タイトル（仮）「Python言語を用いたデータ分析プログラミング〜ビッグデータに対するテキストマイニングからWebスクレイピング技術まで」

一橋大学イノベーション研究センター・田村龍一

## 概要

近年欧米のデータサイエンティストだけでなくエコノミストにも利用されている汎用プログラミング言語「Python（パイソン）」について[[1]](#footnote-1)、特にビッグデータの解析やWeb上に公開されている情報の自動取得（Webスクレイピング）に目的を置いたプログラミング技法の解説を行う。その際、Pythonの言語仕様といった一般的解説は最小限にとどめ、上記2つの技術に必要なプログラミング知識の取得と運用に集中する。そのため講義は演習の形式をとり、ビッグデータやWeb公開情報から必要な情報の抽出、結果の整理・可視化といった課題をPythonプログラミングによって解決していく。

　Webスクレイピング技術は元来検索解析技術のひとつとして発達してきたが、2000年代より経済実証分析の分野でも積極的な活用が行われている。最近の研究を挙げると、Mayzlin, Dover and Chevaluer ([American Economic Review, 2014](http://dx.doi.org/10.1257/aer.104.8.2421))ではホテル評価サイト（TripAdvisorとExpedia）における宿泊レビュー投稿データをスクレイピングによって収集し、業者の「なりすまし」レビューがオンライン評価システムに与える影響を分析している。Roberto Rigobon及びAlberto Cavallo 氏によるMITの「[The Billion Prices Project](http://bpp.mit.edu)」では、数十万種類の製品に関してその特徴と価格を世界中のウェブサイトから日々スクレイピングプログラムによって収集し、チリや東日本で発生した大地震が価格と供給に与えた影響の分析に使用している(Cavallo, Cavallo, and Rigobon, [Review of Income and Wealth, 2014](http://www.mit.edu/~afc/papers/Cavallo-NaturalDisasters.pdf))。その他、Webスクレイピング技術を実証経済分析に応用した例はEdelmanのサーベイ ([Journal of Economic Perspectives, 2012](https://www.google.co.jp/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCMQFjAA&url=https%3A%2F%2Fwww.aeaweb.org%2Ffull_issue.php%3Fdoi%3D10.1257%2Fjep.26.2&ei=ypGHVarTLMjFmAW6x5GYDw&usg=AFQjCNFXcqzvJZQQyc76HmYJQefsYFxvDQ&sig2=qpIuNMwFYHK1Tet9irx9nQ&bvm=bv.96339352,d.dGY))に詳しい。

## 目標

1. Pythonの実行環境（Anaconda）に習熟し、データ解析の定番ライブラリであるNumpy, Scipy, Matplotlib、Webスクレイピングの定番ツールであるBeautifulSoup, SeleniumといったPythonが提供する有用なライブラリの理解と運用能力を涵養する。
2. ビッグデータを始めとする大量の情報の管理と操作に便利なSQLite3データベースの構造を理解し、これをPythonを使って自在に操作できるようになる。
3. Web上に流通するデータフォーマット(XML, SGML, JSON, CSVなど)を理解し、これらに対してPythonのDOMパーサや正規表現を適用することによって、必要な情報を的確に取得できるようになる。

## 想定する受講者

1. 少なくとも数100万件の数値、日付や文字の混在したデータについて、必要な情報を抽出し、各情報の関係性について見通しよく把握したいと希望する方。
2. Web上に公開されているあらゆる情報を、プログラミングによって収集しようという意欲があり、この作業を効率よく運用したいと希望する方。
3. Pythonの利用経験は問わない。しかし、STATA、Matlab、Rといった統計・計量経済分析ソフトウエアを用いたプログラム経験があることが望ましい。

## 要件

ネットワークを利用したプログラミングを行うため、インターネットに接続できるラップトップ（スペックは問わない）を持参のこと。インストール作業を行うので、管理者権限が必要である。

## 教材

スライドなどは当方で準備する。ビッグデータの素材として、USPTO発明者・企業ファイル（約3.9GB、935万件の発明者・企業情報）、Webスクレイピングの対象としてRePEC上に公開されている書誌情報を扱う予定である。

## 参考書

Thomas Sargent and John Stachurski, “QUANTATIVE ECONOMICS with Python,” available at http://quant-econ.net/\_static/pdfs/py-quant-econ.pdf

1. 2014年にはニューヨーク大学Thomas J. Sargent氏らによってPythonを用いた経済分析に関するチュートリアルサイト「Quantitative Economics」が開設されている：http://quant-econ.net/ [↑](#footnote-ref-1)