A Brief Analysis of the first names of Williams College Gradutes 2004-2016

Introdution

The **wgd** package uses publically available data from the Williams College Registrar to load and determine statistics about the first names of Williams College graduates from 2004 to 2016.

Data

The raw data used for this analysis is in the form of PDFs posted to the website of the Office of the Registrar of Williams College. The data was converted into text files using PDFelement, an online PDF to text editor. After the data is read into R, we subset for the first name by subsetting for characters before the first space occurs. Next, non alphanumeric characters are removed to clean up any unwanted characters. In many cases, a "+" or other characters appear before the name of a student to indicate a designation or other acheivement.

```
data <- gsub( " .*$", "", data) ## subset for characters before the first space
data <- gsub("[^[:alnum:] ]", "", data)
```

In many cases, the lines of the original data do not get preserved when copying into the text files. For example, a line might say "Physics" referring to a student who graduated with honors in Physics. We must thus remove cases where the data does not reflect a name, but a subject or designation. We subset for elements that contain a subject name and then remove those elements.

```
rm <- grep("Science|Bachelor|Major|Spanish|Degrees", data) ## many more included in actual function
data <- data[-rm]
```

Next, we create a vector containing the number of characters of each name.

```
len <- numeric(length(data1))
for(i in 1:length(data1)) {
  len[i] <- nchar(data1[i])
}
```

This vector is subsequently binded to the vector of names to create a dataframe.

```
head(readNameData("names2006.txt"))
```

```
##   First Name Length
## 1      Saroj      5
## 2   Jennifer      8
## 3       Ryan      4
## 4      Maria      5
## 5    Phakawa      7
## 6       Ivan      4
```

Use readNameData

The information in the locally stored text files can be loaded and parsed using the readEphData function. The function has one argument and can be used as follows.

```
readNameData("dataset")
```

Use summaryStatistic

The function summaryStatistic() uses the data from the readNameData() function to create summary statistics about the first names of Williams College graduates from 2004 to 2016. The function can be used as follows.

```
summaryStatistic("type")
```

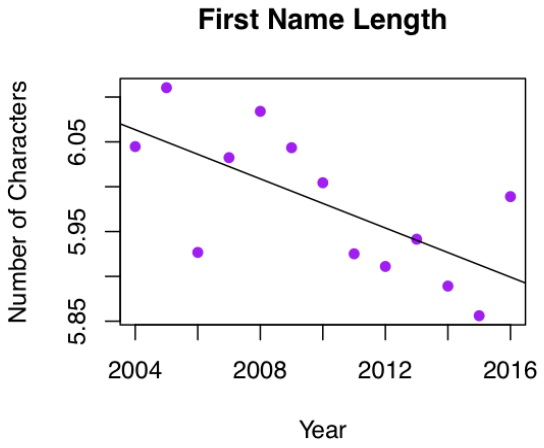
The options for type are "popular", "most popular", "name length", "name length plot", and "least popular." If the user codes popular, the function returns a dataframe with the six most popular names over the past 13 years. If the user codes most popular, the function returns a dataframe reporting all names that have been in the top six over the past 13 years, and how many times each name has been in the top six. If the user codes name length, the function returns a dataframe reporting the average first name length of the graduating class for each year. If the user codes name length plot, the function reports name length data in the form of a scatterplot. If the user codes least popular, the function returns a random sample of 100 first names that have occurred only one time at Williamssince 2004.

```
summaryStatistic("most popular")
```

```
##           Name # times in top 6
## 22   Elizabeth               8
## 19      Daniel               7
## 20     Matthew               7
## 21     Michael               7
## 16 Christopher               6
## 17       Emily               6
## 18       Sarah               6
## 14        John               4
## 15     William               4
## 10      Andrew               3
## 11       David               3
## 12       James               3
## 13      Samuel               3
## 8       Robert               2
## 9       Thomas               2
## 1         Eric               1
## 2       Hannah               1
## 3     Jennifer               1
## 4      Jessica               1
## 5       Joseph               1
## 6        Laura               1
## 7      Rebecca               1
```

We see there are 22 unique names that have been in the top six most popular for Williams College graduating classes between the years of 2004 and 2016. Elizabeth has been in the top six 8 times, and Daniel, Matthew, and Michael, have been in the top six 7 times.

```
summaryStatistic("name length plot")
```



We see that the average first name length has been decreasing over the past 13 years. It would be interesting to compare this trend to the overal trend in the United States.

```
head(summaryStatistic("least popular"))
```

```
## full
##    Donna Jabulani  Caitlyn    Yosef    Ansel Cornelia
##        1        1        1        1        1        1
```

This call returns names that have only occured one time in Williams College graduating classes from 2004 to 2016. These names could be odd spellings of more commmon names, or just uncommon names. For example, Meaghan is an uncommon spelling for a more common name, while Hiteshwar is an uncommon name.

Conclusion

The package wgd allows for easy reading and analysis for first name data of Williams College Graduates between 2004 and 2016. The analysis is limited becuase the package only parses for the first name of the students. This leaves out important information like middle and last name, major, and type of degree earned. Parsing these types of data in generality would allow a much deeper analysis of the data and many more questions to be answered. Regardless, the readNameData and summaryStatistic functions provide a good start for analyzing the first name data of Williams College Graduates between 2004 and 2016.