

テキスト分析のためのR入門: Exercises

Ex.1 ファイル操作の基本

1. デスクトップ上に新しいフォルダを作成しなさい。フォルダ名は `R入門20201228` とすること
2. コマンドプロンプトを起動し、`cd` コマンドでカレントフォルダを確認しなさい
 - ヒント: 「ファイル名を指定して実行」(`Win+R`)で `cmd` と入力(Windowsの場合)
3. `cd` コマンドを使って、(1)で作ったフォルダに移動しなさい
 - ヒント: フォルダを右クリック→プロパティから、親フォルダの場所を確認できる(Windowsの場合)
4. `cd` コマンドを使って、親フォルダに移動しなさい
5. `cd ~` を入力して実行しなさい

Ex.2 RStudio

1. RStudioから新規Rスクリプトを作成しなさい
 - ヒント: `File > New File > R Script`
2. (1)で作ったRスクリプトをEx.1で作ったフォルダに保存しなさい。ファイル名は `r_intro.r` とすること
3. 一度RStudio上でファイルを閉じ、(2)で作ったファイルを開き直しなさい
4. エディタ上でRスクリプトに `print('hello')` と入力し、実行しなさい
 - ヒント: 実行したい部分を選択して `Ctrl(command) + Enter` でスクリプトを実行できる

Ex.3 はじめに覚えておくこと

1. コンソールに `# print('comment')` と入力して実行し、実行結果が **出ない** ことを確認しなさい
2. エディタ上で `?mean` と入力して実行し、ヘルプが表示されることを確認しなさい
3. `getwd()` を実行して作業ディレクトリを確認しなさい
4. `setwd()`関数を使ってEx.1-1で作ったフォルダを作業ディレクトリに設定しなさい
5. 以下を順に実行しなさい
 1. `test = 1`
 2. `print(test)`
 3. `print(TEST)` (エラーが出ることを確認する)
 4. `'TEST' == 'test'`
 5. `test == 'test'`

Ex.4 演算

1. [四則演算] $x = 3, y = 4, z = 5$ として以下の値を求めなさい
 1. x と y と z の和
 2. x と y の和を z で割った値
 3. z の x 乗
 4. z を x で割った商とあまり(整数除算)
2. [数学関数] $x = c(1,2,3,4,5)$ として、以下を実行しなさい

1. `sum(x)`
 2. `mean(x)`
 3. `sum(x) / length(x)`
3. [比較演算子] $x=1, y=2$ として以下の計算を実行しなさい
1. `x == y`
 2. `x != y`
 3. `x > y`
 4. `x < y`
 5. `x %in% c(2,4)`
 6. `y %in% c(2,4)`
4. [論理演算] $x = c(T, T, F, F), y = c(T, F, T, F)$ として、以下を実行しなさい
1. `x & y`
 2. `x && y`
 3. `x | y`
 4. `x || y`
 5. `!x`
5. [NAとの計算] 以下を実行しなさい
1. `1 + NA`
 2. `0 * NA`
 3. `sum(1,2,NA)`

Ex.5 ベクトル

1. [ベクトルの作成] 以下のベクトルを作成しなさい
 1. 1から5までの整数を順に並べたベクトル
 2. a,b,c,dのそれぞれの文字を要素とするベクトル
 3. コロン演算子`:`を使って、1から10までの整数を順に並べたベクトル
 4. コロン演算子`:`を使って、8から2までの整数を大きい順に並べたベクトル
 5. `c(1,2,3)`を4つ並べたベクトル
2. [ベクトル演算] $x = 1:5, y = c(2,4), z=c(5, 3, 1)$ として、以下の計算をしなさい
 1. `x + 1`
 2. `y * 2`
 3. `x + y`
 4. `x * z`
 5. `sqrt(x)`
 6. `mean(y)`
 7. `rev(x)`
 8. `sort(z)`
3. [要素へのアクセス] $x = c('a', 'b', 'c', 'd', 'e'), y = 1:6$ として、以下の操作をしなさい
 1. 添字番号を使ってxの3番目の要素を取り出す
 2. 整数ベクトルを使ってxの2番目と5番目の要素を取り出す
 3. xの1番目の要素をfに変更する
 4. 論理ベクトルを使ってxの1番目と4番目の要素を取り出す
 5. yの偶数である要素を取り出す
 6. yの2番目と3番目以外の要素を取り出す

Ex.6 関数

1. [関数の作成] 以下の関数を作成しなさい

1. 実行すると'done.'と表示する関数
 2. 金額（税抜）を入力すると消費税込の金額を返す関数（消費税率は10%とする）※小数点以下がでてよい
2. [制御構文]
1. 以下の関数を作成しなさい。ただし整数以外の入力値については考慮しなくてよいものとする
 1. 入力された整数が偶数なら'even', 奇数なら'odd'と表示する関数
 2. 前問で作成した関数について、for文を使って1から20までの整数値を入力した結果を表示しなさい

Ex.7 stringrによる文字列操作

1. [基本操作] `month.name` は1年の月名のリストである。stringrを用いて以下の問に答えなさい
 1. `str_subset()`関数を用いて名前に'r'を含む月を抜き出しなさい
 2. `str_detect()`関数を用いて各月が名前に'r'を含むかどうかの論理ベクトルを得なさい
 3. `str_length()`関数を用いて各月の文字数を求めなさい
 4. `str_subset()`関数は正規表現を扱える。正規表現を用いて以下の条件にあてはまる月をそれぞれ求めなさい
 1. 名前が'A'で始まる月
 2. 名前が'r'で終わる月
 3. 名前の2文字目が'a'である月
 4. 名前の最後から3番目の文字が'u'である月
 5. 名前の最初の文字が母音である月
 6. 名前の最後の文字が'r'でも'y'でもない月
 7. 名前の末尾が'er'または'ry'で終わる月
 5. `str_extract_all()`関数は文字列の中から正規表現にマッチした部分をすべて抜き出せる。
`sentence='computational social science'`として以下の要素を抜き出しなさい
 1. 単語(アルファベットの連続)
 2. 's'で始まる単語
 3. 'al'で終わる単語

Ex.8 データフレーム

1. [データフレームの操作] `exercises` フォルダに `ex7_1_a.tsv` - `ex7_1_d.tsv` が入っている（以下、ファイルA-Dと呼ぶ）。これはある年における講義科目の学生の成績が入っている架空のデータである。学生は10人いて、社会学(sociology)と心理学(psychology)は必修であるため全員が受講し、経済学(economics)と政治学(politics)は選択科目のため受講していない学生もいる。以下の問に答えなさい
 1. ファイルA-Dをデータフレームとして読み込みなさい(名前は `dfa` - `dfd` とすること)
 2. ファイルAはID1-ID5の学生の、ファイルBはID6-ID10の学生の社会学と心理学の成績である。`rbind()`関数を用いて両者を結合し、新しいデータフレームを作りなさい（名前は `df_ab` とすること）
 3. ファイルCは経済学、ファイルDは政治学の成績である。`merge()`関数を用いて以下の結合結果を確かめなさい。
 1. 経済学と政治学の両方を受講している学生のデータのみ使う
 2. 経済学を受講している学生のデータはすべて使い、政治学のみ受講している学生のデータは使わない

3. 政治学を受講している学生のデータはすべて使い、経済学のみ受講している学生のデータは使わない
4. ファイルC, ファイルDに含まれるすべての学生のデータを使う
4. 前問(3-4)の結合結果を `df_cd` と呼ぶことにする。merge()関数を用いて、`df_ab` と `df_cd` を結合したデータフレームを作成しなさい(名前は `df_abcd` とする)。ただしすべてのデータを使う(all=T)こと
5. `df_abcd` について、学生ひとりひとりの成績の平均値を求め、列名を `GPA` として新しい列を追加しなさい
6. `df_abcd` をtsvファイルとして書き出しなさい
1. [データフレームと関数] `USJudgeRatings` は弁護士による米国高等裁判所における州裁判官の評価である。なお、各変数の説明は以下の通りである：

1	[,1]	CONT	Number of contacts of lawyer with judge. # 弁護士と裁判官の接触回数
2	[,2]	INTG	Judicial integrity. # 司法の廉潔性
3	[,3]	DMNR	Demeanor. # 態度
4	[,4]	DILG	Diligence. # 勤勉さ
5	[,5]	CFMG	Case flow managing. # 事案管理
6	[,6]	DECI	Prompt decisions. # 即決性
7	[,7]	PREP	Preparation for trial. # 審理への準備
8	[,8]	FAMI	Familiarity with law. # 法律への熟知性
9	[,9]	ORAL	Sound oral rulings. # 口頭決定の健全性
10	[,10]	WRIT	Sound written rulings. # 書面決定の健全性
11	[,11]	PHYS	Physical ability. # 身体能力
12	[,12]	RTEN	Worthy of retention. # 留保の適切さ

- `USJudgeRatings` をデータセットとして以下の問に答えなさい
 1. psychパッケージをインストールしていないならば `install.packages('psych')` を実行し、インストールしなさい
 2. library()関数でpsychパッケージを読み込んだ上でdescribe()関数を適用し、要約統計量を算出しなさい
 3. データセットの行数と列数を求めなさい
 4. colMeans()を用いて、各変数について裁判官全体の平均値を求めなさい

Ex.9 作図

1. [棒グラフと散布図] `USArrests` は1973年の米国における10万人当たりの暴力犯罪の件数および都市人口データである。 `USArrests` をデータセットとして用いて以下の問に答えなさい
 1. データセットの先頭5件を抽出し、殺人(Murder)の値を縦棒グラフで表示しなさい
 2. データセットの末尾5件を抽出し、暴行(Assault)の値を横棒グラフで表示しなさい
 3. plot()関数を用いて三種類の暴力犯罪(Murder, Assault, Rape)について、総当たりで散布図を作成しなさい
 4. pairs()関数を用いてすべての変数についての散布図行列を作成しなさい
2. [時系列] `AirPassengers` は1949年から1960年にかけての月別国際航空旅客者数のデータである。
 1. plot()関数を用いて折れ線グラフを作りなさい
 2. png()関数を用いて(1)で作ったグラフを保存しなさい
3. [群間データ] `PlantGrowth` は植物の成長に関する実験データである。統制条件と2つの実験条件の形三つの群(group)が用意され、植物の重量(weight)が測定された
 1. table()関数を用いて各条件の標本数を求めなさい

2. 条件間を比較する箱ひげ図を作りなさい

Ex.10 dplyrによるデータハンドリング

1. `ToothGrowth` はビタミンCの用量と与え方がモルモットの歯の長さに与える影響を調べた実験データである。これをデータセットとして用い、`dplyr`パッケージを読み込んだ上で以下の問に答えなさい
 1. パイプ演算子を使ってデータセットに`head()`関数を適用しなさい
 2. パイプ演算子と`filter()`関数を使って以下の操作をしなさい
 1. `supp` が `oj` であった個体を抽出する
 2. `supp` が `vc` で、`dose` が2である個体を抽出する
 3. `supp` が `vc` である個体、または `dose` が2である個体を抽出する
 3. パイプ演算子と`select()`関数を使って `len` 列と `supp` 列を抽出し、それぞれ「length」「supp_type」とリネームして表示しなさい
 4. パイプ演算子と`arrange()`関数を使って、`len` 列で昇順にソートした結果を表示しなさい
 5. パイプ演算子と`summarize()`関数を使って以下の操作をしなさい
 1. `len` の最大値・最小値・中央値・平均をそれぞれ求める
 2. `group_by`を `supp` 列に適用し、群ごとの最大値・最小値・中央値・平均をそれぞれ求める