Stat. 653 Project

Mulati Patiguli

May 15, 2019

For my project ,I run the LDA algorithm on 4 different sets of books, which are downloaded from gutenberg package, whose names are “FROM THE EARTH TO THE MOON” ,“LIFE ON THE MISSISSIPPI” ,“Democracy and Education: An Introduction to the Philosophy of Education” ,“History of the Decline and Fall of the Roman Empire — Volume 1”, .

library(gutenbergr)  
library(knitr)  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.5.2

library(methods)  
library(scales)  
library(topicmodels)  
library(tidytext)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyr)

books\_project <- gutenberg\_download(c(83,245,852,890),   
 meta\_fields = "title")

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

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

books\_project

## # A tibble: 57,025 x 3  
## gutenberg\_id text title   
## <int> <chr> <chr>   
## 1 83 " FROM THE… From the Earth to the Moon;…  
## 2 83 "" From the Earth to the Moon;…  
## 3 83 " … From the Earth to the Moon;…  
## 4 83 "" From the Earth to the Moon;…  
## 5 83 " J… From the Earth to the Moon;…  
## 6 83 "" From the Earth to the Moon;…  
## 7 83 "" From the Earth to the Moon;…  
## 8 83 "" From the Earth to the Moon;…  
## 9 83 "" From the Earth to the Moon;…  
## 10 83 " Tab… From the Earth to the Moon;…  
## # ... with 57,015 more rows

Before pre-processing, I need to divide these into chapters, use tidytext’s unnest\_tokens() to separate them into words, then remove stop\_words and I’m treating every chapter as a separate “document”, each with a name .

After using count()function , I found out that the most common words , for the book “Democracy and Education”, are still the words ,such as education ,science ,philosophy and experience , and these are the words that are pretty much related to this book .

For the book “Life on the Mississippi” ,the words ,like “river”,“water” are the most common words since this book is related to Mississippi river .

library(stringr)  
library(tidyverse)

## ── Attaching packages ────────────────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ tibble 1.4.2 ✔ purrr 0.2.5  
## ✔ readr 1.2.1 ✔ forcats 0.3.0

## ── Conflicts ───────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ readr::col\_factor() masks scales::col\_factor()  
## ✖ purrr::discard() masks scales::discard()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(tidytext)  
# divide into documents, each representing one chapter  
by\_chapter\_project<- books\_project %>%  
 group\_by(title) %>%  
 mutate(chapter = cumsum(str\_detect(text, regex("^chapter ", ignore\_case = TRUE)))) %>%  
 ungroup() %>%  
 filter(chapter > 0) %>%  
 unite(document, title, chapter)  
  
# split into words  
by\_chapter\_word\_project <- by\_chapter\_project %>%  
 unnest\_tokens(word, text)  
  
# find document-word counts  
word\_counts\_project <- by\_chapter\_word\_project %>%  
 anti\_join(stop\_words) %>%  
 count(document, word, sort = TRUE) %>%  
 ungroup()

## Joining, by = "word"

word\_counts\_project

## # A tibble: 139,102 x 3  
## document word n  
## <chr> <chr> <int>  
## 1 Democracy and Education: An Introduction to the Philos… education 67  
## 2 Democracy and Education: An Introduction to the Philos… social 63  
## 3 Life on the Mississippi\_121 river 59  
## 4 Democracy and Education: An Introduction to the Philos… science 58  
## 5 Democracy and Education: An Introduction to the Philos… philosop… 58  
## 6 Democracy and Education: An Introduction to the Philos… experien… 57  
## 7 Life on the Mississippi\_121 head 54  
## 8 Life on the Mississippi\_76 associat… 53  
## 9 Democracy and Education: An Introduction to the Philos… individu… 52  
## 10 Democracy and Education: An Introduction to the Philos… method 51  
## # ... with 139,092 more rows

#### LDA on chapters

After making data frame word\_counts in a tidy form, with one-term-per-document-per-row , we still need to cast a one-token-per-row table into a DocumentTermMatrix with tidytext’s cast\_dtm()

chapters\_dtm\_project <- word\_counts\_project %>%  
 cast\_dtm(document, word, n)  
  
chapters\_dtm\_project

## <<DocumentTermMatrix (documents: 265, terms: 22387)>>  
## Non-/sparse entries: 139102/5793453  
## Sparsity : 98%  
## Maximal term length: 20  
## Weighting : term frequency (tf)

I used the LDA() function to create a six-topic LDA model because I downloaded four different books so I set k=4 and set a seed to 1234 so that the output of the model is predictable .

chapters\_lda\_project <- LDA(chapters\_dtm\_project, k =4, control = list(seed = 1234))  
chapters\_lda\_project

## A LDA\_VEM topic model with 4 topics.

#### Word-topic probabilities

After fitting LDA model , we need to explore and interpreting the model using tidying functions from the tidytext package .This package provides for extracting the per-topic-per-word probabilities, called (“beta”), from the model. Now I need to examine the Word-topic probabilities with matirx=beta .

from the below result ,we can see that the term “education” has almost zero probability generated from topic 4 but compare to topic 2 , i think this term most probabily came from topic 2 with the probability of 8.631327e-03.

chapter\_topics\_project <- tidy(chapters\_lda\_project, matrix = "beta")  
chapter\_topics\_project

## # A tibble: 89,548 x 3  
## topic term beta  
## <int> <chr> <dbl>  
## 1 1 education 1.13e- 4  
## 2 2 education 8.63e- 3  
## 3 3 education 2.77e- 4  
## 4 4 education 1.58e-29  
## 5 1 social 5.60e- 5  
## 6 2 social 9.23e- 3  
## 7 3 social 8.07e- 5  
## 8 4 social 3.01e- 5  
## 9 1 river 9.29e- 3  
## 10 2 river 1.89e- 5  
## # ... with 89,538 more rows

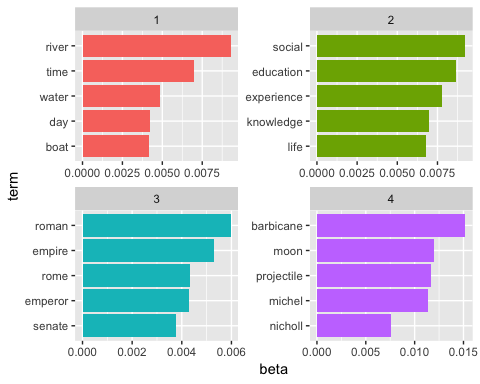
And then I used dplyr’s top\_n() to find the top 5 terms within each topic with a ggplot2 visualizaion .

The most common words in topic1 include “river”,“time”,“water” ,which suggests it may represent the book *LIFE ON THE MISSISSIPPI*. The topic of “social”, “education”, “experience”, and “knowledge” belongs to *“Democracy and Education: An Introduction to the Philosophy of Education”* . And that “roman”, “empire”, and “emperor” belongs to *History of the Decline and Fall of the Roman Empire — Volume 1*. Besides I also think that “moon” and “barbicane” from *FROM THE EARTH TO THE MOON*. And i don’t see any common words within these 4 topics for the top 5 terms .because I choose quite different 4 books for my project .

library(ggplot2)  
top\_terms\_project <- chapter\_topics\_project %>%  
 group\_by(topic) %>%  
 top\_n(5, beta) %>%  
 ungroup() %>%  
 arrange(topic, -beta)  
  
top\_terms\_project

## # A tibble: 20 x 3  
## topic term beta  
## <int> <chr> <dbl>  
## 1 1 river 0.00929  
## 2 1 time 0.00699  
## 3 1 water 0.00485  
## 4 1 day 0.00421  
## 5 1 boat 0.00415  
## 6 2 social 0.00923  
## 7 2 education 0.00863  
## 8 2 experience 0.00778  
## 9 2 knowledge 0.00694  
## 10 2 life 0.00676  
## 11 3 roman 0.00598  
## 12 3 empire 0.00531  
## 13 3 rome 0.00434  
## 14 3 emperor 0.00428  
## 15 3 senate 0.00376  
## 16 4 barbicane 0.0152   
## 17 4 moon 0.0120   
## 18 4 projectile 0.0116   
## 19 4 michel 0.0114   
## 20 4 nicholl 0.00759

top\_terms\_project %>%  
 mutate(term = reorder(term, beta)) %>%  
 ggplot(aes(term, beta, fill = factor(topic))) +  
 geom\_col(show.legend = FALSE) +  
 facet\_wrap(~ topic, scales = "free") +  
 coord\_flip()



#### Per-document classification

we may want to know which topics are associated with each document and we can find this by examining the per-document-per-topic probabilities, (“gamma”).

The model estimates that each word in the *Democracy and Education: An Introduction to the Philosophy of Education* document has only a 0.00000793 probability of coming from topic 1 .

chapters\_gamma\_project<- tidy(chapters\_lda\_project, matrix = "gamma")  
chapters\_gamma\_project

## # A tibble: 1,060 x 3  
## document topic gamma  
## <chr> <int> <dbl>  
## 1 Democracy and Education: An Introduction to the Philos… 1 7.99e-6  
## 2 Life on the Mississippi\_121 1 10.00e-1  
## 3 Democracy and Education: An Introduction to the Philos… 1 1.24e-5  
## 4 Democracy and Education: An Introduction to the Philos… 1 1.39e-5  
## 5 Democracy and Education: An Introduction to the Philos… 1 1.01e-5  
## 6 Life on the Mississippi\_76 1 10.00e-1  
## 7 Democracy and Education: An Introduction to the Philos… 1 9.94e-6  
## 8 Democracy and Education: An Introduction to the Philos… 1 9.58e-6  
## 9 Democracy and Education: An Introduction to the Philos… 1 1.28e-5  
## 10 Democracy and Education: An Introduction to the Philos… 1 1.12e-5  
## # ... with 1,050 more rows

To examine how well our unsupervised learning did at distingushing the four book ,first we re-seperate document name into title and chapter .

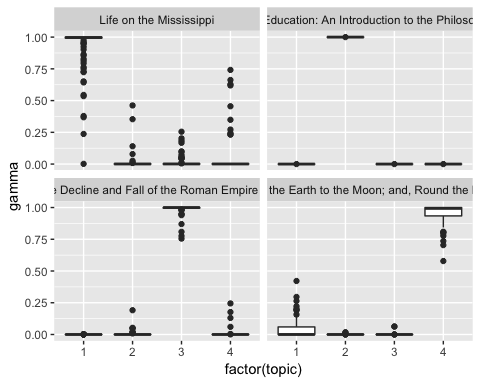
chapters\_gamma\_project <- chapters\_gamma\_project %>%  
 separate(document, c("title", "chapter"), sep = "\_", convert = TRUE)  
  
chapters\_gamma\_project

## # A tibble: 1,060 x 4  
## title chapter topic gamma  
## <chr> <int> <int> <dbl>  
## 1 Democracy and Education: An Introduction to th… 7 1 7.99e-6  
## 2 Life on the Mississippi 121 1 10.00e-1  
## 3 Democracy and Education: An Introduction to th… 17 1 1.24e-5  
## 4 Democracy and Education: An Introduction to th… 24 1 1.39e-5  
## 5 Democracy and Education: An Introduction to th… 20 1 1.01e-5  
## 6 Life on the Mississippi 76 1 10.00e-1  
## 7 Democracy and Education: An Introduction to th… 22 1 9.94e-6  
## 8 Democracy and Education: An Introduction to th… 13 1 9.58e-6  
## 9 Democracy and Education: An Introduction to th… 16 1 1.28e-5  
## 10 Democracy and Education: An Introduction to th… 14 1 1.12e-5  
## # ... with 1,050 more rows

Visualize the per-topic per document probability using boxplot for each . From below boxplot ,I found out that almost all of the chapters from *LIFE ON THE MISSISSIPPI*, *Democracy and Education: An Introduction to the Philosophy of Education*, and *History of the Decline and Fall of the Roman Empire — Volume 1* were uniquely identified as a single topic each.

It does look like some chapters from “from the earth to the moon” (which should be topic 4) were somewhat associated with other topics.

# reorder titles in order of topic 1, topic 2, etc before plotting  
chapters\_gamma\_project %>%  
 mutate(title = reorder(title, gamma \* topic)) %>%  
 ggplot(aes(factor(topic), gamma)) +  
 geom\_boxplot() +  
 facet\_wrap(~ title)



To examine the topic that was most associated with each chapter , I used top\_n(), which is effectively the “classification” of that chapter.

chapter\_classifications\_project <- chapters\_gamma\_project %>%  
 group\_by(title, chapter) %>%  
 top\_n(1, gamma) %>%  
 ungroup()  
  
chapter\_classifications\_project

## # A tibble: 265 x 4  
## title chapter topic gamma  
## <chr> <int> <int> <dbl>  
## 1 Life on the Mississippi 121 1 1.000  
## 2 Life on the Mississippi 76 1 1.000  
## 3 Life on the Mississippi 113 1 1.000  
## 4 Life on the Mississippi 64 1 1.000  
## 5 Life on the Mississippi 61 1 0.651  
## 6 Life on the Mississippi 89 1 1.000  
## 7 Life on the Mississippi 71 1 1.000  
## 8 Life on the Mississippi 69 1 1.000  
## 9 Life on the Mississippi 78 1 1.000  
## 10 Life on the Mississippi 92 1 1.000  
## # ... with 255 more rows

shows the topic for each book .

book\_topics\_project <- chapter\_classifications\_project %>%  
 count(title, topic) %>%  
 group\_by(title) %>%  
 top\_n(1, n) %>%  
 ungroup() %>%  
 transmute(consensus = title, topic)  
book\_topics\_project

## # A tibble: 4 x 2  
## consensus topic  
## <chr> <int>  
## 1 Democracy and Education: An Introduction to the Philosophy of Edu… 2  
## 2 From the Earth to the Moon; and, Round the Moon 4  
## 3 History of the Decline and Fall of the Roman Empire — Volume 1 3  
## 4 Life on the Mississippi 1

commpare each concensus topic for each book .

from below result , four chapters from the book *Life on the Mississippi* was misclassified , as LDA described the first one ,which is chapter 60 ,as coming from the *From the Earth to the Moon; and, Round the Moon* which is topic 4 .

Misclassified the second (chapter33) ,the third (chapter33) and the forth (chapter53) from the book*Life on the Mississippi*  ,as coming from the *From the Earth to the Moon; and, Round the Moon* which is also topic 4.

chapter\_classifications\_project %>%  
 inner\_join(book\_topics\_project, by = "topic") %>%  
 filter(title != consensus)

## # A tibble: 4 x 5  
## title chapter topic gamma consensus   
## <chr> <int> <int> <dbl> <chr>   
## 1 Life on the Miss… 60 4 0.662 From the Earth to the Moon; and, …  
## 2 Life on the Miss… 33 4 0.629 From the Earth to the Moon; and, …  
## 3 Life on the Miss… 38 4 0.619 From the Earth to the Moon; and, …  
## 4 Life on the Miss… 53 4 0.742 From the Earth to the Moon; and, …

#### By word assignments: augment

We use augment ()function to find which words in each document were assignend to which topic . and this will add a new column: .topic

assignments\_project<- augment(chapters\_lda\_project , data = chapters\_dtm\_project)  
assignments\_project

## # A tibble: 139,102 x 4  
## document term count .topic  
## <chr> <chr> <dbl> <dbl>  
## 1 Democracy and Education: An Introduction to the P… educat… 67 2  
## 2 Democracy and Education: An Introduction to the P… educat… 7 2  
## 3 Democracy and Education: An Introduction to the P… educat… 25 2  
## 4 Democracy and Education: An Introduction to the P… educat… 13 2  
## 5 Life on the Mississippi\_76 educat… 1 1  
## 6 Democracy and Education: An Introduction to the P… educat… 8 2  
## 7 Democracy and Education: An Introduction to the P… educat… 4 2  
## 8 Democracy and Education: An Introduction to the P… educat… 4 2  
## 9 Democracy and Education: An Introduction to the P… educat… 12 2  
## 10 Democracy and Education: An Introduction to the P… educat… 49 2  
## # ... with 139,092 more rows

combine assignment table with the concensus book title to find which words were incorrectly identified . true book (title column ) and the book assigned is (consensus column) .

assignments\_project <- assignments\_project %>%  
 separate(document, c("title", "chapter"), sep = "\_", convert = TRUE) %>%  
 inner\_join(book\_topics\_project, by = c(".topic" = "topic"))  
  
assignments\_project

## # A tibble: 139,102 x 6  
## title chapter term count .topic consensus   
## <chr> <int> <chr> <dbl> <dbl> <chr>   
## 1 Democracy and Educa… 7 educa… 67 2 Democracy and Educati…  
## 2 Democracy and Educa… 17 educa… 7 2 Democracy and Educati…  
## 3 Democracy and Educa… 24 educa… 25 2 Democracy and Educati…  
## 4 Democracy and Educa… 20 educa… 13 2 Democracy and Educati…  
## 5 Life on the Mississ… 76 educa… 1 1 Life on the Mississip…  
## 6 Democracy and Educa… 22 educa… 8 2 Democracy and Educati…  
## 7 Democracy and Educa… 13 educa… 4 2 Democracy and Educati…  
## 8 Democracy and Educa… 16 educa… 4 2 Democracy and Educati…  
## 9 Democracy and Educa… 14 educa… 12 2 Democracy and Educati…  
## 10 Democracy and Educa… 23 educa… 49 2 Democracy and Educati…  
## # ... with 139,092 more rows

Almost all of the words were correctly identified .

assignments\_project %>%  
 count(title, consensus, wt = count) %>%  
 group\_by(title) %>%  
 mutate(percent = n / sum(n)) %>%  
 ggplot(aes(consensus, title, fill = percent)) +  
 geom\_tile() +  
 scale\_fill\_gradient2(high = "red", label = percent\_format()) +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1),  
 panel.grid = element\_blank()) +  
 labs(x = "Book words were assigned to",  
 y = "Book words came from",  
 fill = "% of assignments")



For some of these words, such as “subject” and “mind”, were misidentified ,that’s because they’re more common in *Democracy and Education: An Introduction to the Philosophy of Education*

And somewords like “wood”,“trees” ,“cock” are also misidentified , they originally should appear in the book *From the Earth to the Moon; and, Round the Moon* , but they are still more common in *Life on the Mississippi* .

wrong\_words\_project <- assignments\_project %>%  
 filter(title != consensus)  
  
wrong\_words\_project

## # A tibble: 2,897 x 6  
## title chapter term count .topic consensus   
## <chr> <int> <chr> <dbl> <dbl> <chr>   
## 1 Life on the Missis… 74 educa… 1 2 Democracy and Educatio…  
## 2 History of the Dec… 61 educa… 2 2 Democracy and Educatio…  
## 3 History of the Dec… 67 educa… 1 2 Democracy and Educatio…  
## 4 History of the Dec… 15 social 1 2 Democracy and Educatio…  
## 5 History of the Dec… 39 social 1 2 Democracy and Educatio…  
## 6 History of the Dec… 63 social 1 2 Democracy and Educatio…  
## 7 From the Earth to … 13 river 2 1 Life on the Mississippi  
## 8 History of the Dec… 38 scien… 1 4 From the Earth to the …  
## 9 History of the Dec… 15 philo… 1 2 Democracy and Educatio…  
## 10 From the Earth to … 36 philo… 1 2 Democracy and Educatio…  
## # ... with 2,887 more rows

wrong\_words\_project %>%  
 count(title, consensus, term, wt = count) %>%  
 ungroup() %>%  
 arrange(desc(n))

## # A tibble: 2,623 x 4  
## title consensus term n  
## <chr> <chr> <chr> <dbl>  
## 1 History of the Decline an… Democracy and Education: An Int… subj… 9  
## 2 From the Earth to the Moo… Life on the Mississippi wood 8  
## 3 From the Earth to the Moo… Life on the Mississippi alge… 7  
## 4 From the Earth to the Moo… Life on the Mississippi cock 7  
## 5 From the Earth to the Moo… Life on the Mississippi head 7  
## 6 From the Earth to the Moo… Life on the Mississippi pris… 7  
## 7 From the Earth to the Moo… Life on the Mississippi trees 7  
## 8 History of the Decline an… Democracy and Education: An Int… mind 7  
## 9 Life on the Mississippi From the Earth to the Moon; and… sea 7  
## 10 Life on the Mississippi History of the Decline and Fall… roman 7  
## # ... with 2,613 more rows