UNIVERSITEIT TWENTE.

Data Science [201400174]

Course year 2022/2023, Quarter 1B

DATE
November 11, 2022

EXCERPT

Information Extraction Using Natural Language Processing [IENLP]

TEACHERS
Faizan Ahmed
Ellen-Wien Augustijn
Nacir Bouali
Faiza Bukhsh
Rolf de By
Karin Groothuis-Oudshoorn
Maurice van Keulen
Mahdi Khodadadzadeh
Elena Mocanu
Estefania Talavera
Brenda Voorthuis
Shenghui Wang

COURSE COORDINATOR
Karin Groothuis-Oudshoorn (quartile 1A)
Maurice van Keulen (quartile 1B)
Faizan Ahmed (quartile 2A)

PROJECT OWNERS
Faiza Bukhsh
Karin Groothuis-Oudshoorn
Maurice van Keulen
Elena Mocanu
Mannes Poel
Michel van Putten
Mohsen Jafari Songhori
Luc Wismans

Topic 3

Information Extraction Using Natural Language Processing [IENLP]

3.1 Introduction

This course discuses basic approaches in natural language processing, including text normalization, regular expressions, language modeling, text classification and information extraction.

3.1.1 Global description of the practicum and project

In the assignments, you will learn about some basic tasks in Natural Language Processing and Information Extraction. The assignments will also introduce you to some helpful tools in the field. You will learn how to write regular expressions and to build a text classifiers.

You also will learn about named entity recognition (NER) using the Stanford NER tool.

3.1.2 Study material and tools

The tutorials in the nlp-ie folder of `https://github.com/chseifert/tutorials` might come in handy. A very comprehensive resource is the online pyhton nltk book `https://www.nltk.org/book/`. **Note**: Links to those resources are given in the details of each assignment.

3.1.3 Deliverables and obligatory items

The detailed descriptions of the deliverables are shown in each assignment. Upload a ipython notebook (.ipynb) with your solutions and also a PDF printout (File → Print Preview in the ipython Notebooks.) for each assignment separately.

3.2 Description of the practical assignments

3.2.1 Datasets

You will need the following data sets. All of them are available as corpora in python (so no need to download them manually)

- Shakespeare's Hamlet (see `https://www.nltk.org/book/ch02.html`, Section 1.1)
- The firefox discussion forum (see `https://www.nltk.org/book/ch02.html`, Section 1.2)
- The Reuters corpus (see `https://www.nltk.org/book/ch02.html`, Section 1.4)

You can download them using the `nltk.download()` command, after loading the nltk library (`import nltk`).

3.2.2 Prerequisites

For the code examples to work we need python, its scientific libraries and a jupyter notebook server. There are many possibilities to install python, SciPy and the jupyter notebook server. We recommend to use the "all-in-one"-solution Anaconda, which is available for all (major) platforms, including Linux, OS X and Windows.

Anaconda can be downloaded from `https://www.anaconda.com/download`. Once installed and started, you should see the screen in figure 3.1. Clicking on `jupyter notebook` starts the notebook server in the background an opens a browser window (see figure 3.2) in which you could load an existing notebook or write your code from scratch.

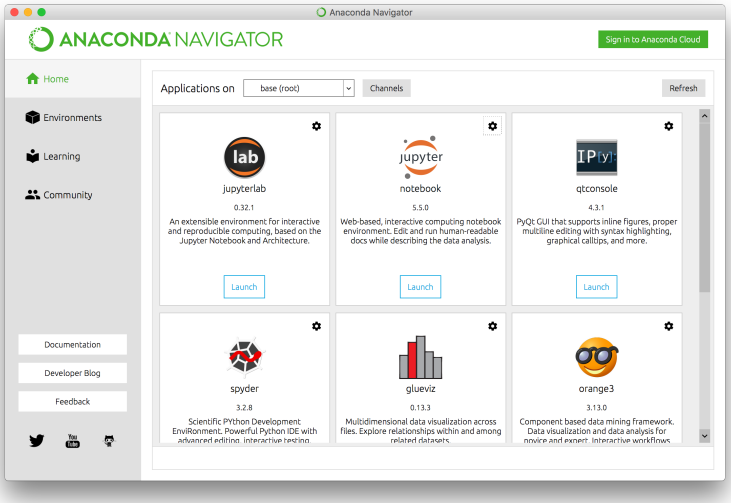

Figure 3.1: Start screen of Anaconda Navigator

If you have want to try whether some code would run on a "clean" installation, you could try the jupyter test webpage `http://jupyter.org/try` (see figure 3.3)[1].

Jupyter notebooks not only contain the python code, but also textual annotations in the markdown annotation language. Thus, a jupyter notebook (indicated by the file ending "ipynb") can not directly be run by a python interpreter (file ending "py").

[1] Depending on the server load it might happen that you can't connect to the server, in which case an informative error message is displayeed

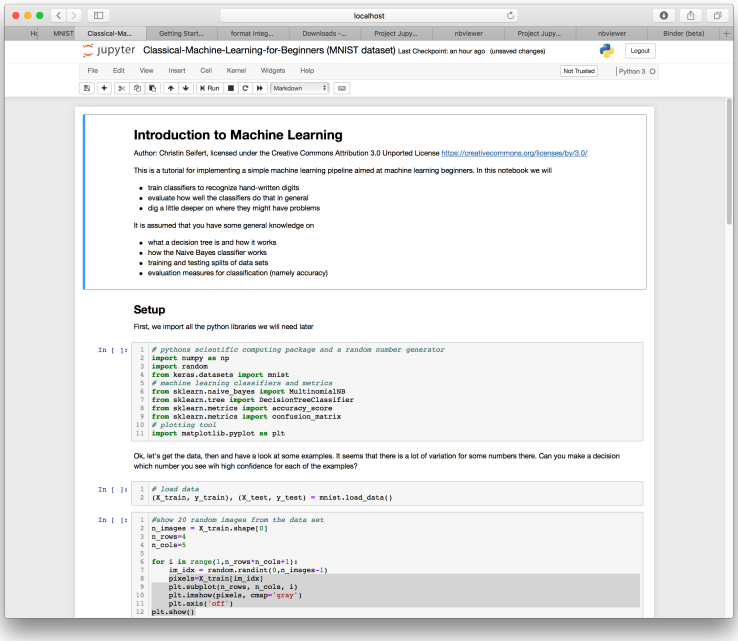

Figure 3.2: A jupyter notebook in the browser

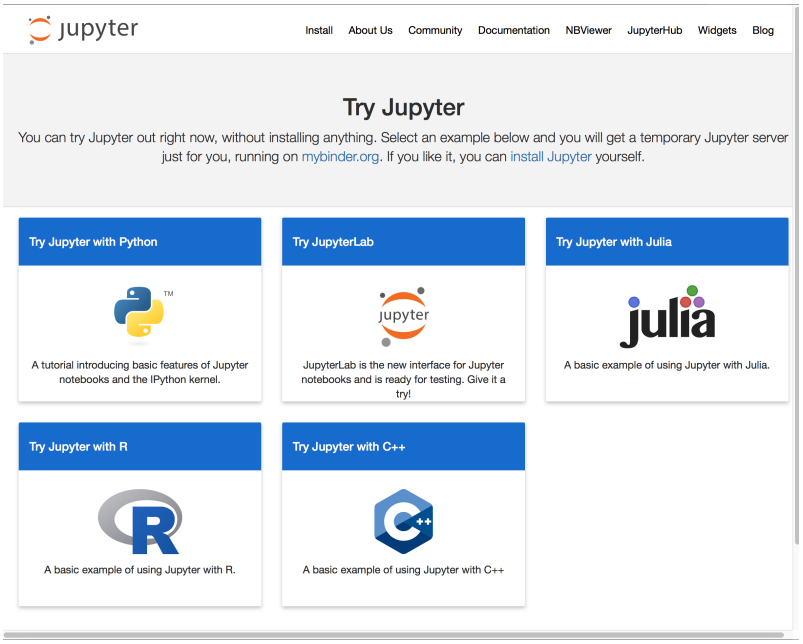

Figure 3.3: Online server for jupyter notebooks

3.2.3 Assignment 1: Regular Expressions

Introduction: Regular expressions is a basic method of extraction information from text. In this task you should learn how to write a regular expression to extract specific information from text.
Task: Create one ore more regular expressions to extract i) all URLs and ii) all keyboard shortcuts (e.g. CRTL+A) from the **firefox discussion forums**. Upload a ipython notebook (.ipynb) with your solutions and also a PDF(File → Download as → PDF via PDFLatex, in the ipython Notebooks.)

3.2.4 Assignment 2: Named Entity Recognition

Introduction: Named entity recognition (NER) is an important task of many information extraction systems. NER seeks to locate and classify elements in text into pre-defined categories such as the names of persons, organizations, locations, etc. In this task you should learn how to apply existing NER systems on your dataset and how to evaluate them.

Task: Find out how many different persons are in the **Hamlet corpus**. How many if you use the 3, 4 and 7-classes tagger? Upload a ipython notebook (.ipynb) with your solutions and also a PDF (File → Download as → PDF via PDFLatex, in the ipython Notebooks.)

3.2.5 Assignment 3: Text Classification

Introduction: Text classification is widely used, for instance for sorting news articles, deciding whether emails are spam or for detecting offensive tweets.

Task: Perform text classification on the **Reuters corpus**. Try at least two different ways of text preprocessing and compare the results. What can you conclude? Also include a short discussion on possible improvements in your notebook. Upload a ipython notebook (.ipynb) with your solutions and also a PDF (File → Download as → PDF via PDFLatex, in the ipython Notebooks.)

3.2.6 Some Hints:

- If you get an error indicating that the stanford NER tools can not be found, you might want to check the path specification. Paths on Windows need to be escaped (since the backslash is itself a escape character in python).
- There have been some troubles loading the Reuters corpus in ipyhton-Notebooks (and just there). Try running the same code in a python environment (without the notebooks). The corpus that is downloaded with the corpus downloader from nltk comes as a zipped version. On some operating-system/python configurations the source files of the corpus need to be unzipped, so locate the files and unzip them in the same directory.
- If you are not yet familiar with python, the following concepts should be mastered before diving into NLP with python.
  - Basic pyhton concepts: numbers, strings, loops (`https://docs.python.org/3/tutorial/introduction.html`) and more complex control flow including basic knowledge about functions (`https://docs.python.org/3/tutorial/controlflow.html`)
  - Data structures: especially lists, sets and dictionaries. Also lists of lists come in handy (tutorial here `https://docs.python.org/3/tutorial/datastructures.html`)
  - You might also want to know some basics about exceptions to be able to figure out what went wrong when something went wrong (`https://docs.python.org/3/tutorial/errors.html`)