Workshop on "Medium Data" and Text Analysis

Jan R. Riebling

January 21, 2018

Summary

Concepts like "Big Data" and "Data Science" have received considerable attention in recent years both in the world of business as well as among academics. But how well is sociology prepared to deal with new data types and methods, resulting from the rise of digital information and communication technologies? Through practical exercises and critical discourse this workshop tries to show the possibilities of these new data sources as well as their limitations. The main focus will be on the analysis of textual data, since this is one of the most prevalent data types produced by the new information technologies. Because of the limited time frame and the extensive topic only a broad overview can be given. Therefore, this workshop aims to be a starting point rather than a all inclusive manual. Instead of detailed cookbooks and recipes, general pythonic thinking and the ability to find your own solutions are fostered. The workshop programm consists of three parts. A general yet brief introduction to the Python programming language, a discussion of the properties of "Medium Data" as well as how to manage it and an overview of methods for the quantitative analysis of text. This programm is spread out over two workshop sessions each providing six hours of content. A informal Q&A session will be provided on the two days following the workshop sessions. Workshop participants are encouraged to bring their own specific problems and questions to the forefront of the discussion.

Syllabus

This is a preliminary syllabus and can be subject to change.

1. chapter (2018/01/25)

1. The "Medium Data" problem.
    - Complex and process-generated data vs. "Big Data".
    - Definitions, problems and solutions.
2. Introduction to Anaconda Python I.
    - Basic tools and checking the installation.

- Primitive types, control flow and loops.

3. Introduction to Anaconda Python II.
   - Arrays and DataFrame.
   - Web crawling and scraping.
4. Text and data mining.
   - String operations in Pandas.
   - Regular expressions.

2. chapter (2018/02/21)

1. Introduction to computational linguistics.
   - The Natural Language Toolkit (NLTK) for Python.
   - Simple text analysis.
2. Advanced text handling.
   - Tokenization.
   - Stemming vs. Lemmatization.
3. Topic Modelling
   - Latent dimensions in text.
   - The use of the gensim package.
4. Machine Learning.
   - Training and testing classifiers.
   - Practical applications in SciKit-Learn.

Installation guide

The workshop uses Anaconda Python as package manager/development environment. This also includes the to Jupyter Notebook as the main environment for programming and analysis. It is strongly recommended to have a working installation of Anaconda Python before the workshop starts, since I will be unable to give extensive techsupport once the workshop starts, because of the limited time available. In order to install the programm, follow these steps:

1. Go to https://www.continuum.io/downloads. Select and download the Anaconda 5.0.1 (64bit) installer suitable to your operating system.
2. Follow the provided instructions on the website to install Anaconda.
3. Open a command-line interface to your OS (shell). Depending on your operating system, this could be a bash shell(Linux), a terminal(OSX) or a PowerShell(Windows). If you are unsure what a shell and respectively a command-line interface are, this Wikipedia article should provide a good starting point.

4. Enter the following command on the shell:
   `conda -V`
   The output should confirm the installation of the `conda`package manager by returning its name and version number.
5. Troubleshooting: If conda can not be found after the installation try the *Anaconda Prompt*. This is a pre-configured shell for Windows Users, which sets the correct environmental variables. It can be found in the "Anaconda 64bit" startmenu entry.
6. Start a Jupyter Notebook by typing:
   `jupyter notebook`
   Your standard browser should open and display the content of the directory in which the shell command was executed. If this step fails, use the Anaconda Launcher delivered with the installation (only available on Mac and Windows).
7. Go back to the shell window and shut down the Notebook server by pressing `ctrl+C` twice.
8. Feel free to experiment with the ressources below or write me an email if any problems manifest themselves during the installation process.

Resources

Because of the limited amount of time in the workshop it is strongly advised to familiarize yourself with the basics of Python before the start of the workshop. The Python 3 Tutorial is a good starting place for general Python training, which should be combined with the Introduction to the Jupyter Notebook.

Essentials

Some online resources to prepare for the workshop as well as for further studies:

- Introduction to the Jupyter Notebook.
- Python 3 Documentation.
- Python 3 Standard Library. Keep this under your pillow.
- Conda Documentation.
- Gallery of Notebooks.
- Python 3 Tutorial.
- 10 Minutes to Pandas.

Anaconda IPython

- Anaconda Download
- IPython Docs
- Notebook Gallery

Scientific computing

- Numpy for fast, n-dimensional arrays
- SciPy for linear algebra and some statistics.
- StatsModels provides wide variety of statistical models.
- Pandas for flexible and fast data structures.
- scikit-learn for machine learning algorithms and toolchains.
- Matplotlib for all your graphics needs.
- seaborn for additional aesthetics.
- ggplot implements the *Grammar of Graphics* in Python.
- NetworkX implements graphs and algorithms for network analysis.

Books on the subject

- Downey, Allen B. 2012. Think Python. How to Think Like a Computer Scientist. O'Reilly Media, Incorporated. (online).
- Downey, Allen B. 2014. Think Stats: Exploratory Data Analysis. 2 edition. Sebastopol, CA: O'Reilly Media. (online).
- Downey, Allen B. 2012. Think Complexity: Complexity Science and Computational Modeling. 1 edition. Beijing u.a.: O'Reilly Media. (online).
- and basically everything else from Green Tea Press.
- Sweigart, Al. 2015. Automate the Boring Stuff with Python: Practical Programming for Total Beginners. 1 edition. San Francisco: No Starch Press. (online).

General Q&A ("techsupport")

- "Try turning it on and off again!"
- Online communities: e.g. stackoverflow.com, stackexchange.com, etc.
- Project repositories: e.g. Github