

NOVA

IMS

Information
Management
School

Data Mining I

Practical Class #1

João Fonseca - jpfonseca@novaims.unl.pt

Farina Pontejos - fpontejos@novaims.unl.pt

About us

Farina Pontejos

Academic background:

- Graduation in Mining Engineering (2011)
- MSc in Advanced Analytics (writing thesis – NOVA IMS)

Professional Experience:

- Mining Engineer
- Web Developer
- IT Consulting (Systems Analysis, ERP implementation)

About us

João Fonseca

Academic background:

- Graduation in Economics (2016 – NOVA SBE)
- MSc in Management (2019 – NOVA SBE)
- MSc in Information Management (2019 – NOVA IMS)
- PhD in Information Management (Since 2020 [Ongoing] – NOVA IMS)

Professional Experience:

- Research in Tourism Management
- Research in Remote Sensing, Natural Language Processing and Machine Learning methods (data augmentation, oversampling and active learning)

Resources

- Bibliography
- Data Mining I Github repo:
 - <https://github.com/joaopfonseca/Data-Mining-22-23>
 - Class slides and Jupyter Notebooks
- Google, Stack Overflow, documentations, Github and YouTube

Our working environment

- We will be using Anaconda: Currently one of the most popular Python distributions.
- Sets up a data science oriented working environment in Python
- It installs a set of libraries (for now, think of libraries as programming tools – like a toolbox in a woodshop)
- But it can be used for many different purposes (all it takes is installing the necessary libraries)

Anaconda

www.anaconda.org

- Anaconda is one of the most popular Python distributions for Data Science
- Comes with most of the main libraries for data manipulation
 - Pandas
 - Numpy
 - Matplotlib
 - Scipy
 - ...
- Easy to use and install



Virtual environments

<https://docs.conda.io/projects/conda/en/latest/user-guide/concepts/environments.html>

<https://docs.conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html>

- Isolated spaces that contain per-project dependencies (specific collection of installed conda packages)
- Using conda to manage environments:
 - Create, export, list, remove, and update environments
 - Switching or moving between environments (conda activate)
 - You can also share an environment file
- You can also use pip to manage environment

Python packages



Git and GitHub

<https://guides.github.com/activities/hello-world/>

<https://docs.github.com/en/github/getting-started-with-github>

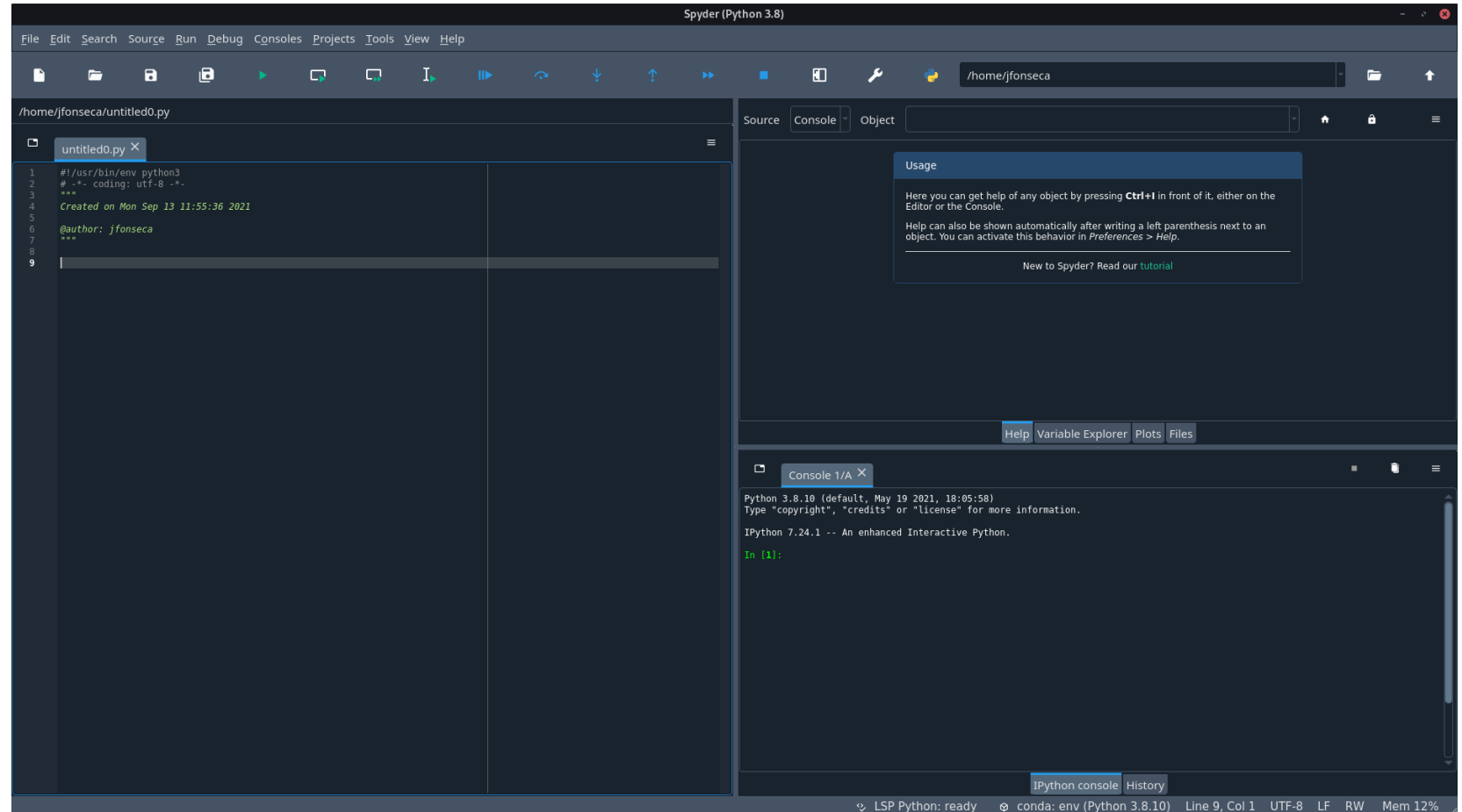
- What is GitHub?
 - Code hosting platform for version control and collaboration
- What is Git?
 - At the heart of GitHub is an open source **version control system** (VCS) called Git. Git is responsible for everything GitHub-related that happens locally on your computer.
- Why Git and GitHub?
 - You will need to use Git and GitHub for collaborating and version control in your projects. Also we have a GitHub repository with all the practical class contents:
 - <https://github.com/joaopfonseca/Data-Mining-22-23>

Main ways to access Python

- Python Shell and IPython
 - An interactive environment for writing and running code
- Jupyter Notebooks
 - A notebook that weaves code, data, prose, equations, analysis, and visualization
 - A tool for prototyping new code and analysis
 - A method for creating a reproducible workflow for scientific research
- IDE (Integrated Development Environment):
 - A software that helps you build code

Integrated Development Environment (IDE)

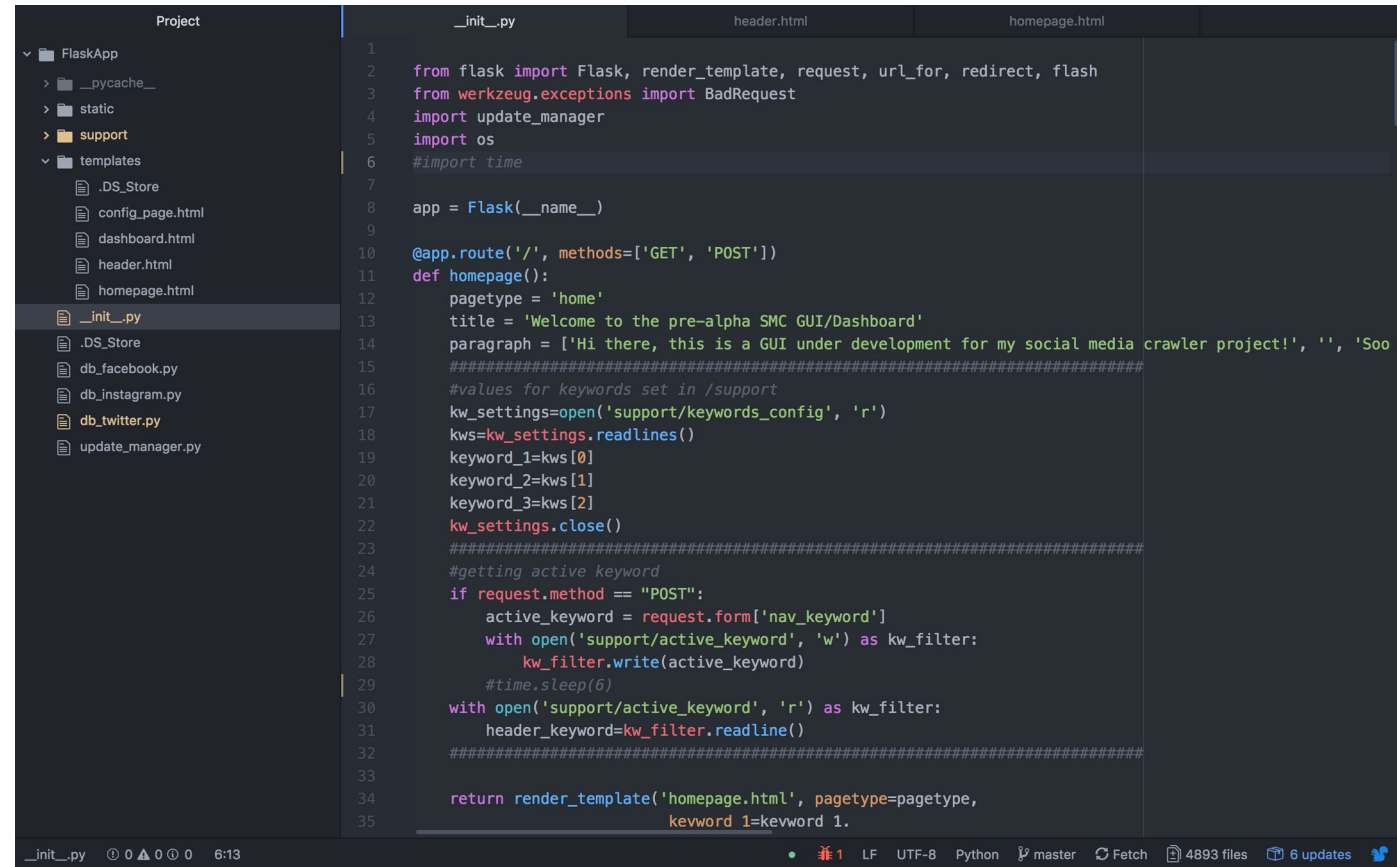
- Popular IDE's:
 - Spyder
 - PyCharm
 - VSCode
 - Rodeo
- Anaconda comes with Spyder and VSCode



Spyder

Text Editors

- Another method to write python scripts is using text editors
- Some popular text editors:
 - Vim (Linux terminal text editor)
 - Atom (popular open source editor)
 - Sublime Text (popular proprietary text editor)
 - Notepad ++ (Windows only)
- Usually highly customizable
- Usage of IDE and/or Text editor (and which ones to use) comes down to personal preference



```

1
2  from flask import Flask, render_template, request, url_for, redirect, flash
3  from werkzeug.exceptions import BadRequest
4  import update_manager
5  import os
6  #import time
7
8  app = Flask(__name__)
9
10 @app.route('/', methods=['GET', 'POST'])
11 def homepage():
12     pagetype = 'home'
13     title = 'Welcome to the pre-alpha SMC GUI/Dashboard'
14     paragraph = ['Hi there, this is a GUI under development for my social media crawler project!', ' ', 'Soo']
15     #####
16     #values for keywords set in /support
17     kw_settings=open('support/keywords_config', 'r')
18     kws=kw_settings.readlines()
19     keyword_1=kws[0]
20     keyword_2=kws[1]
21     keyword_3=kws[2]
22     kw_settings.close()
23     #####
24     #getting active keyword
25     if request.method == "POST":
26         active_keyword = request.form['nav_keyword']
27         with open('support/active_keyword', 'w') as kw_filter:
28             kw_filter.write(active_keyword)
29         #time.sleep(6)
30     with open('support/active_keyword', 'r') as kw_filter:
31         header_keyword=kw_filter.readline()
32     #####
33
34     return render_template('homepage.html', pagetype=pagetype,
35                           keyword_1=keyword_1.

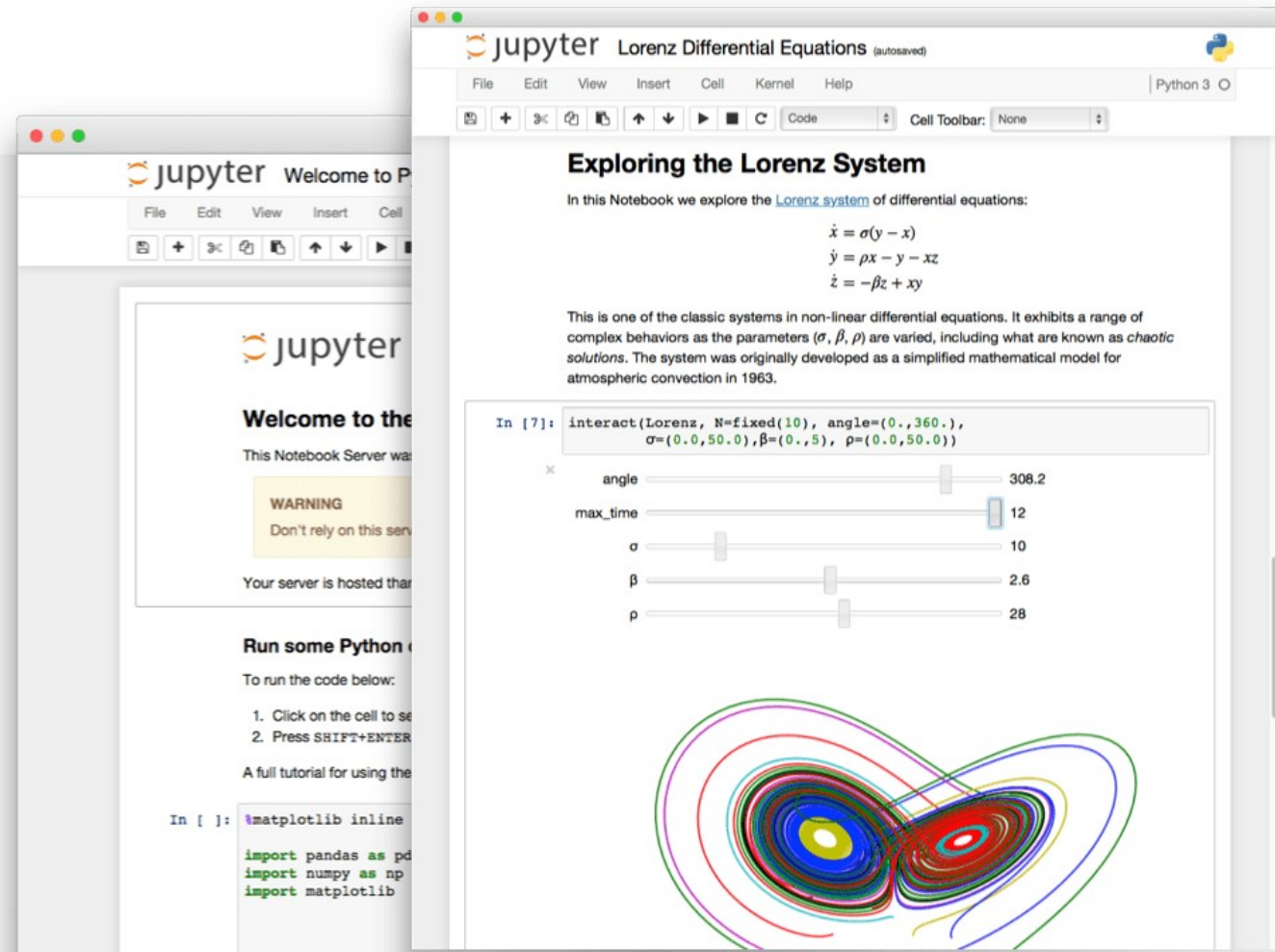
```

Atom Text Editor

The Jupyter Notebook

<http://jupyter.org/>

- Let's try it out!
 - Open your Anaconda Navigator
 - Start Jupyter Notebook



Data Mining Project

- Groups on Moodle (up to 3 students)
- Project guidelines on Moodle
- Anonymized data from a real-world **insurance company** (henceforth called "A2Z Insurance")
- Your goal is to develop a **Customer Segmentation** in such a way that it will be possible for the Marketing Department to better understand all the different Customers' Profiles.