自然言語処理
―準備、正規表現―

https://github.com/satoyoshiharu/nlp

100本ノック第3章の位置づけ

- 100本ノック第3章の課題集は、Pythonの正規表現です。
- 課題の内容は、自然言語処理の貴重なデータ源として巨大な言語テキストデータである Wikipedia からテキストを抜きだすために、正規表現を利用しています。Wikipediaのソースは、HTMLをはじめとするいろんなタグが複雑に混ざっているからです。ただ、BeatifulSoupというパッケージがあり、それを使うだけで、Wikipediaテキストを入手するという目的は達成できます。
- それよりも、正規表現は、実は、デバッグやネットワークサーバーのログから欲しい情報だけ抜き出すなど、様々な場面で日常的に活躍します。
- Pythonの正規表現は、基本的な部分だけ押さえて、あとは力業の世界です。必要になったときに、調べながら書くといいです。今は基本だけ押さえましょう。

正規表現資料

- Python reパッケージのマニュアル
  - https://docs.python.org/ja/3/library/re.html
- わかりやすい入門解説
  - https://userweb.mnet.ne.jp/nakama/
  - https://takano.hatenablog.jp/entry/2019/03/22/053026
  - https://dotinstall.com/lessons/basic_regexp_v2

課題20～23,29

- 以下のスライドおよびリンクした資料などを参考に、「100本ノック」の３章の課題20～23,29 を解いてみましょう。正規表現は、基本的な部分に慣れれば、あとは力ワザなので、基本的な課題20～23と、WikipediaのAPIに触れる29だけやりましょう。
- 「正規表現.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。
- *正規表現は、ソフトエンジニアやる限りどこにでも出てきて、コードを書く機会が多いし、逆に特にPythonで慣れる必要度が小さいです。それに、基本がわかっていればあとは組み合わせで力ざわの世界。マニュアルがあれば、やればできる世界なので、課題ノートの基本事項をやったら、あとは軽めに流して結構です。*