Week 6 Problem Set

Ben Arancibia

October 11, 2014

Problem Set 1

1 When you roll a fair die there are 6^3 which is 216 possible outcomes.

2 The probability of getting a sum total of 3 when you roll a die two times is first $(1, 2)$ then $(2, 1)$. So there are 2/36 changes of getting it, which reduces to 1/18.

3 If you are in a room of 25 people the probability that two of them have the same birthday is:

$$\frac{365!}{365^{25}(340)!}$$

This is based on the assumption that all birthdays are equally as likely and equal to 1/365. If you add 25 people so that are 50 people in the room

$$\frac{365!}{365^{50}(315)!}$$

Results:

```
birthday <- function(n){
   return (1 - (prod(365:((365-n)+1))/(365^n)))
}

birthday(25)
```

```
## [1] 0.5687
```

```
birthday(50)
```

```
## [1] 0.9704
```

Problem Set 2

Write a program to take a document in English and print out the estimated probabilities for each of the words that occur in that document. Your program should take in a file containing a large document and write out the probabilities of each of the words that appear in that document.

```
setwd("/users/bcarancibia/CUNY_IS_605/assign6") #YOU WILL NEED TO CHANGE THIS

file <- "assign6.sample.txt"
probability_words <- function(file){
  file_read <- readChar(file, file.info(file)$size, useBytes=F)
  Encoding(file_read) <- "UTF-8" #allows me to account for aposotrophe
  words <- (strsplit(file_read, "\\W"))[[1]]
  words <- subset(words, words != "") #issue with regex, i think fixed.....
  words <- tolower(words) #lower case
  probability <- table(words) / length(unique(words))
  return (probability) #answer
}
final <- probability_words(file) #all probability

head(final)
```

```
## words
##       10      100       18     1942       20     2009
## 0.003448 0.001724 0.001724 0.001724 0.001724 0.001724
```

```
#final #IF YOU WANT TO SEE ALL PROBABILITY UNCOMMENT OUT THIS LINE
```

Extend your program to calculate the probability of two words occurring adjacent to each other. It should take in a document, and two words (say the and for) and compute the probability of each of the words occurring in the document and the joint probability of both of them occurring together. The order of the two words is not important.

```
#Use a lot of the same code
setwd("/users/bcarancibia/CUNY_IS_605/assign6") #YOU WILL NEED TO CHANGE THIS

file <- "assign6.sample.txt"
probability_words_pair <- function(file){
  file_read <- readChar(file, file.info(file)$size, useBytes=F)
  Encoding(file_read) <- "UTF-8" #allows me to account for aposotrophe
  words <- (strsplit(file_read, "\\W"))[[1]]
  words <- subset(words, words != "") #issue with regex, i think fixed.....
  words <- tolower(words) #lower case
  vector_pairs <- c() #create vector and then iterate through document....
  for (i in 1:(length(words)-1)){
    vector_pairs[i] <- paste(words[i], words[i+1], sep=" ")
  }
  probability_pair<- table(vector_pairs) / length(unique(vector_pairs))
  return(probability_pair)
}

final2 <- probability_words_pair(file)

head(final2)
```

```
## vector_pairs
##   10 percent     10 years 100 officers     18 years     1942 and
##    0.0008039    0.0008039    0.0008039    0.0008039    0.0008039
##     20 years
##    0.0008039
```

```
#final2 #IF YOU WANT TO SEE ALL PROBABILITY UNCOMMENT OUT THIS LINE
```