Data Science Foundations

Master in Big Data Solutions 2019-2020

Víctor Pajuelo

victor.pajuelo@bts.tech

Today’s class

Contents

2. Loading and processing images and text

- Image loading, pre-processing and filtering
- Image pre-processing for object detection and segmentation
- Text pre-processing, normalization, stemming, stopword removal
- Converting text to vectors and computing text similarity

Today’s Objective

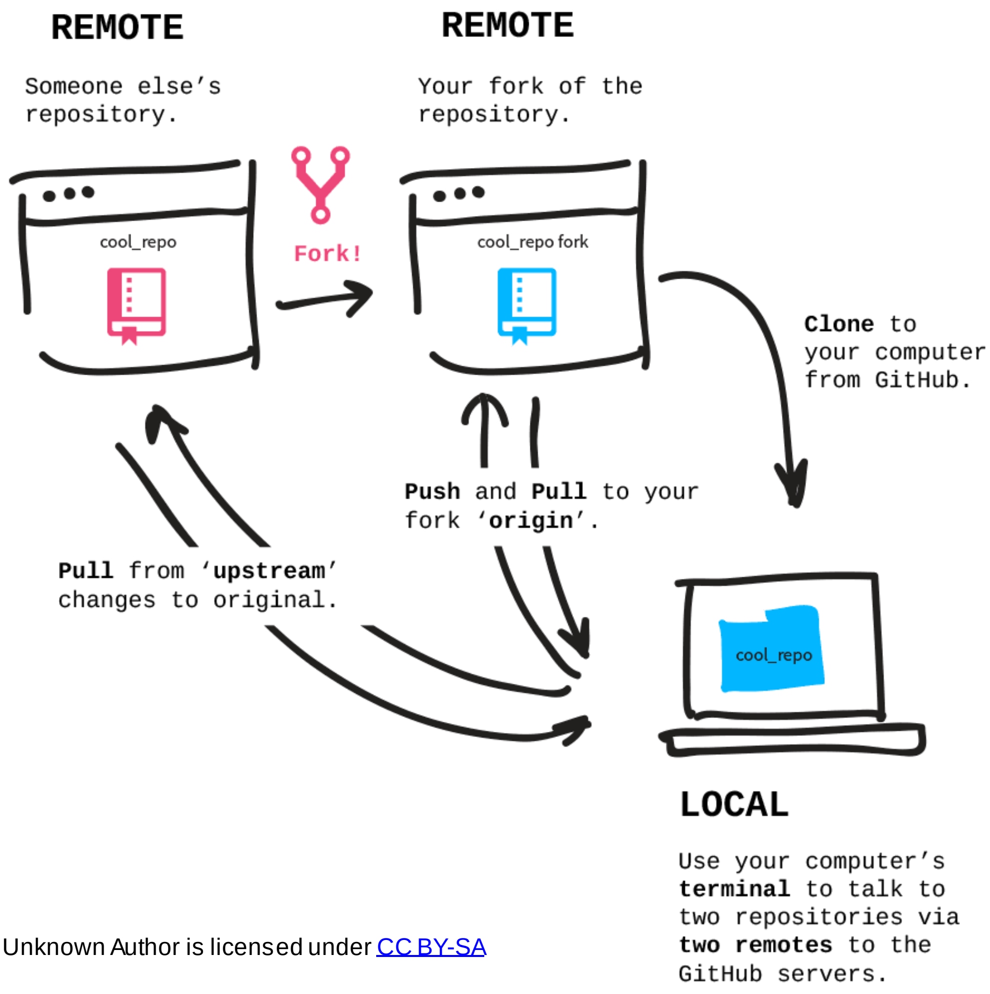
- Learn to process text in spaCy
- Starting to get used with text processing for analysis

- Why is this useful for a digital project?
  - Sentiment analysis
  - Text analytics
  - Recommender systems

Let's git things done!

Or, in case that you preferred a fork…

- https://help.github.com/en/articles/syncing-a-fork



REMOTE

Someone else's repository.

REMOTE

Your fork of the repository.

Clone to your computer from GitHub.

Push and **Pull** to your fork '**origin**'.

Pull from '**upstream**' changes to original.

LOCAL

Use your computer's **terminal** to talk to two repositories via **two remotes** to the GitHub servers.

This Photo by Unknown Author is licensed under CC BY-SA

Let’s see it again

```
$ git clone https://github.com/vfp1/bts-mbds-data-science-foundations-2019.git
# Some time passes...
$ git fetch upstream # There are changes!
$ git pull # Pull the changes
$ git checkout master  # Just in case
$ git merge --ff-only origin/master  # If in error, you probably made some commits to master
```

About the IoT visit

Our visit to IoT

- What are we going to do there?

Text mining and NLP

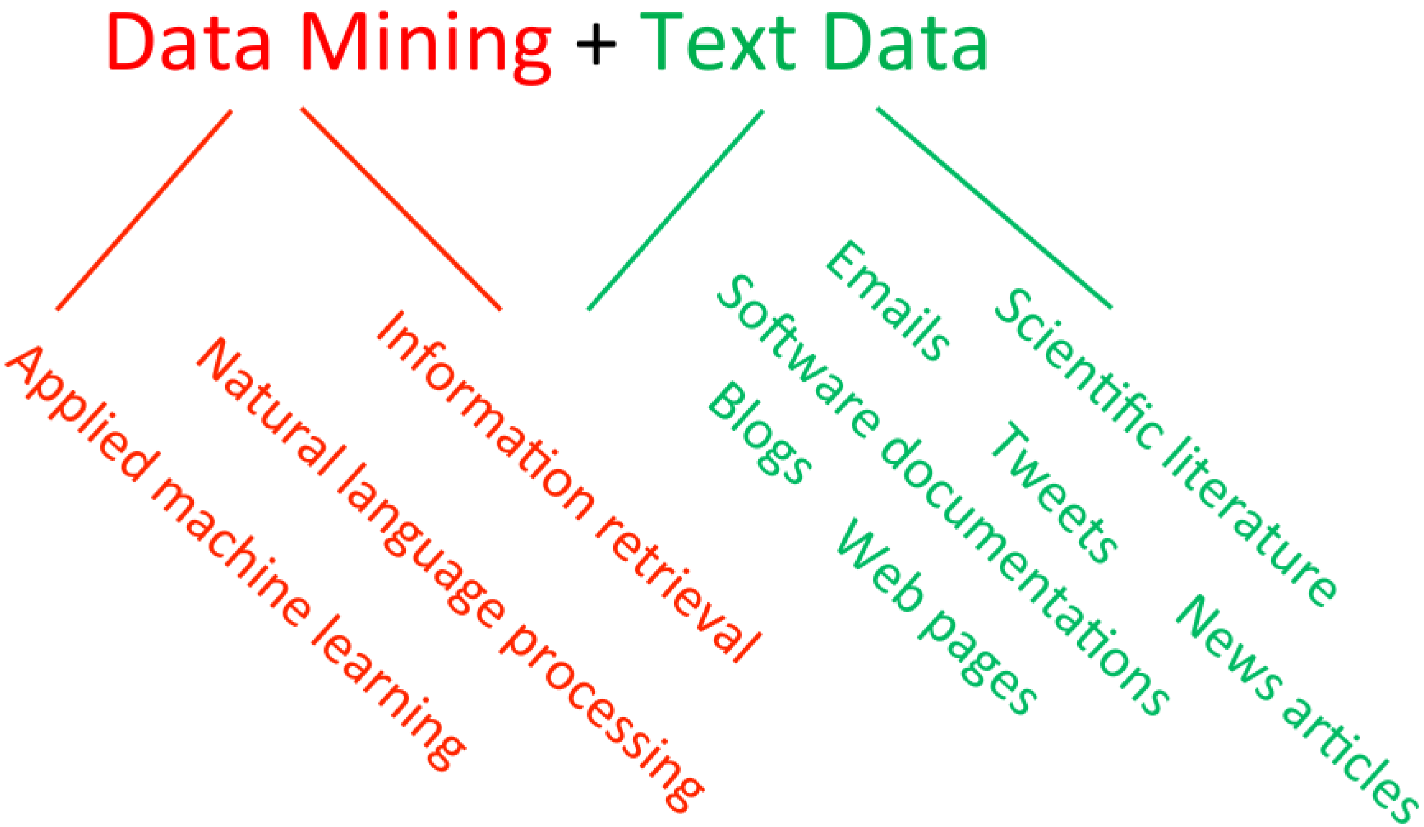
Introduction

- **Text mining**: *"the process of deriving high-quality information from text. […] The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods."*
- **Natural Language Processing**: *"a subfield of computer science, information engineering, and artificial intelligence concerned with […] how to program computers to process and analyze large amounts of natural language data."*

Applications

- Sentiment analysis (social networks, marketing)
- Document summarization
- Product recommendation (movies, books, anything that has reviews)
- Digital healthcare
- Many more!

Text mining visually



Data Mining + Text Data

Applied machine learning

Natural language processing

Information retrieval

Emails

Software documentations

Blogs

Scientific literature

Tweets

Web pages

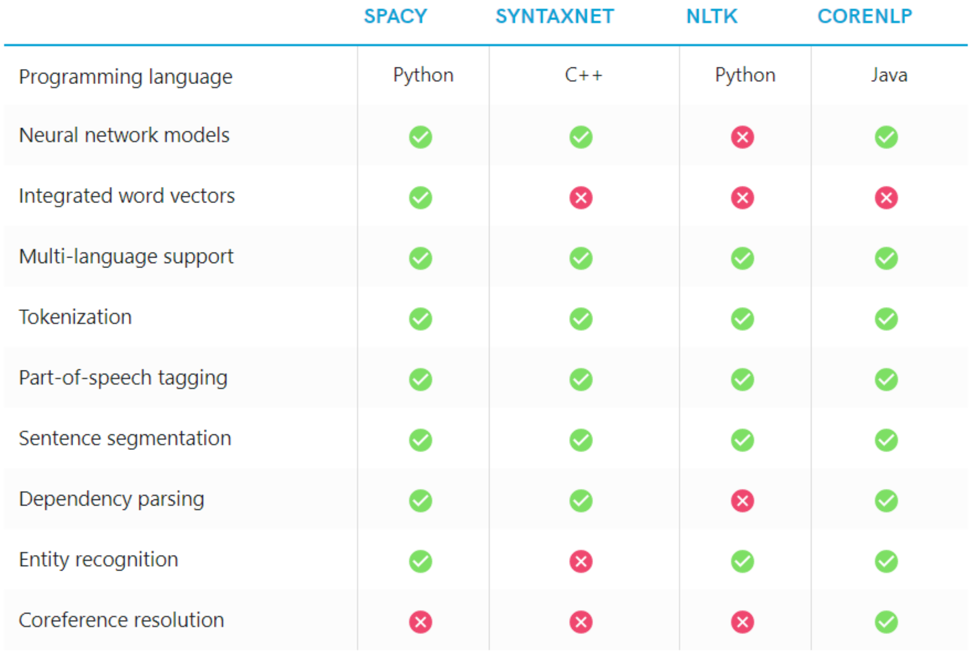
News articles

Challenges in text mining

- Mostly unstructured data, semi-structured at best
- Natural language contains lots of ambiguities on many levels
  - How to detect irony or sarcasm? How to know whether something is funny or not? What about slang?
- Annotated data depends on context and is hard to find

Python libraries

- **NLTK** (Natural Language ToolKit), the oldest one
- **TextBlob**, an easy to use library based on NLTK
- **spaCy**, a more modern alternative oriented towards deep learning



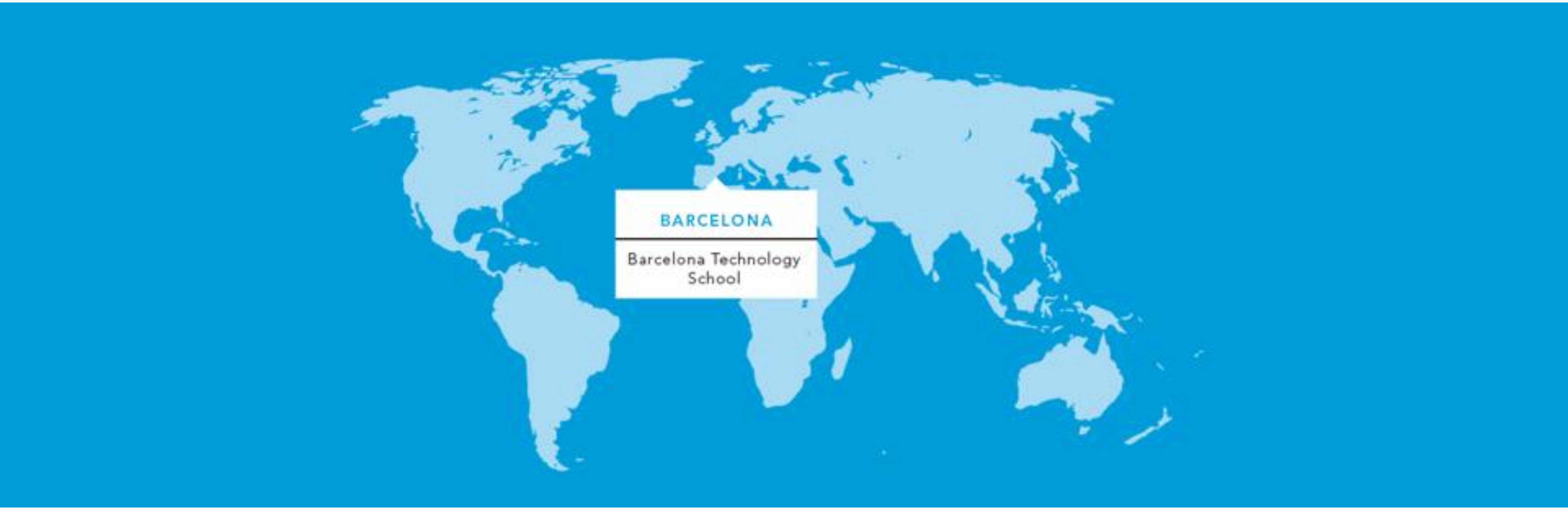
	SPACY	SYNTAXNET	NLTK	CORENLP
Programming language	Python	C++	Python	Java
Neural network models	✓	✓	✗	✓
Integrated word vectors	✓	✗	✗	✗
Multi-language support	✓	✓	✓	✓
Tokenization	✓	✓	✓	✓
Part-of-speech tagging	✓	✓	✓	✓
Sentence segmentation	✓	✓	✓	✓
Dependency parsing	✓	✓	✗	✓
Entity recognition	✓	✗	✓	✓
Coreference resolution	✗	✗	✗	✓

Installation

```
$ conda activate bts36
$ conda install -c conda-forge spacy
```

Let’s get to code!

Go to the notebook


BARCELONA
Barcelona Technology School