STAT 482 Project

Hidden Markov Chains in Dickens' Novels

Submitted to
Dr. Olga Korosteleva

Report Prepared By
Ethan Huang

November 29, 2022

Table of Contents

I. Objective

As I was looking through the course textbook for project ideas, I stumbled upon Application 1.1, Markov's initial application of transition probabilities to the first 20,000 letters of Alexander Pushkin's novel *Eugene Onegin*. For this project, I apply Markov's analysis toward Charles Dickens' novels to compute the transition probability matrices. I then compare them to Moby *Dick* (entire novel) and Markov's analysis of *Eugene Onegin*. In doing so, I hope to understand Dickens' signature pattern of vowel/consonant transitions.

II. The Data

I used Project Gutenberg to find full texts of A Tale of Two Cities, Great Expectations, A Christmas Carol, and Moby Dick. I copied and pasted the full texts onto a notepad text file, excluding the chapter titles. Afterwards, I used R to remove all punctuation, numbers, and stray markings. I then broke the clean string into letters and shifted the text to the right.

III. Results

To compute the transition probability matrix, I first defined the vowels, consonants, and the patterns (vv,cc,cv,vc). Then, I used the following formula to get my results:

```
matrix(c((sum(vv)/(sum(vv)+sum(vc))), (sum(vc)/(sum(vv)+sum(vc))),
(sum(cv)/(sum(cv)+sum(cc))),
(sum(cc)/(sum(cv)+sum(cc))), nrow=2, ncol=2, byrow=TRUE))
```

A Tale of Two Cities

	[v]	[c]
[v]	0.1464071	0.8535929
[c]	0.5206246	0.4793754

Great Expectations

	[v]	[c]
[v]	0.1509735	0.8490265
[c]	0.5243050	0.4756950

A Christmas Carol

	[v]	[c]
[v]	0.1503024	0.8496976
[c]	0.5112999	0.4887001

Moby Dick

	[v]	[c]
[v]	0.1422454	0.8577546
[c]	0.5133822	0.4866178

IV. Conclusion

We notice that the transition probability matrices for the Dickens' novels are very similar to each other. When I compare these three novels to Moby Dick, the matrices are similar. However, Markov's transition matrix for *Eugene Onegin* significantly differs from the rest of the novels as shown below.

Markov's Analysis of Eugene Onegin

	[v]	[c]
[v]	0.1278	0.8722
[c]	0.6631	0.3369

Source: (Korosteleva 29)

We can conclude that Dickens' and Melville's signature patterns are different from that for Pushkin. However, a possible explanation to this pertains to language differences between English and Russian. This is as each language has differences in word and sentence patterns.

Appendix

A Tale of Two Cities

```
library(tidyverse)

library(gsubfn)

ATOTC <- read_file("C://Users//colle//Downloads//TaleofTwoCities.txt")

lowercase <- tolower(ATOTC)

omitblanks<- gsub(" ","",lowercase)

omitlinebreaks <- gsub("\r\n", "", omitblanks)

omitdashes <- gsub("-","",omitlinebreaks)

omitdashes2 <- gsub("—","",omitdashes)

omitapostrophe <- gsub("’","",omitdashes2)

omitapostrophe2 <- gsub("‘","",omitapostrophe)

omitleftquotation <- gsub("“","", omitapostrophe2)

omitrightquotation <- gsub("”","",omitleftquotation)

NoPunctuation <- gsub("[[:punct:]]","",omitrightquotation)

clean.string <- gsub("[0-9]","",NoPunctuation)


#We then shift the text

x2<- strsplit(clean.string, "")

no.last<- substr(clean.string, 1, nchar(clean.string)-1)

first.blank<- str_c(" ", no.last)
```

```
x1<- strsplit(first.blank,"")


#We define vowels and consonants
vowels<-c("a","e","i","o","u")
consonants<- c("b","c","d","f","g","h","j","k","l","m","n","p","q","r","s","t",
               "v","w","x","y","z")


#We set the patterns
for (counter in 1:nchar(x2)){
  v<- ifelse(x2[[counter]] %in% vowels,1,0)
  c<- ifelse(x2[[counter]] %in% consonants,1,0)
  vv<- ifelse(x1[[counter]] %in% vowels & x2[[counter]] %in% vowels,1,0)
  vc<- ifelse(x1[[counter]] %in% vowels & x2[[counter]] %in% consonants,1,0)
  cv<- ifelse(x1[[counter]] %in% consonants & x2[[counter]] %in% vowels,1,0)
  cc<- ifelse(x1[[counter]] %in% consonants & x2[[counter]] %in%
consonants,1,0)
}


library(markovchain)
#We set up a transition matrix
transitionATOTC<- matrix(c((sum(vv)/(sum(vv)+sum(vc))),
(sum(vc)/(sum(vv)+sum(vc))),
```

```
(sum(cv)/(sum(cv)+sum(cc))),

(sum(cc)/(sum(cv)+sum(cc))), nrow=2, ncol=2, byrow=TRUE))
```

```
            [v]         [c]
[v]  0.1464071   0.8535929
[c]  0.5206246   0.4793754
```

Great Expectations

```
library(tidyverse)

library(gsubfn)

GreatExpectations <-
read_file("C://Users//colle//Downloads//GreatExpectations.txt")

lowercase <- tolower(GreatExpectations)


NoLetters <- gsub("[[:alpha:]]","",lowercase)

omitblanks<- gsub(" ","",NoLetters)

omitlinebreaks <- gsub("\r\n", "", omitblanks)

omitdashes <- gsub("-","",omitlinebreaks)

omitdashes2 <- gsub("—","",omitdashes)

omitapostrophe <- gsub("’","",omitdashes2)

omitapostrophe2 <- gsub("‘","",omitapostrophe)

omitleftquotation <- gsub("“","", omitapostrophe2)

omitrightquotation <- gsub("”","",omitleftquotation)
```

```
NoPunctuation <- gsub("[[:punct:]]","",omitrightquotation)

clean.string <- gsub("[0-9]","",NoPunctuation)


#We get rid of blanks, line breaks,dashes, apostrophe, quotation marks, rest punctuation

lowercase <- tolower(GreatExpectations)

omitblanks<- gsub(" ","",lowercase)

omitlinebreaks <- gsub("\r\n", "", omitblanks)

omitdashes <- gsub("-","",omitlinebreaks)

omitdashes2 <- gsub("—","",omitdashes)

omitapostrophe <- gsub("’","",omitdashes2)

omitapostrophe2 <- gsub("‘","",omitapostrophe)

omitleftquotation <- gsub("“","", omitapostrophe2)

omitrightquotation <- gsub("”","",omitleftquotation)

NoPunctuation <- gsub("[[:punct:]]","",omitrightquotation)

clean.string <- gsub("[0-9]","",NoPunctuation)


#We then shift the text

x2<- strsplit(clean.string, "")

no.last<- substr(clean.string, 1, nchar(clean.string)-1)

first.blank<- str_c(" ", no.last)

x1<- strsplit(first.blank,"")
```

```
#We define vowels and consonants
vowels<-c("a","e","i","o","u")
consonants<- c("b","c","d","f","g","h","j","k","l","m","n","p","q","r","s","t",
               "v","w","x","y","z")


#We set the patterns
for (counter in 1:nchar(x2)){
  v<- ifelse(x2[[counter]] %in% vowels,1,0)
  c<- ifelse(x2[[counter]] %in% consonants,1,0)
  vv<- ifelse(x1[[counter]] %in% vowels & x2[[counter]] %in% vowels,1,0)
  vc<- ifelse(x1[[counter]] %in% vowels & x2[[counter]] %in% consonants,1,0)
  cv<- ifelse(x1[[counter]] %in% consonants & x2[[counter]] %in% vowels,1,0)
  cc<- ifelse(x1[[counter]] %in% consonants & x2[[counter]] %in%
consonants,1,0)
}
library(markovchain)
#We set up a transition matrix
transitionGreatExpectations<- matrix(c((sum(vv)/(sum(vv)+sum(vc))),
(sum(vc)/(sum(vv)+sum(vc))),
(sum(cv)/(sum(cv)+sum(cc))),
(sum(cc)/(sum(cv)+sum(cc))), nrow=2, ncol=2, byrow=TRUE))
```

```
            [v]       [c]
[v] 0.1509735 0.8490265

[c] 0.5243050 0.4756950
```

A Christmas Carol

```
library(tidyverse)

library(gsubfn)

XmasCarol <- read_file("C://Users//colle//Downloads//ChristmasCarol.txt")

lowercase <- tolower(XmasCarol)

NoLetters <- gsub("[[:alpha:]]","",lowercase)

omitblanks<- gsub(" ","",NoLetters)

omitlinebreaks <- gsub("\r\n", "", omitblanks)

omitdashes <- gsub("-","",omitlinebreaks)

omitdashes2 <- gsub("—","",omitdashes)

omitapostrophe <- gsub("’","",omitdashes2)

omitapostrophe2 <- gsub("‘","",omitapostrophe)

omitleftquotation <- gsub("“","", omitapostrophe2)

omitrightquotation <- gsub("”","",omitleftquotation)

NoPunctuation <- gsub("[[:punct:]]","",omitrightquotation)

clean.string <- gsub("[0-9]","",NoPunctuation)


#We get rid of blanks, line breaks,dashes, apostrophe, quotation marks, rest
punctuation
```

```
lowercase <- tolower(XmasCarol)

omitblanks<- gsub(" ","",lowercase)

omitlinebreaks <- gsub("\r\n", "", omitblanks)

omitdashes <- gsub("-","",omitlinebreaks)

omitdashes2 <- gsub("—","",omitdashes)

omitapostrophe <- gsub("’","",omitdashes2)

omitapostrophe2 <- gsub("‘","",omitapostrophe)

omitleftquotation <- gsub("“","", omitapostrophe2)

omitrightquotation <- gsub("”","",omitleftquotation)

NoPunctuation <- gsub("[[:punct:]]","",omitrightquotation)

clean.string <- gsub("[0-9]","",NoPunctuation)


#We then shift the text

x2<- strsplit(clean.string, "")

no.last<- substr(clean.string, 1, nchar(clean.string)-1)

first.blank<- str_c(" ", no.last)

x1<- strsplit(first.blank,"")


vowels<-c("a","e","i","o","u")

consonants<- c("b","c","d","f","g","h","j","k","l","m","n","p","q","r","s","t",

              "v","w","x","y","z")
```

```
#We set the patterns

for (counter in 1:nchar(x2)){

  v<- ifelse(x2[[counter]] %in% vowels,1,0)

  c<- ifelse(x2[[counter]] %in% consonants,1,0)

  vv<- ifelse(x1[[counter]] %in% vowels & x2[[counter]] %in% vowels,1,0)

  vc<- ifelse(x1[[counter]] %in% vowels & x2[[counter]] %in% consonants,1,0)

  cv<- ifelse(x1[[counter]] %in% consonants & x2[[counter]] %in% vowels,1,0)

  cc<- ifelse(x1[[counter]] %in% consonants & x2[[counter]] %in%
consonants,1,0)

}

library(markovchain)

transitionXmasCarol<- matrix(c((sum(vv)/(sum(vv)+sum(vc))),

(sum(vc)/(sum(vv)+sum(vc))),

(sum(cv)/(sum(cv)+sum(cc))),

(sum(cc)/(sum(cv)+sum(cc))), nrow=2, ncol=2, byrow=TRUE))
```

```
             [v]          [c]
[v]  0.1503024    0.8496976

[c]  0.5112999    0.4887001
```

Moby Dick

```
library(tidyverse)

library(gsubfn)


MobyDick <- read_file("C://Users//colle//Downloads//MobyDick.txt")

lowercase <- tolower(MobyDick)

omitblanks<- gsub(" ","",lowercase)

omitlinebreaks <- gsub("\r\n", "", omitblanks)

omitdashes <- gsub("-","",omitlinebreaks)

omitdashes2 <- gsub("—","",omitdashes)

omitapostrophe <- gsub("’","",omitdashes2)

omitapostrophe2 <- gsub("‘","",omitapostrophe)

omitleftquotation <- gsub("“","", omitapostrophe2)

omitrightquotation <- gsub("”","",omitleftquotation)

NoPunctuation <- gsub("[[:punct:]]","",omitrightquotation)

omitOE <- gsub("œ","",NoPunctuation)

clean.string <- gsub("[0-9]","",omitOE)


#We then shift the text

x2<- strsplit(clean.string, "")

no.last<- substr(clean.string, 1, nchar(clean.string)-1)
```

```
first.blank<- str_c(" ", no.last)

x1<- strsplit(first.blank,"")


#We define vowels and consonants

vowels<-c("a","e","i","o","u")

consonants<- c("b","c","d","f","g","h","j","k","l","m","n","p","q","r","s","t",

              "v","w","x","y","z")


#We set the patterns

for (counter in 1:nchar(x2)){

  v<- ifelse(x2[[counter]] %in% vowels,1,0)

  c<- ifelse(x2[[counter]] %in% consonants,1,0)

  vv<- ifelse(x1[[counter]] %in% vowels & x2[[counter]] %in% vowels,1,0)

  vc<- ifelse(x1[[counter]] %in% vowels & x2[[counter]] %in% consonants,1,0)

  cv<- ifelse(x1[[counter]] %in% consonants & x2[[counter]] %in% vowels,1,0)

  cc<- ifelse(x1[[counter]] %in% consonants & x2[[counter]] %in%
consonants,1,0)

}


library(markovchain)

#We set up a transition matrix

transitionMobyDick<- matrix(c((sum(vv)/(sum(vv)+sum(vc))),
```

```
(sum(vc)/(sum(vv)+sum(vc))),
(sum(cv)/(sum(cv)+sum(cc))),
(sum(cc)/(sum(cv)+sum(cc))), nrow=2, ncol=2, byrow=TRUE))
```

```
              [v]         [c]
[v] 0.1422454   0.8577546
[c]  0.5133822  0.4866178
```

Works Cited

Dickens, Charles. *A Christmas Carol*. https://www.gutenberg.org/files/46/46-h/46-h.htm. 1843.

Dickens, Charles. *A Tale of Two Cities*. https://www.gutenberg.org/files/98/98-h/98-h.htm. 1859.

Dickens, Charles. *Great Expectations.* https://www.gutenberg.org/files/1400/1400-h/1400-h.htm. 1860.

Korosteleva, Olga. *Stochastic Processes With R: An Introduction*. CRC Press, 2022.

Melville, Herman. *Moby Dick.* https://www.gutenberg.org/files/2701/2701-h/2701-h.htm. 1851.