Corpus-Linguistics-2023-12-08-regexes

Adam Tallman

2023-12-06

## Regexes and Corpus Linguistics

#whitman <- scan(file.choose(), what=character(), sep="\n")  
whitman <- scan("/Users/Adan Tallman/Desktop/Corpus\_Linguistics\_FSU\_2023-2024/10\_Desagulier/chap 04/whitman.txt", what="char", sep="\n") # Windows  
head(whitman)

## [1] "When I heard the learn'd astronomer;"   
## [2] "When the proofs, the figures, were ranged in columns before me;"   
## [3] "When I was shown the charts and the diagrams, to add, divide, and measure them;"   
## [4] "When I, sitting, heard the astronomer, where he lectured with much applause in the lecture-room,"  
## [5] "How soon , unaccountable, I became tired and sick;"   
## [6] "Till rising and gliding out, I wander'd off by myself,"

#dracula <- scan(file.choose(), what=character(), sep="\n")  
dracula <- scan(file="/Users/Adan Tallman/Desktop/Corpus\_Linguistics\_FSU\_2023-2024/10\_Desagulier/chap 05/dracula.txt", what="char", sep="\n") # Windows  
head(dracula)

## [1] "The Project Gutenberg EBook of Dracula, by Bram Stoker"   
## [2] "This eBook is for the use of anyone anywhere at no cost and with"   
## [3] "almost no restrictions whatsoever. You may copy it, give it away or"  
## [4] "re-use it under the terms of the Project Gutenberg License included"   
## [5] "with this eBook or online at www.gutenberg.org"   
## [6] "Title: Dracula"

tail(dracula)

## [1] "Most people start at our Web site which has the main PG search facility:"  
## [2] " www.gutenberg.org"   
## [3] "This Web site includes information about Project Gutenberg-tm,"   
## [4] "including how to make donations to the Project Gutenberg Literary"   
## [5] "Archive Foundation, how to help produce our new eBooks, and how to"   
## [6] "subscribe to our email newsletter to hear about new eBooks."

##Splitting

The function strsplit() can split texts according to some delimiter. Its a very basic function used to ``parse” texts into elements

comment\_01 <- "Many people think that words are delimited by spaces in English."  
comment\_01

## [1] "Many people think that words are delimited by spaces in English."

strsplit(comment\_01, " ")

## [[1]]  
## [1] "Many" "people" "think" "that" "words" "are"   
## [7] "delimited" "by" "spaces" "in" "English."

whitman\_words <- strsplit(whitman, " ")  
whitman\_words

## [[1]]  
## [1] "When" "I" "heard" "the" "learn'd"   
## [6] "astronomer;"  
##   
## [[2]]  
## [1] "When" "the" "proofs," "the" "figures," "were"   
## [7] "ranged" "in" "columns" "before" "me;"   
##   
## [[3]]  
## [1] "When" "I" "was" "shown" "the" "charts"   
## [7] "" "" "and" "the" "diagrams," "to"   
## [13] "add," "divide," "and" "measure" "them;"   
##   
## [[4]]  
## [1] "When" "I," "sitting," "heard"   
## [5] "the" "astronomer," "where" "he"   
## [9] "lectured" "with" "much" "applause"   
## [13] "in" "the" "lecture-room,"  
##   
## [[5]]  
## [1] "How" "soon" "," "unaccountable,"  
## [5] "I" "became" "tired" "and"   
## [9] "sick;"   
##   
## [[6]]  
## [1] "Till" "rising" "and" "gliding" "out," "I"   
## [7] "wander'd" "off" "by" "myself,"   
##   
## [[7]]  
## [1] "In" "the" "mystical" "moist" "night-air,"  
## [6] "and" "from" "time" "" "to"   
## [11] "time,"   
##   
## [[8]]  
## [1] "Look'd" "up" "in" "perfect" "silence" "at" "the"   
## [8] "stars."

whitman\_words\_unlisted <- unlist(strsplit(whitman, " "))  
whitman\_words\_unlisted

## [1] "When" "I" "heard" "the"   
## [5] "learn'd" "astronomer;" "When" "the"   
## [9] "proofs," "the" "figures," "were"   
## [13] "ranged" "in" "columns" "before"   
## [17] "me;" "When" "I" "was"   
## [21] "shown" "the" "charts" ""   
## [25] "" "and" "the" "diagrams,"   
## [29] "to" "add," "divide," "and"   
## [33] "measure" "them;" "When" "I,"   
## [37] "sitting," "heard" "the" "astronomer,"   
## [41] "where" "he" "lectured" "with"   
## [45] "much" "applause" "in" "the"   
## [49] "lecture-room," "How" "soon" ","   
## [53] "unaccountable," "I" "became" "tired"   
## [57] "and" "sick;" "Till" "rising"   
## [61] "and" "gliding" "out," "I"   
## [65] "wander'd" "off" "by" "myself,"   
## [69] "In" "the" "mystical" "moist"   
## [73] "night-air," "and" "from" "time"   
## [77] "" "to" "time," "Look'd"   
## [81] "up" "in" "perfect" "silence"   
## [85] "at" "the" "stars."

table(whitman\_words\_unlisted)

## whitman\_words\_unlisted  
## , add, and applause   
## 3 1 1 5 1   
## astronomer, astronomer; at became before   
## 1 1 1 1 1   
## by charts columns diagrams, divide,   
## 1 1 1 1 1   
## figures, from gliding he heard   
## 1 1 1 1 2   
## How I I, in In   
## 1 4 1 3 1   
## learn'd lecture-room, lectured Look'd me;   
## 1 1 1 1 1   
## measure moist much myself, mystical   
## 1 1 1 1 1   
## night-air, off out, perfect proofs,   
## 1 1 1 1 1   
## ranged rising shown sick; silence   
## 1 1 1 1 1   
## sitting, soon stars. the them;   
## 1 1 1 9 1   
## Till time time, tired to   
## 1 1 1 1 2   
## unaccountable, up wander'd was were   
## 1 1 1 1 1   
## When where with   
## 4 1 1

length(whitman)

## [1] 8

length(whitman\_words)

## [1] 8

length(whitman\_words\_unlisted)

## [1] 87

grep("tired", whitman) #It just tells you the vector number if you do this

## [1] 5

grep("tired", whitman, value=TRUE) #this is what you need for the full vector

## [1] "How soon , unaccountable, I became tired and sick;"

grep("tired", whitman\_words\_unlisted)

## [1] 56

grep("heaven", dracula, value=TRUE)

## [1] "so pious that she’d be sure to go to heaven, an’ he didn’t want to"   
## [2] "and other women that there is a heaven where we can enter, and that its"  
## [3] "must have one hour that will make the very face of heaven grow black to"  
## [4] "To us for ever are the gates of heaven shut; for who shall open them to"  
## [5] "touch--then pouf! and there comes a flash of light, heaven’s wide,"   
## [6] "great dark came upon us--for even after down-sun the heavens reflected"

sub("'d", "ed", whitman, ignore.case = FALSE)

## [1] "When I heard the learned astronomer;"   
## [2] "When the proofs, the figures, were ranged in columns before me;"   
## [3] "When I was shown the charts and the diagrams, to add, divide, and measure them;"   
## [4] "When I, sitting, heard the astronomer, where he lectured with much applause in the lecture-room,"  
## [5] "How soon , unaccountable, I became tired and sick;"   
## [6] "Till rising and gliding out, I wandered off by myself,"   
## [7] "In the mystical moist night-air, and from time to time,"   
## [8] "Looked up in perfect silence at the stars."

gsub(",", "", whitman)

## [1] "When I heard the learn'd astronomer;"   
## [2] "When the proofs the figures were ranged in columns before me;"   
## [3] "When I was shown the charts and the diagrams to add divide and measure them;"   
## [4] "When I sitting heard the astronomer where he lectured with much applause in the lecture-room"  
## [5] "How soon unaccountable I became tired and sick;"   
## [6] "Till rising and gliding out I wander'd off by myself"   
## [7] "In the mystical moist night-air and from time to time"   
## [8] "Look'd up in perfect silence at the stars."

sub("and", "<conj>and</conj>", whitman, ignore.case = FALSE) #This one just does it once per line

## [1] "When I heard the learn'd astronomer;"   
## [2] "When the proofs, the figures, were ranged in columns before me;"   
## [3] "When I was shown the charts <conj>and</conj> the diagrams, to add, divide, and measure them;"   
## [4] "When I, sitting, heard the astronomer, where he lectured with much applause in the lecture-room,"  
## [5] "How soon , unaccountable, I became tired <conj>and</conj> sick;"   
## [6] "Till rising <conj>and</conj> gliding out, I wander'd off by myself,"   
## [7] "In the mystical moist night-air, <conj>and</conj> from time to time,"   
## [8] "Look'd up in perfect silence at the stars."

gsub("and", "<conj>and</conj>", whitman, ignore.case = FALSE) #This does it through out whole text

## [1] "When I heard the learn'd astronomer;"   
## [2] "When the proofs, the figures, were ranged in columns before me;"   
## [3] "When I was shown the charts <conj>and</conj> the diagrams, to add, divide, <conj>and</conj> measure them;"  
## [4] "When I, sitting, heard the astronomer, where he lectured with much applause in the lecture-room,"   
## [5] "How soon , unaccountable, I became tired <conj>and</conj> sick;"   
## [6] "Till rising <conj>and</conj> gliding out, I w<conj>and</conj>er'd off by myself,"   
## [7] "In the mystical moist night-air, <conj>and</conj> from time to time,"   
## [8] "Look'd up in perfect silence at the stars."