Text Mining

Emanuele Guidotti

2019/2020

Why text mining?

- Text is everywhere
- Text data is growing fast
- Approximately 80% of all data is estimated to be unstructured, text-rich data

What can be done with text?

- Parse text
- Find / Identify / Extract relevant information from text
- Classify text documents
- Search for relevant documents
- Sentiment analysis
- Topic modelling
- . . .

Let's start!

Python Installation

Python Installation

Install **Anaconda** from https://www.anaconda.com/distribution/ (Python version 3.7). This will set up:

- Python
- Jupyter Notebook for interactive coding
- Conda for package, dependency and environment management

Install **PyCharm** from https://www.jetbrains.com/pycharm/ (free version), an integrated development environment for Python.

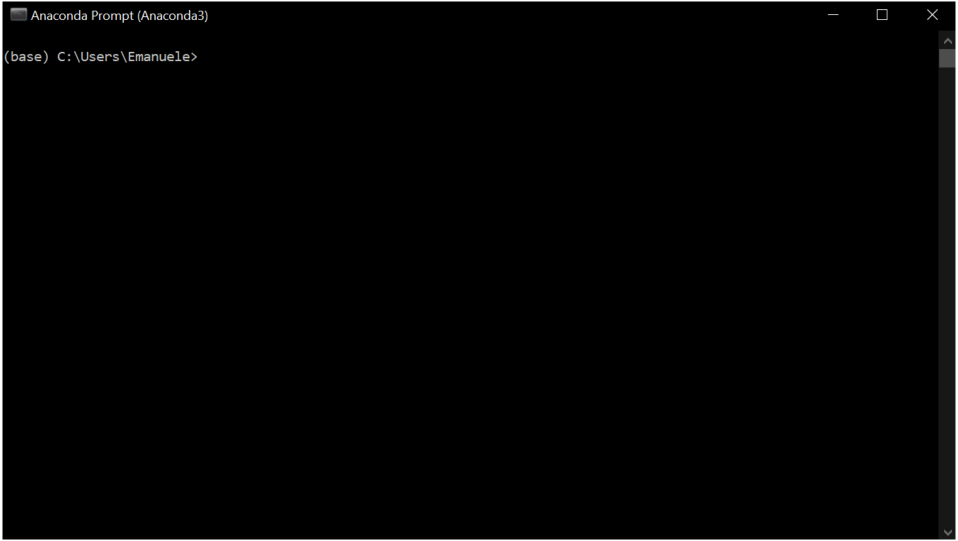
Python environments

A virtual environment is a tool that helps to keep dependencies required by different projects separate by creating isolated python virtual environments for them. This is one of the most important tools that most of the Python developers use.

Virtual Environment should be used whenever you work on any Python based project. It is generally good to have one new virtual environment for every Python based project you work on. So the dependencies of every project are isolated from the system and each other.

Python environments with Conda

Open Anaconda Prompt



```
(base) C:\Users\Emanuele>
```

Check conda is installed and in your PATH

```
conda -V
```

Check conda is up to date

```
conda update conda
```

Create a virtual environment for your project

```
conda create -n yourenvname python=x.x
```

Install additional Python packages to a virtual environment

```
conda install -n yourenvname [package]
```

Delete a no longer needed virtual environment

```
conda remove -n yourenvname --all
```

List environments

```
conda env list
```

Viewing a list of the packages in an environment

```
conda list -n yourenvname
```

Additional resources:

- Cheat Sheet
- Conda Docs

Set up the ‘text-mining’ environment

Conda environment

Create the environment

```
conda create -n text-mining python=3.7
```

Activate the environment

```
conda activate text-mining
```

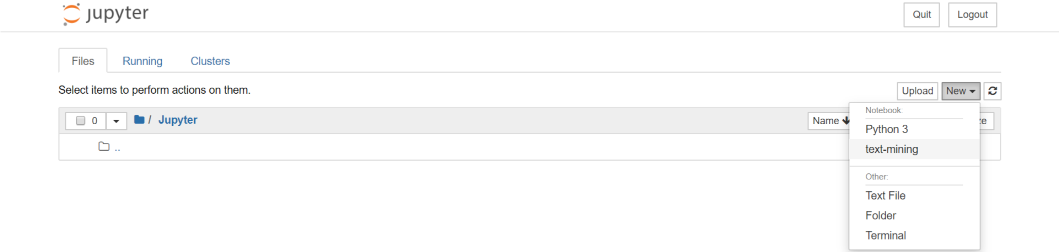
Link the environment to Jupyter

Install additional packages needed for Jupyter

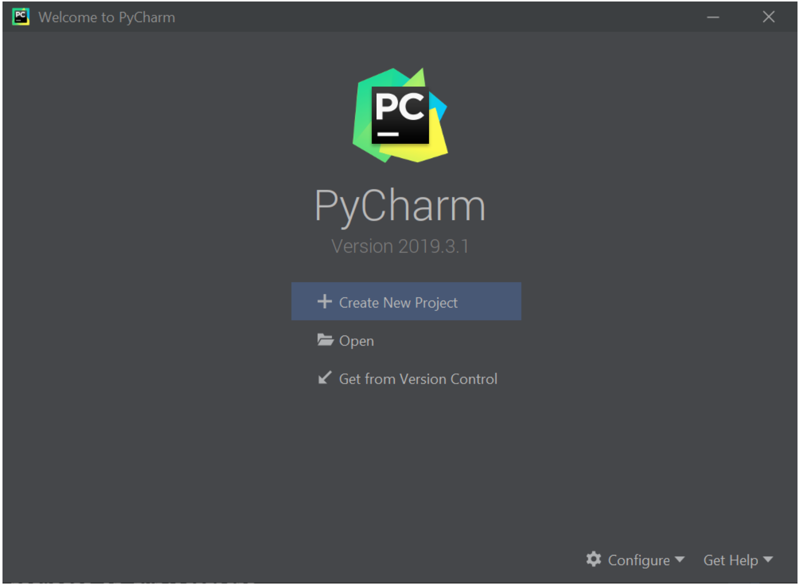
```
conda install -c anaconda ipykernel
```

Make the environment available in Jupyter

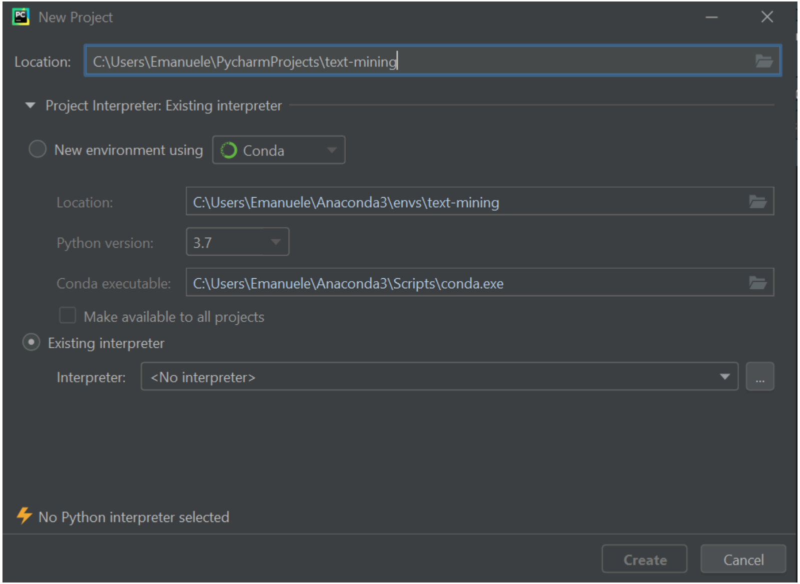
```
python -m ipykernel install --user --name=text-mining
```



Link the environment to PyCharm



Link the environment to PyCharm



New Project

Location: C:\Users\Emanuele\PycharmProjects\text-mining

Project Interpreter: Existing interpreter

○ New environment using Conda

Location: C:\Users\Emanuele\Anaconda3\envs\text-mining

Python version: 3.7

Conda executable: C:\Users\Emanuele\Anaconda3\Scripts\conda.exe

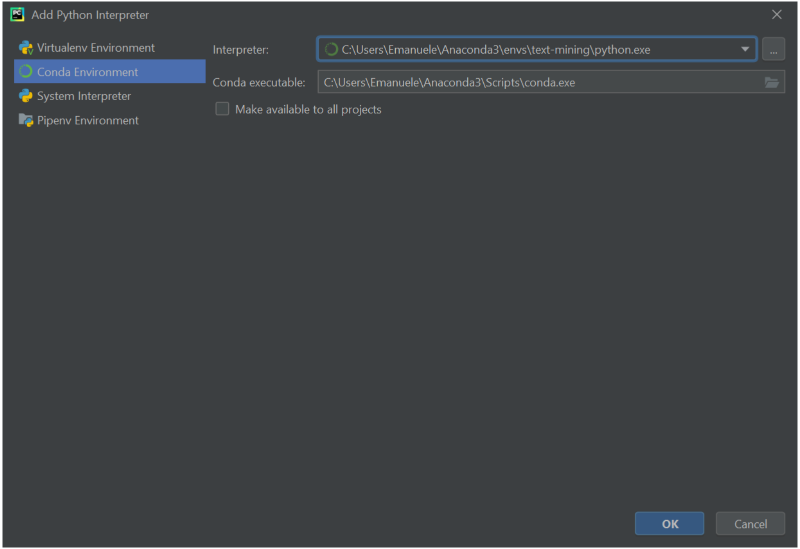
□ Make available to all projects

● Existing interpreter

Interpreter: <No interpreter>

No Python interpreter selected

Link the environment to PyCharm

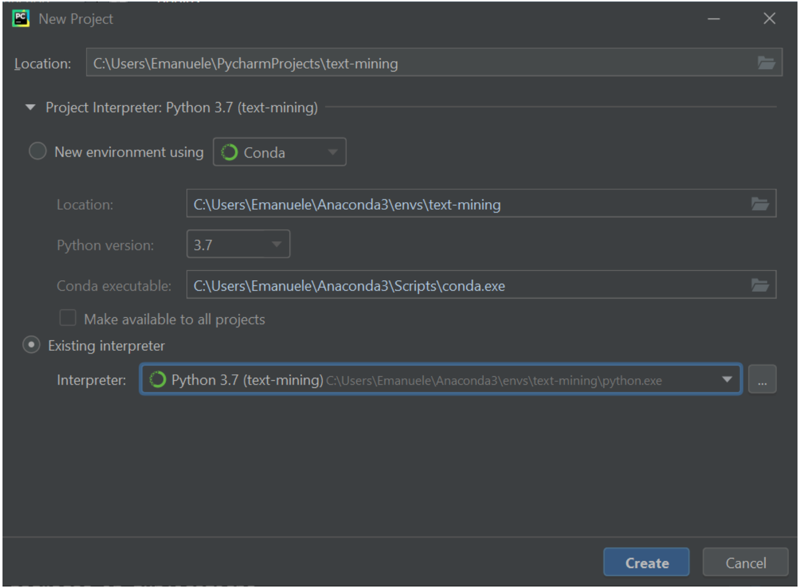


Interpreter: C:\Users\Emanuele\Anaconda3\envs\text-mining\python.exe

Conda executable: C:\Users\Emanuele\Anaconda3\Scripts\conda.exe

□ Make available to all projects

Link the environment to PyCharm



New Project

Location: C:\Users\Emanuele\PycharmProjects\text-mining

Project Interpreter: Python 3.7 (text-mining)

○ New environment using Conda

Location: C:\Users\Emanuele\Anaconda3\envs\text-mining

Python version: 3.7

Conda executable: C:\Users\Emanuele\Anaconda3\Scripts\conda.exe

□ Make available to all projects

● Existing interpreter

Interpreter: Python 3.7 (text-mining) C:\Users\Emanuele\Anaconda3\envs\text-mining\python.exe

Install additional packages

Must-have packages

```
conda install -n text-mining numpy pandas
```

NLP packages

```
conda install -n text-mining nltk spaCy
conda install -n text-mining -c conda-forge textacy
```

ML packages

```
conda install -n text-mining scikit-learn
```

SpaCy models: https://spacy.io/models/

```
conda activate text-mining
python -m spacy download en_core_web_sm
```

Additional packages

```
conda install -n text-mining beautifulsoup4
```

Character Encodings

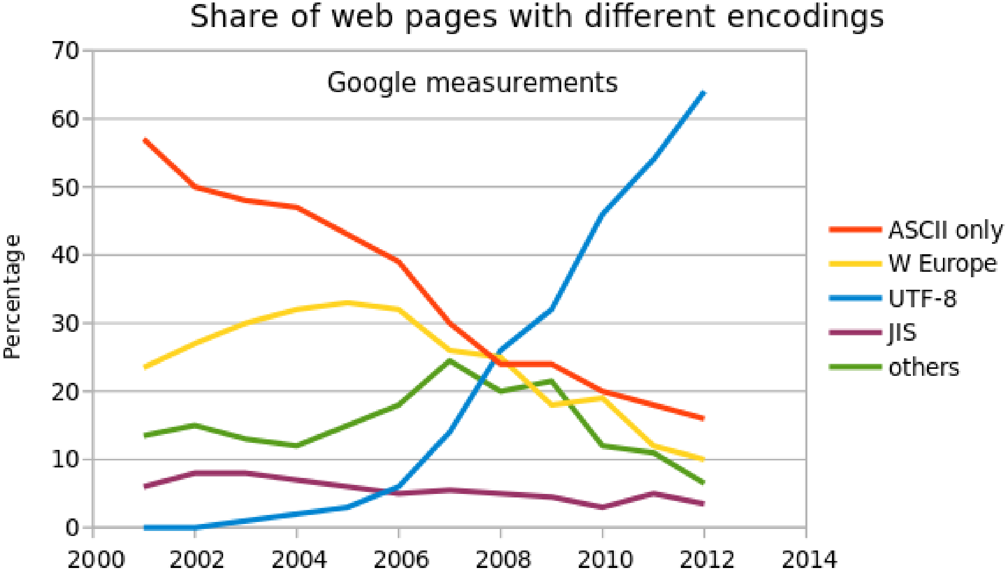
What is a character encoding?

- Words and sentences in text are created from **characters** (á, h, 8, ô, . . .)
- The characters are stored in the computer as one or more **bytes**
- A **character encoding** provides a key to unlock the code. It is a set of mappings between the bytes in the computer and the characters

When you input text using a keyboard or in some other way, the character encoding maps characters you choose to specific bytes in computer memory, and then to display the text it reads the bytes back into characters.

Character Encodings

- ASCII: American Standard Code for Information Interchange
- IBM EBCDIC
- Latin-1
- JIS: Japanese Industrial Standards
- CCCII: Chinese Character Code for Information Interchange
- EUC: Extended Unix Code
- Numerous other national standards
- **Unicode and UTF-8**


Share of web pages with different encodings
Google measurements
Percentage
70
60
50
40
30
20
10
0
2000
2002
2004
2006
2008
2010
2012
2014
ASCII only
W Europe
UTF-8
JIS
others

Unicode (UTF-8)

- Industry standard for encoding and representing text
- Over 128,000 characters from 130+ scripts and symbol sets
- Backward compatible with ASCII
- Dominant character encoding for the Web

UTF-8 is the default character encoding in Python 3. All Python code is in UTF-8 and, ideally, all your data should be as well. It's when things aren't in UTF-8 that you run into trouble.

Take Home Concepts

Take Home Concepts

- Use Python environments
- Work in UTF-8