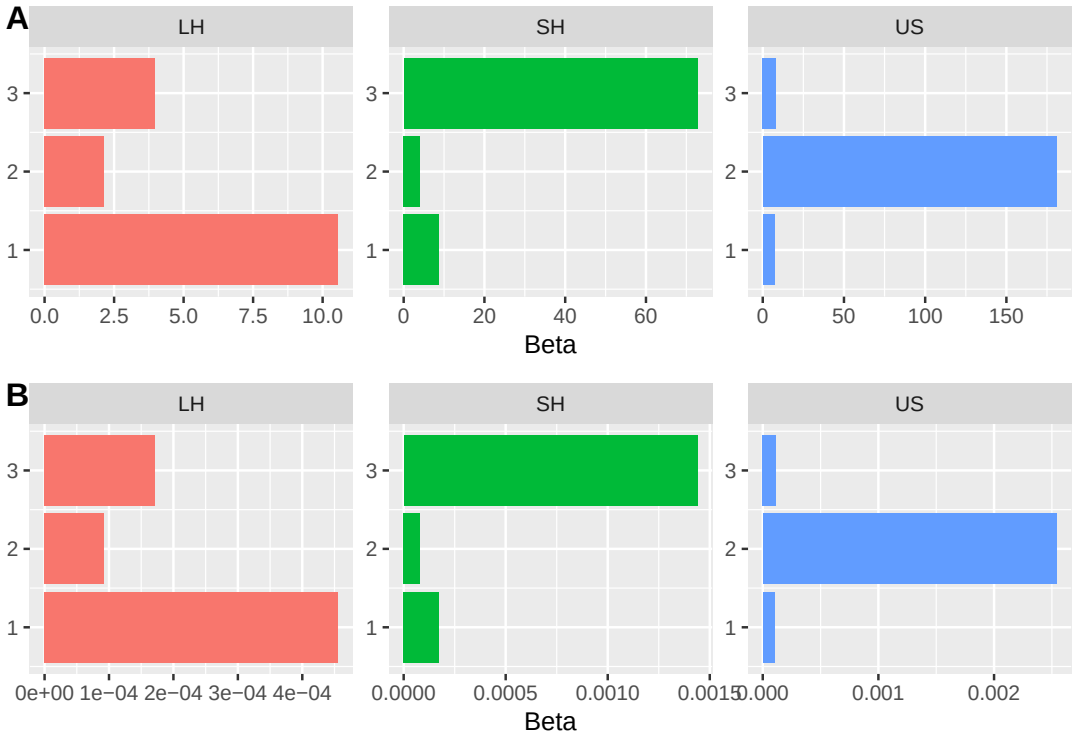

Figure 3: A: Cummulative topic for the three authors. B: Normalized topic for the three authors. We observe that the trend is consistent with or without normalization. Also, the LDA separates the three authors, with LH on 1st topic, SH on 3rd topic, and US ont 2nd topic.

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i'll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i'll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i'll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i'll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i'll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i'll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i'll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
```

```
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>
```

I then analyze the contrast between the topics. I let the first ratio be topic2/topic1, second be topic3/ topic1, and third topic3/topic2. I plot the results in Figure 5. It is clear that the US (represented by the numbers) shows great positive difference in ratio 1, which is as expect, relatively less difference in ratio 2 but still positive, which make sense as some of the “royal” term such as lord can appear in law & government related work, and has no difference in ratio 3. Similarly, LH shows great negative difference in ratio 1 and 2 as expected, but also great postive difference in ratio 3. This might resort to the fact that some of the greatest horrors in the novel are addressed as lord. Lastly, SH has no difference in ratio 1, has great difference in ratio 2 and 3. To conclude, we would argue (only heuristically) that the topic law and horror are relatively independent, and topic royal roughly bridges the two topics.

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>
```

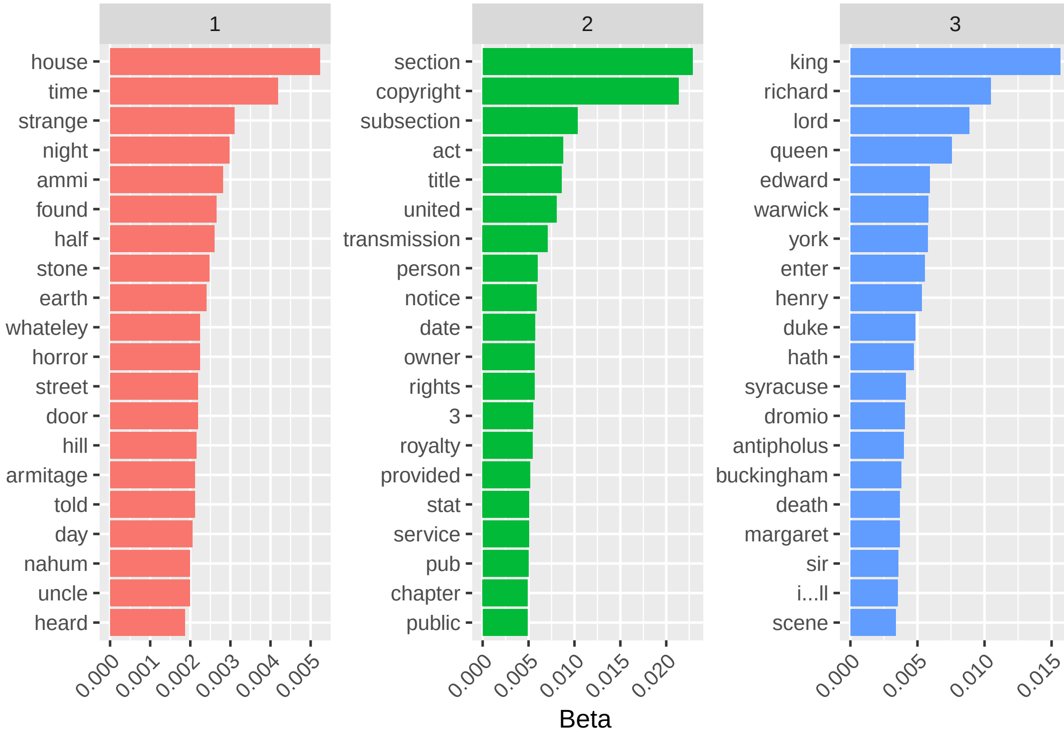

Figure 4: Word by the three topics. Once again, and consistent as in previous figure, 1st topic contains words mostly from LH, 2nd topic contains words mostly from US, and 3rd topic contains words mostly from SH

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
```

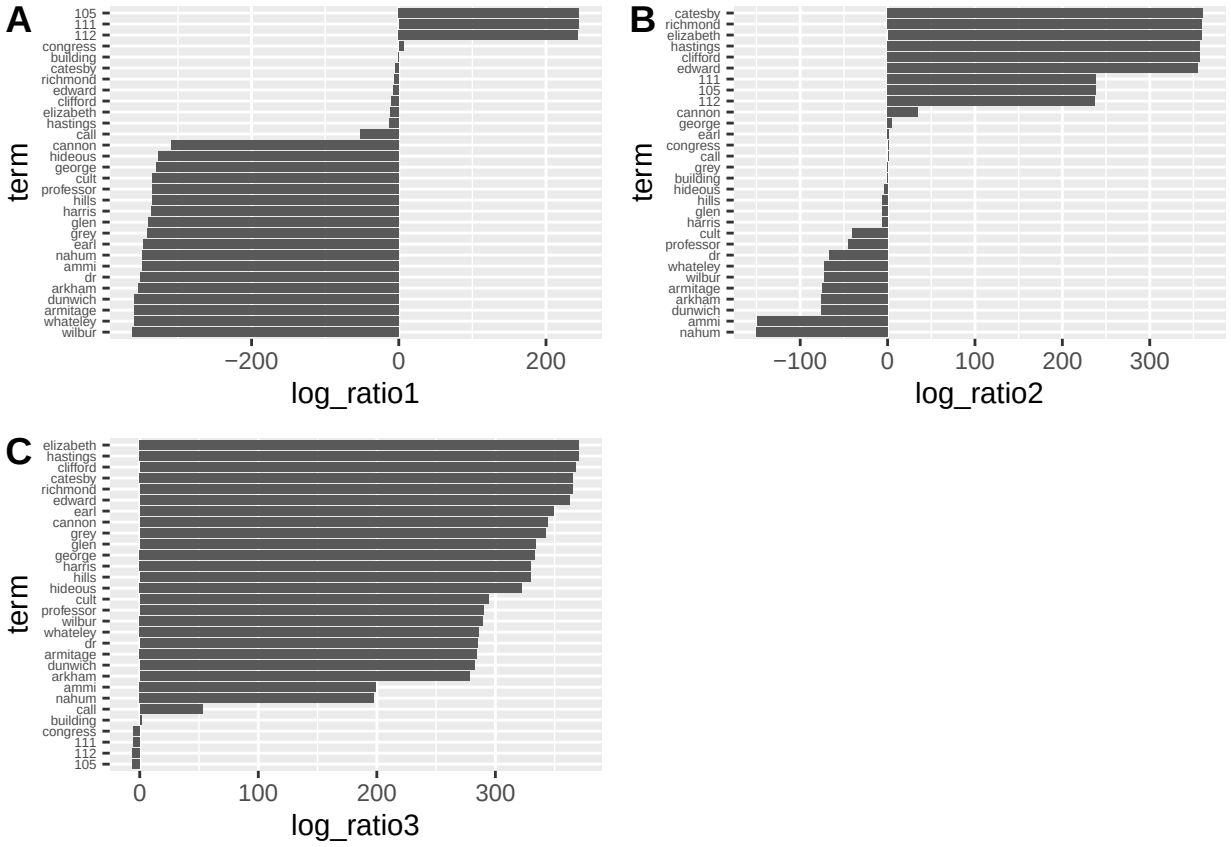

Figure 5: A: difference with ratio = topic2/topic1; B: difference with ratio = topic3/topic1; A: difference with ratio = topic3/topic2.

```
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i’ll' in 'mbcsToSbcs': dot substituted for <e2>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i'll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i'll' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i'll' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i'll' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'i'll' in 'mbcsToSbcs': dot substituted for <99>
```

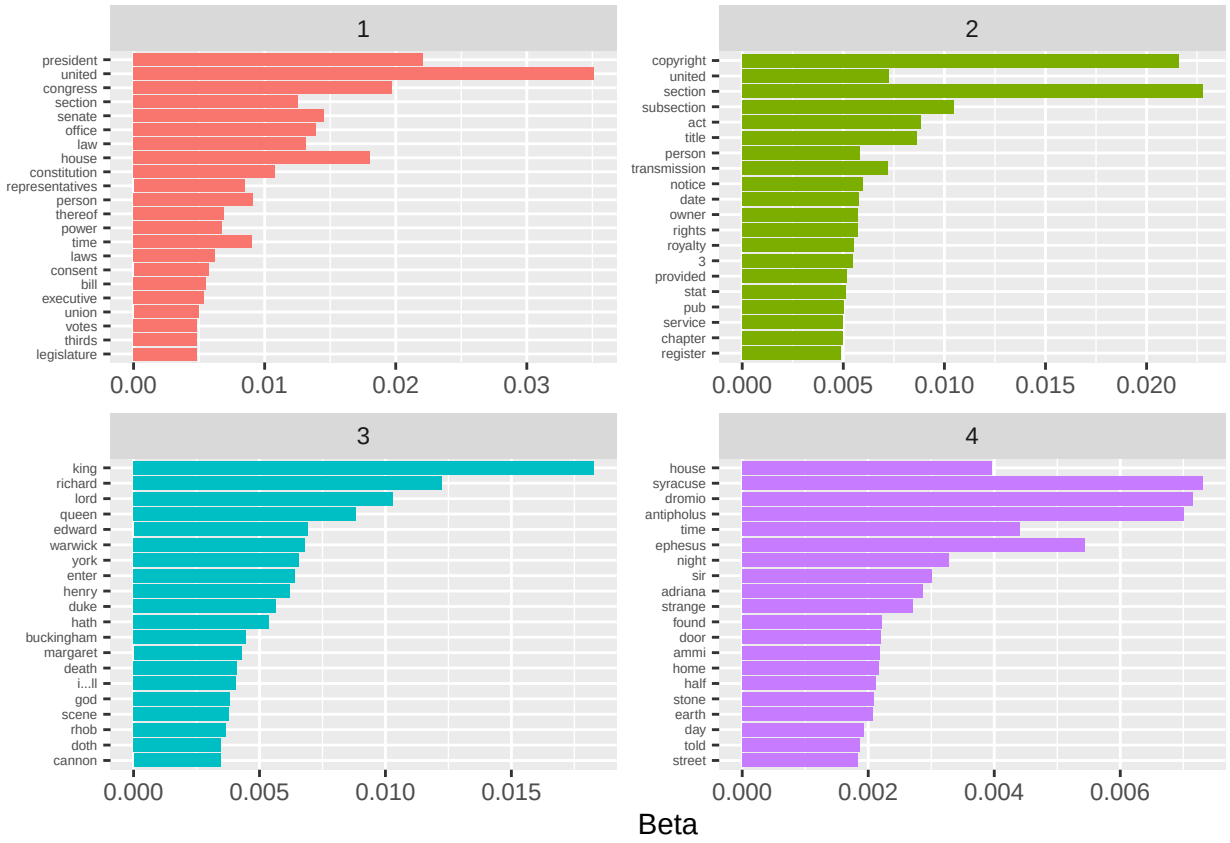

Figure 6: We now use four topics for LDA analysis. We observe that the topic law is separated in two: the new topic 1 and 2, one more on the government side and another more on copyright. The new topic 3 and 4 corresponds to royal and horror which seems to change very little.

I then tried for the four topic analysis. We observe that topic law is further separated into two: the new topic 1 and 2, one more on the government side and another more on copyright. The new topic 3 and 4 corresponds to royal and horror which seems to change very little. The difference patterns consists with the three topic case but seems to be more clear. To be more specfic, we observe postive difference in C, E, F, all involves with log ratio with topic 4 in the nominator, etc.

I think the choice of topic number is very similar as in PCA or hierarchical clustering - if you know the ground truth (i.e. if you know the books in authors are consistent and the authors are different) I would suggest sticking with the number of author. If you don't know the ground truth, or the authors are very similar, I would say using a forward selection starting with 2 topics and incearse the topic number to see if

there are any differences. Note that, when trying for more topics, the difference between topics and authors becomes more clear compared to the 3 authors case. Yet still, too many topics will make analysis difficult, as the comparison plot is in n choose 2 and grows in quadratic. I don't think the difference plots are a good way of doing topic classification - some loss functions or goodness-of-fit method should works better. I think there should be methods like these already developed, but I am new to the field and not so sure.

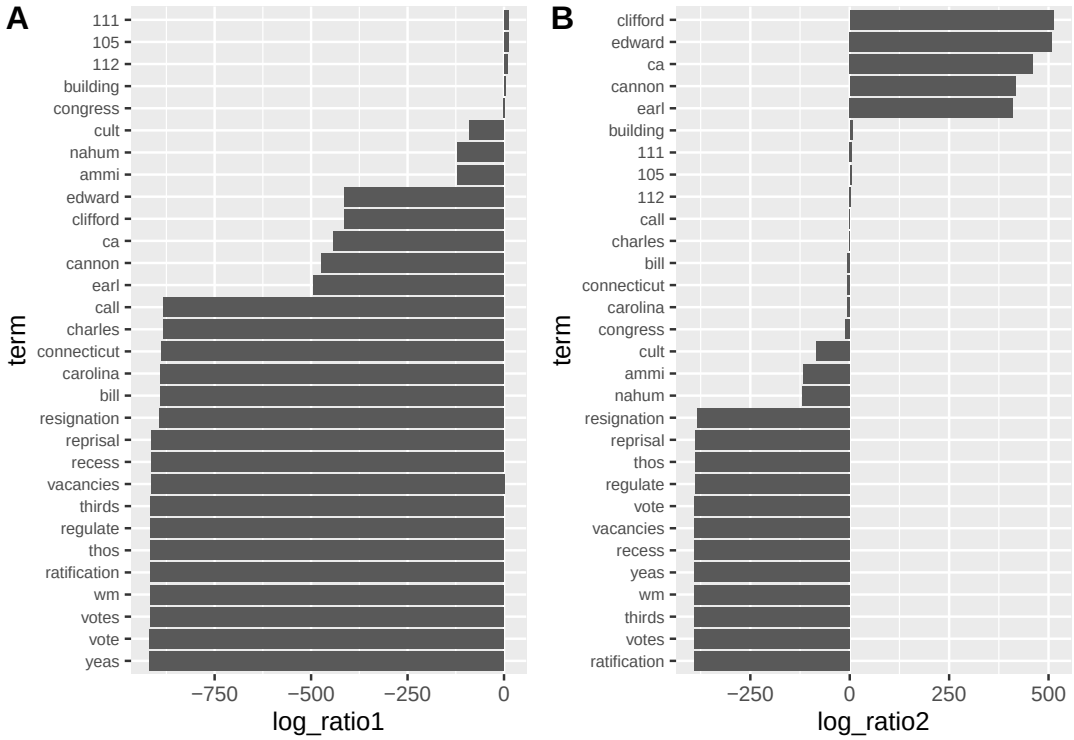

Figure 7: A: difference with ratio = topic2/topic1; B: difference with ratio = topic3/topic1.

```
cat('\\pagebreak')
```

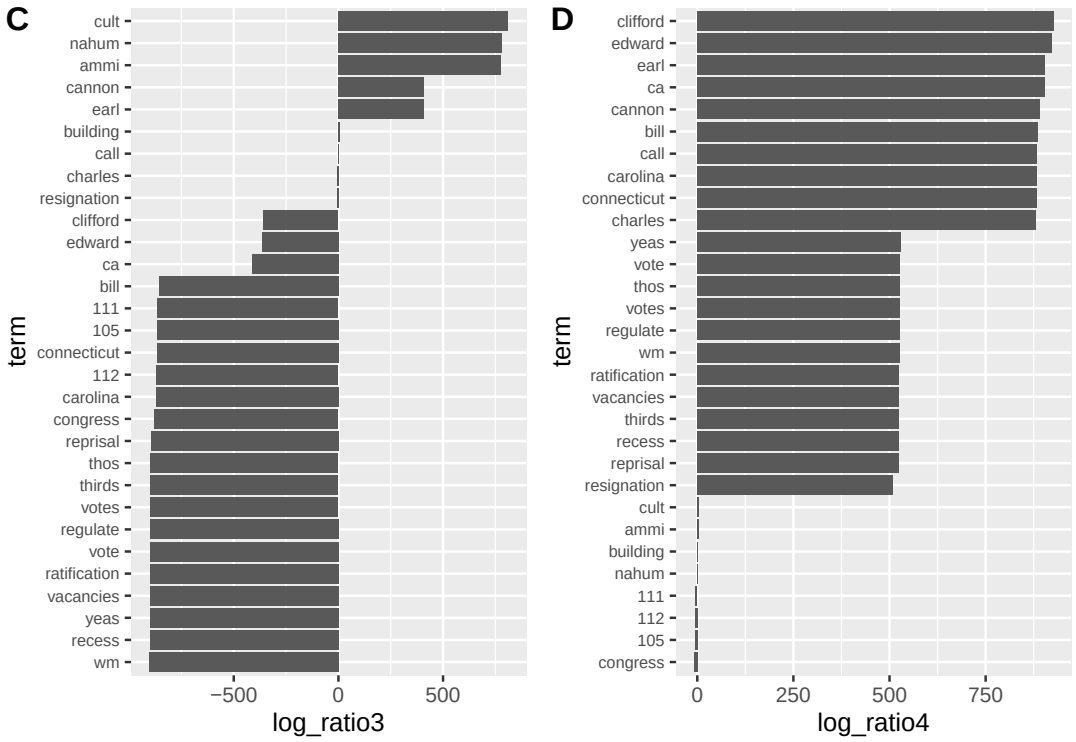

Figure 8: C: difference with ratio = topic4/topic1; D: difference with ratio = topic3/topic2.

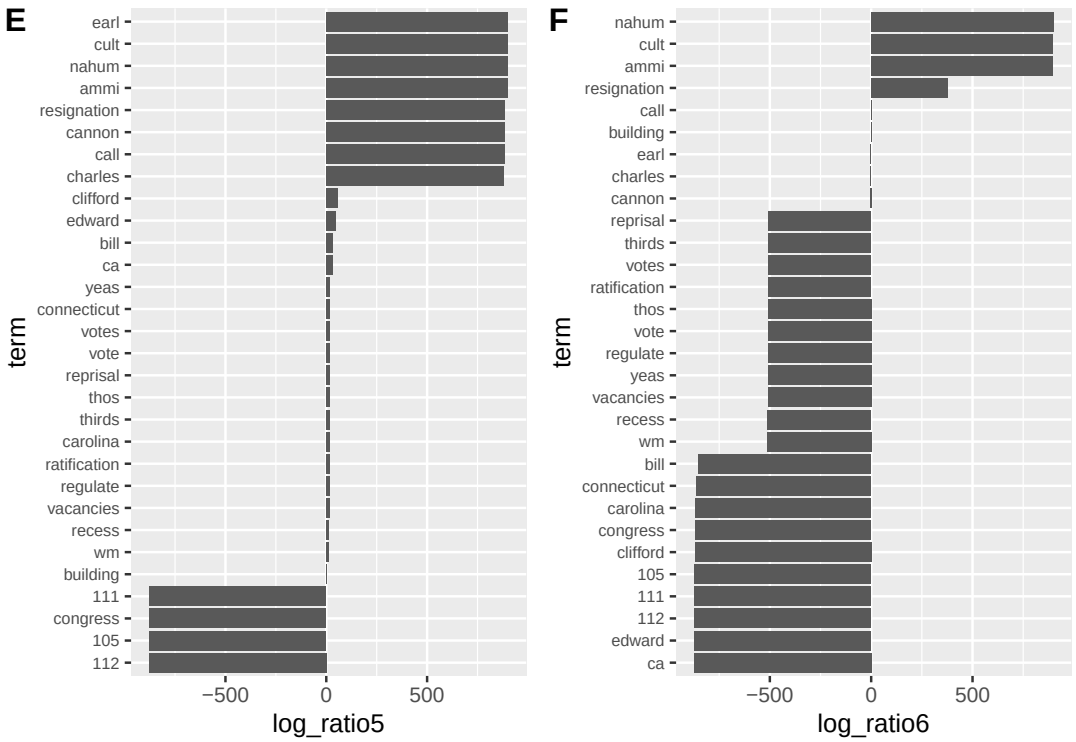

Figure 9: E: difference with ratio = topic4/topic2; F: difference with ratio = topic4/topic3.

Appendix

book	word	n	sentiment
The colour out of space	ammi	68	NA
The Shunned House	house	61	NA
The Dunwich Horror	whateley	54	NA
The Dunwich Horror	armitage	51	NA
The colour out of space	nahum	48	NA
The Shunned House	street	42	NA
The Dunwich Horror	wilbur	41	NA
The Dunwich Horror	dunwich	38	NA
The call of Cthulhu	cult	35	NA
The Dunwich Horror	hill	34	NA

book	word	n	sentiment
Title 17, Circular 92	section	1190	NA
Title 17, Circular 92	copyright	1018	NA
Title 17, Circular 92	subsection	550	NA
Title 17, Circular 92	act	544	NA
Title 17, Circular 92	title	459	NA
Title 17, Circular 92	united	389	NA
Copyright Law of the United States of America Contained in Title 17 of the United States Code	copyright	388	NA
Title 17, Circular 92	transmission	358	NA
Title 17, Circular 92	stat	332	NA
Title 17, Circular 92	pub	328	NA

book	word	n	sentiment
The Tragedy of King Richard III	richard	380	NA
The History of King Henry the Sixth, Third Part	king	364	NA
The Tragedy of King Richard III	king	268	NA
The History of King Henry the Sixth, Third Part	edward	246	NA
The History of King Henry the Sixth, Third Part	warwick	239	NA
The Tragedy of King Richard III	lord	232	NA
The Comedy of Errors	syracuse	227	NA
The Comedy of Errors	dromio	222	NA
The Comedy of Errors	antipholus	218	NA
History of King Henry the Sixth, Second Part	king	216	NA