Universität des Saarlandes
Prof. Dr. Dietrich Klakow
Lehrstuhl für Signalverarbeitung
NNIA Winter Term 2019/2020

Project Instructions (part 1.2)

Creating Git Repository for the Project

Deadline: 23.02.2021, 23:59

Step 1.2.1 - Get to know git

Study this tutorial to get to know how to work with git. Create a new repository for your project.

Step 1.2.2 - Create README.md

An essential part of each project is a README file which provides the description of the project and instructions on how to work with it. Create a README.md file of your project, here are some useful links:

- How to write a good readme
- Template
- Basic writing and formatting syntax
- Mastering markdown

At this point your README file should contain the following:

- Name of the project + short description
- Table of Contents
- General Information (update it during the course of implementing the project)
- Data Preprocessing (see Step 4 for further instructions on filling in this section)

If you want to add any other piece of information, feel free to do so.

Important Points will be subtracted from the final grade if there is no README.md (note the extension, it should be a markdown file) or it is incomprehensible.

Step 1.2.3 - Creating environment

We need to create an environment of our project that would contain all the dependencies and packages needed for the project.

If you are not familiar with the concept of environments, here is a great intro.

Here is a guide on how to create and manage environments using conda. If you prefer to use pip for creating environments, feel free to use pip. In any case, check if you have anaconda or pip installed by typing conda –version or pip –version into your terminal. Install one of those if it's not installed.

Your project should contain an environment.yaml file which specifies your environment. List all the packages you use for your project in this file, so that I could create environments for each individual project and not install the dependencies you use by hand each time. Update your environment.yaml file during the course of the project if needed.

The name of the environment: firstname1_firstname2_firstname3 (first names of each member).

Step 1.2.4 - Data preprocessing

For this project we will be working with pre-trained models provided by HuggingFace: Transformers. Carefully study the README, example code in examples folder, especially token-classification. Concentrate on the structure of the project.

- Create a file or directory with the code for preprocessing data, e.g. data_preprocess.py The file should contain main() method. For inspiration see the structure of the run_ner.py file.
- The arguments for the main() method—the input .conll file and the output directory—should be passed through the terminal.
- An example of how the command from terminal should look like you can find in the file run.sh You can create a similar .sh file and just run it or insert the contents in the terminal.
- Use argparse to parse the arguments passed through the terminal, or use any other method you prefer. You can have a look at the implementation in the run_ner.py file.
- In the README.md specify how to run data preprocessing (see the README of the token_classification example).

Project Contents

At this point your project should contain:

- README.md
- environment.yaml
- File or directory with your code for preprocessing. You may want to break the code into separate files (modules), in this case save all the files in a folder 'data'.

Submission instructions

The following instructions are mandatory. If you are not following them, tutors can decide to not correct your submission.

- Your submission for this assignment in Teams should contain only the link to your project and the names of the team members. Only one team member should make the submission.
- Don't include any files in your submission: I will create .tsv and .info files from the .conll file by running your code. Make sure you have all the instructions in README.
- Run the preprocessing on the sample data provided right now in Teams. The whole dataset will be available later.
- If you have any problems or questions, **use the channel 'Project Questions'** in Teams. Feel free to answer the questions of your fellow students.