自然言語処理
―準備、Pandas―

https://satoyoshiharu.github.io/nlp/

100本ノック第2章の位置づけ

- 100本ノック第2章の課題集は、Unixコマンドというタイトルになっています。しかし、unixコマンドは確認のために使う補助的な位置づけとし、Pythonで書く課題とみなします。

- 課題の内容は、自然言語処理の準備として、テキストデータのファイルを操作するものです。PythonのPandasというパッケージを利用します。データサイエンス系で重宝されているものです。

- 機械学習やニューラルネットのプログラミングは、ほぼライブラリメソッドを使うだけで、実際にはデータを準備し加工する工程に時間がかかります。その加工で活躍します。

Unix コマンド

UNIXコマンド？

- 正確に言うと、Unix/Linux で使える bash (Bourne Again Shell) コマンド・ライン・インターフェイスのコマンドのことである。Unix コマンド shell は、たくさんの新しいアイデア（pipe, redirection, stdin/out/err など）をコンピュータ業界に普及させた、いまだ、OSコマンドshellの標準である。
  - ただし、1970年代からの技術なので、覚えにくい（各コマンドの名前が学者趣味、各コマンドの細かい引数がわかりにくい、viがWYSWYGの前時代）。
- Linux系のVMをいじるときなどに必要となる。

よく使う unix コマンド

- cd パス：パスへ移動する
  - 例：cd ~：ユーザのデフォルトルートへ移動する、cd /：OSのルートに移動する、cd ..：一つ上へ移動する
- pwd：現在のフォルダーのパスを表示する
- ls：現在のフォルダーのファイルの一覧を表示
- cat file-path：fileの中身を表示する
  - 例： cat ./file | wc ：fileに含まれる単語数を数える
- パス/プログラム名：プログラムを起動する
  - .例： /myprogram

よく使う unix コマンド

- pipe：あるコマンドの標準出力を、別のコマンドの標準入力として、ストリーム結合する。
  - 例：man xxx | more ： コマンドxxxのマニュアルを、ページごとに表示する
  - 例：ps –ax | grep python ：pythonという文字列を含むプロセスを表示する
- redirection：あるコマンドの標準出力や標準エラーを、指定ファイルに流す。
  - 例：sort ./file > ./file2：fileの中身をソートしてfile2へ
  - 例：cat ./file >> ./file2 ：fileの中身をfile2へ追記する

Pandasパッケージ

pandas

- Python package の pandas は、データフレームという、EXCELのような表形式データに対し、SQLのような演算を提供するもの。

- 参考（英語）
  - https://pandas.pydata.org/pandas-docs/stable/getting_started/intro_tutorials/index.html
  - https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html#min
  - https://pandas.pydata.org/pandas-docs/stable/reference/index.html

pandas参考資料

- データ構造、参照、IOなどの基本がコード付きでまとまっている
  https://aiacademy.jp/texts/show/?id=23
- 描画機能 pyplotとからむサンプル
  https://qiita.com/sqrtxx/items/a5e234c9deb5e503ca7b
- ChainerのPandas入門
  https://tutorials.chainer.org/ja/11_Introduction_to_Pandas.html
- 視聴回数の多い教育動画
  - https://www.youtube.com/watch?v=jJtOroFFYxU
  - https://www.youtube.com/watch?v=EI40pSM8TJc
  - 長いの　https://www.youtube.com/watch?v=XfoYeWCzjac

課題10～19

- 「100本ノック」の２章の課題10～19を解いてみましょう。
- 「Pandas.ipynb」というノートをコピーし、冒頭の準備、基本事項をやった後、各課題のセクション下のコードセルに解答コードを書き、実行ログを残してください。
- ネットに解答集がいくつか見つけられます。講師の解答案も提供します。力をつけるため、なるべくそれらを見ないで自力でやってから、見てください。回答をコピペするのでは、力は少しもつかず、やった感を作るだけの、時間の無駄です。一方、考えながら、書くために必要なことを調べるのは成長につながる大切な時間となります。

確認クイズ

- Pandas_確認クイズ.ipynb で力を試してください。出力を指定しているので、それが出力できればOKです。