Assignment lab 4 - Clustering trials and multidimensional features visualisation

Version: 21.11.2021

The purpose of this assignment

The goal of this lab is to check that the student has knowledge in the following topics:

- Dataset preprocessing.
- k-means clustering.
- Density-based spatial clustering of applications with noise (DBSCAN).
- Multidimensional features visualisation.

Business problem description

You can freely choose clustering related business problem and topic to be solved in assignment **from topics above**. Business problem does not have to be "new". You can use any example business problem found in Internet, excluding those what were already used in course examples.

Chosen business problem shall be not too complex, solvable with machine learning and development effort should be in range 6 .. 8 development hours.

Dataset description

You can use any data set available in Internet. You can also use your own data set if it does not contain any confidential and personally identifiable information.

Possible data set sources:

- Kaggle
- Opinion Mining, Sentiment Analysis, and Opinion Spam Detection
- Movie Review Data
- Kenya openData
- Estonian Open Government Data Portal
- Machine learning datasets
- Awesome Machine Learning for Cyber Security

Do not use Kaggle dataset Credit Card Dataset for Clustering which was used in previous years.

Task

1. Install Python Anaconda distribution (or Python with required modules) if it was not installed before.
2. Create software project in GitLab. Use one of to https://gitlab.cs.ttu.ee or https://gitlab.com. See class 1 material for details.
3. Print out python and available modules versions.
4. Read dataset file to pandas data frame. **See lab1 for CSV file handling**
5. Save dataset description to file in results directory. **See lab3 for guideline and implementation.**
6. Preprocess dataset if needed. **See lab1 and class 3 materials for details.**
7. Find possible suitable number of clusters with help of elbow method. **WCSS plot shall be saved to results folder for review. See lab1 how to save plot to file.**
8. Visualise dataset with help of t-SNE or PCA dimensions reduction to 2 dimensions. **See class 9 materials and examples for details.**
9. Select suitable clustering algorithm for your business problem and data set.
10. Find clusters. **See class 8 materials and examples for details.**
11. Visualise dataset with found clusters with help of t-SNE dimensions reduction by adding different colour and symbol to each cluster. **See class 9 materials and classes 8, 9 examples for details.**

Guidelines

Project repository structure and files

Project shall consist of following files (excluding directories `.git` and also `builds` if local gitlab-runner is used).

```
.
├── .gitignore
├── .gitlab-ci.yml
├── .pylintrc
├── common
|   ├── describe_data.py
|   └── test_env.py
├── data
|   └── .placeholder
├── lab4.py
├── Readme.markdown or Readme.md
└── results
    └── .placeholder
```

`lab4.py` shall be created by student.

For `.gitignore`, `.gitlab-ci.yml`, `pylintrc`, `data` and `common` files from lab4 template shall be used. Student shall add at least dataset file into data directory and lab4.py file into project root. Plotting functions available in classes examples can be added to common directory.

NB! Be aware that if you want to use different file names you need to modify CI configuration and tests accordingly.

Automation and GitLab CI stages

- Check-files
  - Tests existence of required files and fail if all files are not present.
  - List repository files excluding `.git` and `build` directories.
- Lint
  - Test `lab4.py` formatting with `pep8`.
  - Lint `lab4.py` with `pylint` by using configuration from file `.pylintrc`.
- Run-lab
  - Run `lab4.py`

Content of results directory is archived as build artefacts and can be downloaded.

Formatting and lint

autopep8 is used to test code formatting. autopep8 is supported by VS Code. For other editors it can be installed with conda:

```
$ pip install --upgrade autopep8
```

To run formatter from command line:

```
$ autopep8 --in-place lab4.py
```

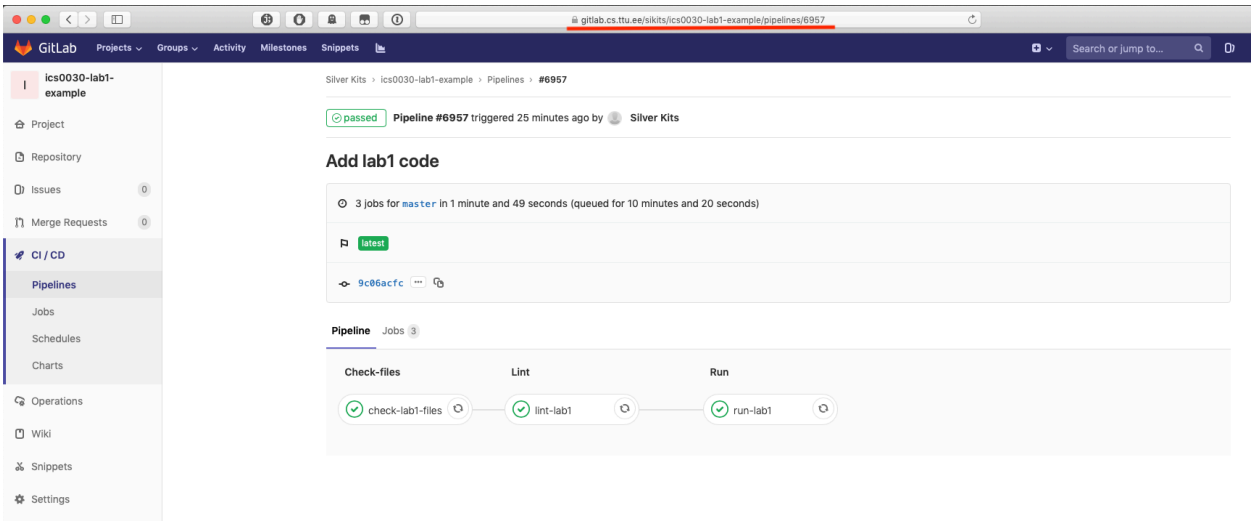
pylint is used for lint. Project template contains `.pylintrc`. Settings in this file are inline with VSCode default settings.

To run pylint from command line:

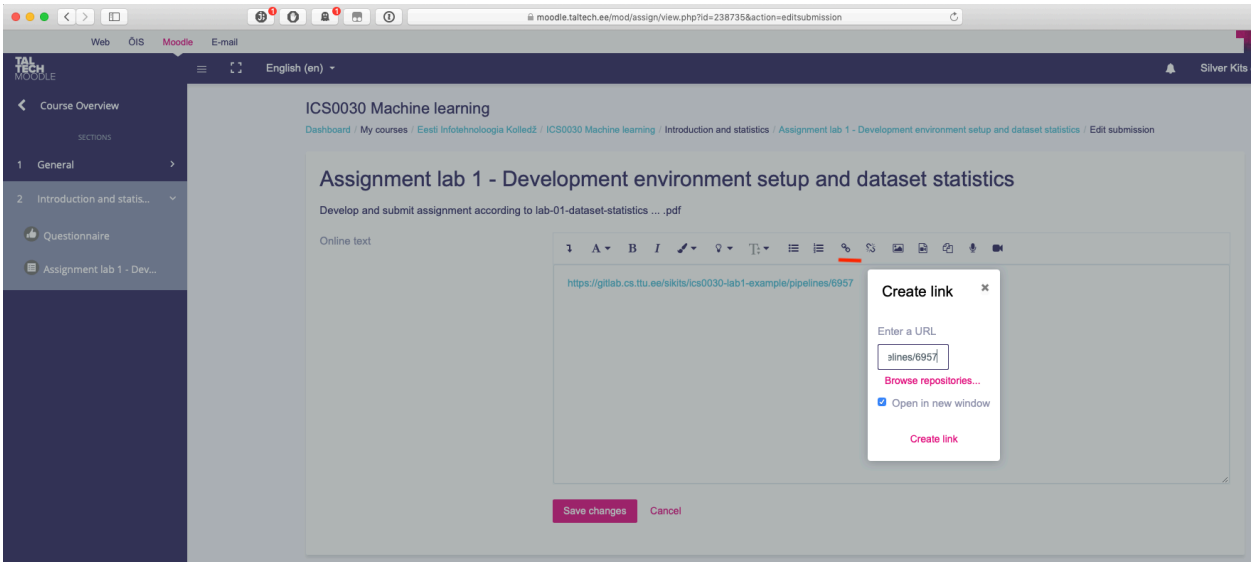
```
$ pylint lab4.py
```

Submission instructions

1. Be sure that your pipeline succeeds before submitting assignment in Moodle.



2. Submit link to the pipeline as an answer in Moodle. **Please make link HTML URL!**



3. Submit written description and analysis including following as an answer in Readme.markdown or Readme.md in your gitlab project:

    - Describe your business problem - What do you want to achieve?
    - Describe your approach - How you try to achieve possible solution to your problem?
    - Describe your results selection. - What "evidence" do you need to validate your solution?
    - Evaluate and describe your results. - Have you solved your problem? If yes then what is the solution. If no then why you most likely have not solved your problem.