Graduating Classes

by Nam Nguyen

Abstract The **gradinfo** package uses data published online by the Williams College Registrar to compute numerous statistics about the graduating classes of Williams from 2001 to 2016. The data were acquired by scanning the text versions of the published documents and reorganized to serve the purpose of computing relevant statistics. The package also provides tools to visualize the variation of these statistics over time, which is helpful in revealing some important characteristics of the student body at Williams College.

Introduction

The list of graduating seniors annually released by Williams College Registrar is a rich source of information for those who are interested in learning about the academic environment and the student body of Williams. However, the sheer volume of information contained in the list can sometimes stand as a discouragement to those who want to study the list closely. The motivation of this package is to help the user get access to information about Williams College graduating classes from 2001 to 2016 and to provide some tools that aid the process of analyzing the data in this list.

Included in the database of **gradinfo** package are data frames that store information about names, Latin honors, department honors, Sigma Xi membership and Phi Kappa Beta membership of Williams students graduating between 2001 and 2016. The creator of this package believes that an anlytical study of these information will help identify important characteristics of Williams College student body, which are a valuable knowledge to those who want to learn more about Wiliams such as high school students in college admission process or scholars looking to teach at Williams.

Data

The information used to construct the data in this package was extracted from annual course catalogs published on the website of The Office of The Registrar of Williams College. The documents for each year between 2001 to 2016 were downloaded and manually converted to text format using an online pdf to text converter. These text files were then manually edited to filter out all irrelevant content so that only information about the graduating class is left.

However, reading all information from these texts in a systematic way is not an easy task, given the inconsistent and convoluted format in which the data is published. To ensure that the information about past graduating classes at Williams College is read in accurately, the package provides function **readStudentInfo**, which extracts data from the locally stored text files in the following manner:

First, the function reads in the list of students graduating in a particular year line by line:

```
strwrap(head(readLines(system.file("extdata", "2000-01.txt", package = "gradinfo"), warn = FALSE)))
```

```
#> [1] "Conferring of the Degree of Master of Arts"
#> [2] "Katherine Anne Bussard"
#> [3] "Lisa Beth Dorin"
#> [4] "Alanna Erin Gedgaudas"
#> [5] "*Robert Gordon Glass"
#> [6] "Elyse Amparo Gonzales"
```

Each of these characters vector was then broken down into 6 shorter vectors corresponding to 6 categories within the list of students: students who received master degree of Art, students who received master degree of Economics Development, students who were awarded Summa Cum Laude, students who were awarded Magna Cum Laude, students who were awarded Cum Laude and finally, students who only graduated with a Bachelor Degree of Arts.

```
summa <- input[(brks[2] + 2):(brks[3] - 1),]
head(summa)
```

```
#> [1] "*DoHyun Tony Chung, with honors in Political"
#> [2] "Economy"
#> [3] "+*Rebecca Tamar Cover, with highest honors in"
#> [4] "Astrophysics"
#> [5] "*Amanda Bouvier Edmonds"
#> [6] "*Douglas Bertrand Marshall III, with highest"
```

Next, each of these 6 vectors was passed into a helper function named **process_data** which creates a dataframe with 6 columns from its argument vector. In order split each row of the argument vector into multiple chunks of data, function **gsub** was used. In particular, first, the name of the student was extracted by replacing all characters behind the first comma (if there is any) by an empty string. After that, the function checks if the row contains the word "with". If there is, then the characters following the word "with" will be extracted, giving us the name of the major for which the student received an honor. Sometimes, the name of the major and the name of the student are not written on the same row. For such situation, the function also checks the next row everytime it finds the word "honor" in the current row. If the next row has the word "in" or contains just a single word, then the next row would be merged with the current row before the subject is read in. An illustration of this process is provided below:

```
i <- 1
temp <- dataset[i,] # getting raw content from a specific row of the read-in dataset
temp # typical format of a row in the text file

#> [1] "*DoHyun Tony Chung, with honors in Political"
```

Getting the name of the student by removing all symbols that indicate Phi Kappa Beta and Sigma Xi membership then replacing all characters behind the first comma by empty string:

```
name <- gsub("[+*]", "", temp) # removing symbols that indicate student's Phi Kappa Beta
                                 and Sigma Xi membership
name <- gsub(",.*$", "", name) # extracting characters in front of the first comma

name

#> [1] "DoHyun Tony Chung"
```

Getting the major for which the student received an honor:

```
if (grepl(" in ", temp) && (grepl(" ", dataset[i + 1,]) || i == nrow(dataset))) {
  temp <- gsub("[\r\n]", "", temp)      # remove end of line character (if there is any)
  subject <- gsub("^.*\\ in ","", temp)  # read in the subject
} else {
 subject <- paste(temp, dataset[i + 1,], sep = " ")  # merge the next line and the current line
  subject <- gsub("^.*\\ in ","", subject)           # read in the subject
}

subject

#> [1] "Political Economy"
```

A big drawback of this scraping method is its consistency. The **process_data** function was designed based on certain assumptions about the format of the text files. For instance, the function heavily relies on the assumption that the word "with" always follows right after the names of the students who received a department honor. Sometimes, this may not be the case. It could be possible that the name of the student is too long and hence, the word "with" is placed in the following line. If that happens, then it is impossible to distinguish whether it is the student whose name is listed in the current line receiving the honor or the student whose name is listed in the next line.

Another dataset used in this package is the numbers of Williams College students by majors over the course of 10 years from 2007 to 2016. The data were provided by Mary L. Morrison in Williams College Registrar Office in pdf format. Since the data were organized in tabular form, they were directly read in using the function **read.delim** after being converted to text. Then the dataset was modified so that the years, which originally were column names, became values of a variable named `Year`. The dataset **sum_majors** shows total number of majors by graduation year while the dataset **majors** shows the number of students in each department for every year from 2007 to 2016. Below is a condensed overview of two datasets:

```
sum_majors

#>    Year Number.of.Majors
#> 1  2007              691
#> 2  2008              707
#> 3  2009              711
#> 4  2010              724
#> 5  2011              714
#> 6  2012              734
#> 7  2013              735
#> 8  2014              746
#> 9  2015              732
#> 10 2016              771

head(majors)

#>   Majors Year Number.of.Students  Percentage
#> 1   AMST 2007                 11 0.015918958
#> 2   ANTH 2007                  4 0.005788712
#> 3   ARAB 2007                  0 0.000000000
#> 4    ART 2007                 56 0.081041968
#> 5   ASPH 2007                  1 0.001447178
#> 6   ASST 2007                  8 0.011577424
```

Use readStudentInfo

The information within the locally stored text files can be accessed by using the **readStudentInfo** function which reads in and processes the data as described above. This function outputs a single data frame that contains 9 columns: Name, Dept.honor, Dept.honor.lv, sigma.xi, PKB, Clark.Fellow, Latin.honor, Grad.Year. More information about these columns can be found by typing the command **?williams_grad**.

However, please be noted that the dataset **williams_grad** provided by this package has slightly different structure from the dataset returned by **readStudentInfo** function. In particular, in **williams_grad**, the **Name** column is divided into the **First.Name** and **Middle.and.Last.Name** columns. This serves the purposes of identifying the students' gender by first name, which is included in the last column of **williams_grad**.

The purpose of providing the dataset **williams_grad** along with the function **readStudentInfo** is to make sure that the user can get access to data for a particular year as well was data for all 16 years (from 2001 to 2016), thus saving the amount of time that would be spent on truncating or merging the data otherwise.

readStudentInfo requires two arguments:

- dataset: Provide the name of the text file from which the data will be read. The format of the entry is of the form "20XX-YY.txt", where XX and YY are two pairs of last digits of two consecutive years. For instance, if the user wants to read in the information of the class graduating in 2001, the argument would be "2000-01.txt". Selection can vary from 2000-01 to 2015-16.
- grad.year: grad.year is the calendar year in which the interested class graduated. The format should be a 4 digit number between 2001 and 2016.

Use data_scraping

This function reads in all locally stored text files at the same time using **readStudentInfo** function then combines all returned data frames into a single data frame. After that, the **Name** column is broken down into two columns **Firt.Name** and **Middle.and.Last.Name**. Then, the gender of each student is determined from his or her first name using the **gender** function from **gender** package. Finally, the **Gender** column is added to the previously produced dataset. This function was provided so that the user can update the dataset **williams_grad** of this package.

Figure 1: Number of undergraduate students graduating from Williams from 2001 to 2016

Figure 2: Distribution of the graduating class size at Williams

Use statsummary

This function is used to generate various summary statistics of the data in **williams_grad** dataset. This function has two arguments:

- type: This argument receives one of these five values (`undergrad,grad,latin,department,gender`).
  - undergrad: generate summary statistics about undegraduate students
  - grad: generate summary statistics about graduate students
  - latin: generate summary statistics about the distribution of Latin honors at Williams
  - department: generate summary statistics about the distribution of department honors at Williams
  - gender: generate summary statistics about the overall gender ratio and the gender ratios of each department.
- format: This argument has three values (`numeric,timeseries,distribution`).
  - numeric: return a data frame containing relevant summary statistics
  - timeseries: graphical summary that displays changes over time
  - distribution: graphical summary that displays statistics related to the distribution of the interested population.

Summary

According to data in **williams_grad**, the average number of undergraduate students graduating from Williams College each year is approximately 523 students. Below are some key statistics about the size of the graduating classes from 2001 to 2016:

```
#>  min     Q1 median    Q3 max   mean       sd  n missing
#>  506 512.75  521.5 526.5 560 522.75 12.88669 16       0
```

Looking at Figure 1, we can see that 2001 is the year with largest graduating class (560 students) while 2006 is the year with smallest graduating class(506 students). However, we cannot see drastic change in the number of students graduating from Williams each year. Overall, the number of students graduating from Williams remains quite stable over from 2001 to 2016. Another look at the histogram of the graduating class size at Williams (Figure 2) helps us double check this assertion. Most of the area under the density plot concentrates on the range between 520 to 540, indicating the low variance of the graduating class size at Williams.

We can also look at the variation over time of the number of students in Williams College's graduate school of Art and graduate school of Economics Development for a comparison. From Figure 4, it seems that the distribution of the number of students who received a master degee of Art from Williams is quite skewed. This can be explained by looking at Figure 3. Although for most of the years the number of students receiving master degree of Art from Williams lies between 10 and 14, there is a particular year when this number is 17. That is 2001. Meanwhile, the distribution of students who graduated from Williams' Graduate School of Economics Development has two peaks, one approximately at 23 and the other approximately at 29. There is no particular explanation for this observation, however.

Figure 3: Number of graduate students graduating from Williams from 2001 to 2016

Figure 4: Distribution of number of students who received MA and MAED degrees at Williams

Figure 5: The distribution of Latin honors at Williams Collge over time

Another interesting information that can be attained from the dataset is the distribution of Latin honors among each graduating class. In Figure 5, we can see that the percentages of students who received Summa Cum Laude, Magna Cum Laude and Cum Laude stayed relatively stable over time. However, there is a noticeable observation: the *percentage* of students receiving Magna Cum Laude in 2013 is slightly higher than that of the other years.

The histograms in Figure 6 show us the distributions of the *numbers* of students receiving Summa Cum Laude, Magna Cum Laude, Cum Laude, and no Latin honor over time. From these histograms, we can tell that the average number of students who received Summa Cum Laude, Magna Cum Laude, Cum Laude and no Latin honor are about 10, 70, 105 and 335 respectively. The histograms for Magna Cum Laude, Cum Laude and no Latin honor all contain an outlier. The outliers in the histograms for the number of magna cum laude and cum laude receivers are, not surprisingly, coressponding to the year 2013 while the outlier in the histogram for the number of students who did not receive any Latin honor corresponds to the year 2001. So, in summation, in respect to the common trends exhibited by the histograms, we can say that there was a surge in the number of students receiving Magna Cum Laude in 2013 and a noticeable drop in the total number of students receiving Latin honors in 2001.

Figure 6: Histograms of the numbers of students receiving and not receiving Latin honors

Figure 7: Number of majors at Williams over time

Figure 8: Top 5 most popular majors at Williams from 2007 to 2016

Figure 7 shows us the number of majors at Williams from 2007 and 2016. From the diagram, we can see that the number of majors in Williams follows a generally increasing trend. However, there is two noticeable excpetions: the drops in the number of majors in the year 2011 and 2015.

So how about the most popular majors over this period of time? This piece of information can be found in Figure 8. Over the course of ten years from 2007 to 2016, we can see that Economics has consistently been the most popular major, followed by English, Psychology and Biology. The last spot in top 5 has alternatively been taken by Art, History and Math. Math is consistently one of the top 5 most popular majors at Williams College from 2009 to 2014 but loses its popularity to other majors in 2015 and 2016.

Figure 9: Gender distribution and gender ratio of graduating classes at Williams College over time

Figure 10: Department with the most skewed gender ratio at Williams from 2001 to 2016

Next, we can also look at the gender distribution of the graduating class over time and the gender distribution of students receiving department honors by their respective department. Figure 9 displays two graphs. The first graph tells us the percentage of male and female in the graduating class of each year from 2001 to 2016. Meanwhile, the second graph displays the change in gender ratio of the graduating class over time. Looking at these two graphs, one can tell that the male to female ratio at Williams is relatively high, especially for the year 2016 when the male to female at ratio in Williams is almost 2. This is unlikely to be accurate. According to Williams College Office of Communications, out of all students getting accepted to the class of 2016, 609 are women and 573 are men. Even though not all of these students ended up matriculating at Williams and not all of those who matriculated graduated from the college, it is quite implausible that the male to female ratio drastically changed in favor of male.

The reason for such discrepancy can come from the method in which gender ratio was calculated. First, the gender of each student was determined by his or her first name using function **gender** in **gender"** package. The problem is that function **gender** is not always able to determine a gender for a given name. Sometimes it returns `NA` or `"either"`. Because the **gradinfo** package only aims to provide summary on the gender ratio between male and female, the genders of those students whose genders were originally determined by function **gender** as `NA` or `"either"` were randomly chosen between `"male"` and `"female"`. This method does not guarantee that the numbers of female and male students in the dataset match the numbers of female and male students in reality.

In Figure 10, we can see the department with most skewed gender ratio for each year from 2001 to 2016. However, it is necessary to note that the gender ratio by department here is calculated based on the number of male and female students receiving department honor each year. Since we only know the information about departments of students who received a department honor, this is the only way to come up with rough estimates of gender ratio by department.

Conclusion

The package **gradinfo** includes functions that allow easy reading and analysis of data published by Williams College. This allows the study of various attributes of Williams College graduating classes to be done in a more user friendly manner. However, there are various aspects that this package can improve on. The most necessary improvement would be an update in the numbers of males and females for each year. As commented above, the estimates currently used by this package are unlikely to match the real numbers due to the inconsistency in the gender-predicting algorithm. Another important improvement would be the ability to provide statistical summary on the number of students by majors. This would allow the study of how academic preferences of Williams College students change over time, which majors are rising in popularity and which majors are dropping in popularity,etc... Additionally, the package can be further developed to read in or calculate the number of professors in each department. This would open the scope of modelling how student-faculty ratio affects the performance of students in each department, which can be measured by the percentage of students in the department who received Latin honors.

Nam Nguyen
Williams College
P.O 1871, Paresky Center, Williams College, Williamstown, Massachusetts
ntn3@williams.edu