Andy Zhu

```
from google.colab import drive
from google.colab import files
drive.mount("/content/drive/", force_remount=True)
```

```
    Mounted at /content/drive/
```

```
import pandas as pd
df = pd.read_csv("/content/drive/My Drive/Sac State/CSC177/data/output.csv")
df
```

	Organization	Content	Revenue
0	THE WICKHAM FOUNDATION INC	DIFFERENCE BETWEEN FMV AND COST OF	24780.0
1	A&M FELDMAN FOUNDATION INC	PRIOR PERIOD ADJUSTMENT	109601.0
2	HAROLD AND RENEE BERGER FOUNDATION	MISC. REVENUE	112073.0
3	The Joelson Foundation	PARTNERSHIP INCOME/LOSS	652757.0
4	THE RONALD AND JANE OLSON FOUNDATION	OTHER INCOME - BLACKSTONE GROUP L.P. K-1	300079.0
...	...	...	...
475570	THE AUGUST AICHHORN CENTER FOR	OUR MISSION IS TO PROVIDE A THERAPEUTIC, CONSI...	NaN
475571	NESTUCCA NESKOWIN AND SAND LAKE	TO PROVIDE A FORUM FOR PUBLIC PARTICIPATION AN...	NaN
		COMMONWORKS PROVIDES	

Script

Write a Julia script to process the data, including the following steps:

- Extract from the 2019 IRS 990 data the organization name, one or more text elements that describe the purpose of the organization (description), and one or more elements that can be used as a proxy for size, such as revenue or number of employees.
- Process the text descriptions that describe each organization using a similar process as described in class.
- Create and save the term document matrix for the processed text descriptions.

```
import Pkg; Pkg.add("EzXML");Pkg.add("DataFrames"); Pkg.add("TextAnalysis"); Pkg.add("Query")
using EzXML, DataFrames, TextAnalysis, Query, Serialization, CSV

"""
    Extract org name xpath from the xml doc
"""
function getname(xmlfile)
  doc = parsexml(xmlfile)
  d = findfirst("//BusinessName/BusinessNameLine1Txt/text()", doc)

  if isnothing(d)
    return("NA")
  else
    nodecontent(d)
  end
end

"""
    Extract org content xpath from the xml doc
"""
function getcontent(xmlfile)
    doc = parsexml(xmlfile)
    d = findfirst("//MissionDesc/text()", doc)

    if isnothing(d)
      d = findfirst("//Desc/text()", doc)
    end
    if isnothing(d)
        d = findfirst("//Description/text()", doc)
    end

    if isnothing(d)
      return("NA")
    else
      nodecontent(d)
    end
end

"""
    Extract org revenue/expenses xpath from the xml doc
"""
function getrevenue(xmlfile)
    doc = parsexml(xmlfile)
    d = findfirst("//AnalysisOfRevenueAndExpenses/TotalRevAndExpnssAmt/text()", doc)

    if isnothing(d)
      d = findfirst("//AnalysisOfRevenueAndExpenses/TotalExpensesRevAndExpnssAmt/text()", doc
    end

    if isnothing(d)
      return("NA")
    else
      nodecontent(d)
    end
end

function main()
  files2019 = readdir("2019/", join=true)
  df = DataFrame(Organization = String[], Content = String[], Revenue = String[])

  @time for f in files2019
    doc = read(f, String)

    name = getname(doc)
    content = getcontent(doc)
    revenue = getrevenue(doc)
    push!(df, (name, content, revenue))
  end

  CSV.write("output.csv", df)

  """
    Query all non-empty organization content.
      Empty @collect returns Query as an array.
  """
  @time q1 = @from i in df begin
    @where i.Content != "NA"
    @select i.Content
    @collect
  end

  c = Corpus([])
  for x in q1
   doc = StringDocument(x)
    push!(c,doc)
  end

  open("preprocess.txt", "w") do file
    write(file, text(c[8]))
  end



  @time remove_case!(c)
  @time prepare!(c, strip_punctuation)
  @time stem!(c)
  @time update_lexicon!(c)

  open("postprocess.txt", "w") do file
    write(file, text(c[8]))
  end

  @time d = DocumentTermMatrix(c)
  serialize("dtm.jls", dtm(d, :sparse))
end

if abspath(PROGRAM_FILE) == @__FILE__
  main()
end
```

Questions

1. How many of the returns were you able to process?

> 475,575

2. Show and interpret one explicit example of what you extracted from one tax return, including the text description before and after processing.

```
df.iloc[25]
```

```
Organization                                COMMUNITY OPTIONS INC
Content         COMMUNITY OPTIONS BELIEVES IN THE DIGNITY OF E...
Revenue                                                       NaN
Name: 25, dtype: object
```

Pre-processed:

> COMMUNITY OPTIONS BELIEVES IN THE DIGNITY OF EVERY PERSON, AND IN THE FREEDOM OF ALL PEOPLE TO EXPERIENCE THE HIGHEST DEGREE OF SELF-DETERMINATION. EMBRACING THIS PHILOSOPHY, COMMUNITY OPTIONS PROVIDES HOUSING, SUPPORT SERVICES AND ADVOCACY ASSISTANCE TO HELP EMPOWER PEOPLE WITH DISABILITIES. COMMUNITY OPTIONS, INC. DEVELOPS RESIDENTIAL AND EMPLOYMENT SUPPORTS FOR PEOPLE WITH SEVERE DISABILITIES,UTILIZING TECHNOLOGY AND TRAINING. AS A NATIONAL AGENCY,COMMUNITY OPTIONS HAS PARTICIPATED IN INSTITUTIONAL CLOSURE AND COMMUNITY RESIDENTIAL PLACEMENT FOR THOUSANDS OF PEOPLE ACROSS SEVERAL STATES. COMMUNITY OPTIONS DOES NOT ADMINISTER ANY LARGE CONGREGATE PROGRAMS, RECOGNIZING THAT PEOPLE WITH THE MOST SEVERE DISABILITIES NEED ENVIRONMENTS, EQUIPMENT, CLINICAL AND STAFF SUPPORT THAT ARE TAILORED TO THEIR VERY SPECIFIC NEEDS. IN ITS HISTORY, THE AGENCY HAS DEVELOPED A REPUTATION FOR QUALITY, COST EFFECTIVE ADMINISTRATION THAT ENCOURAGES INDIVIDUAL CHOICE AND FLEXIBILITY.

Post-processed:

> communiti option believ in the digniti of everi person and in the freedom of all peopl to experi the highest degre of selfdetermin embrac this philosophi communiti option provid hous support servic and advocaci assist to help empow peopl with disabl communiti option inc develop residenti and employ support for peopl with sever disabilitiesutil technolog and train as a nation agencycommun option has particip in institut closur and communiti residenti placement for thousand of peopl across sever state communiti option doe not administ ani larg congreg program recogn that peopl with the most sever disabl need environ equip clinic and staff support that are tailor to their veri specif need in it histori the agenc has develop a reput for qualiti cost effect administr that encourag individu choic and flexibl

3. What are the dimensions of your term document matrix?

   > A 323172 X 82625 DocumentTermMatrix

4. How long did your program take to run? (Less than 30 minutes is easily attainable, but no problem if it takes longer, either.)

   > 31 minutes

5. Which parts of the program took the longest time to run?

   > Processing the xml files to get the required data fields into a dataframe took the longest. 27mins of the entire 31min duration of the program.

✓ 0s completed at 9:34 PM ● ×