University of St. Gallen

Institute of Computer Science

Fundamentals of Computer Science (3,125), HS2019
T. Huber, C. Niklaus, B. Bermeitinger, J. Sales, S. Handschuh
thomas.huber@unisg.ch

Assignment 7: Data Science: Advanced

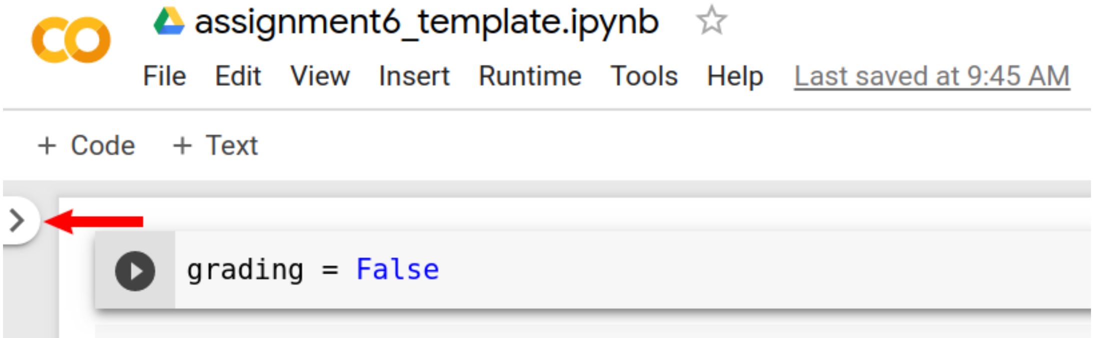
Deadline: Nov 26, 2019; 12:00 CET

Introduction In this assignment you will get firsthand experience with Data Science. You will work with a real dataset of 50.000 movie reviews. Your task is to build a Naive Bayes classifier that can predict whether a given review is positive or negative.
This is a practical implementation of an example from the lecture, but bigger.
After building your classifier you will learn how to evaluate how good or bad it is by calculating the well-known metrics Precision, Recall and F1-Measure.

Rationale It is always a good idea to get practical experience to reinforce the theoretical constructs you have learned in the lecture. Furthermore building a classifier without evaluating how good it is makes little sense. You will be able to apply the evaluation metrics in other areas as well.

Prerequisites Create a Google account if you do not already have. Then upload the notebook for this assignment on Google Colab: https://colab.research.google.com/.

- Click `File` in the top-right menu
- Click `Upload notebook...`
- In the menu select the template file from the .zip file that you downloaded from Canvas `assignment7_template.ipynb`
- Click the arrow on the right and upload the other files from the .zip file in the `Files` tab



```
grading = False
```

Figure 1: How to open the sidebar with the `Files` tab

For this assignment, you **ONLY** need to edit the notebook!
To run your code, click into one of the cells (the blocks of text that you can edit) and then click the *Run* button on the side of the cell.
When the code in a cell depends on a previous cell (for example one cell has a class *Villain*) then you need to run the previous cell before running the cell that depends on it! This also applies when you change code in a cell, you need to run that cell again and also any cells that depend on it (for example when they use a function defined in that cell).
You can run multiple cells by using the `Runtime` menu at the top.

1 Naive Bayes Classification and Evaluation (10 Points)

In this task you will learn how to build a simple classifier and how to evaluate it. Detailed instructions can be found in the notebook.

(a) Implement the `calculate_priors` method.

(b) Implement the `calculate_class_term_frequency` method.

(c) Implement the `calculate_class_term_probs` method.

(d) Implement the `classify` method.

(e) Implement the `evaluate` method.

Hand-in Instructions Create a file `README.txt` that includes 1.) answers to the questions, 2.) any reference that you used to complete this assignment, 3.) pitfalls you encountered and 4.) short explanations of your solution if necessary. Fill in your name and student ID in the comment header of files you edited.

Download your notebook by clicking `File` at the top and then select `Download .ipynb` and `Download .py`. Include both of these files in your `zip` file.

Compress the whole folder with the Python files and the `README.txt` to a `zip` file named "`assignment6.zip`". Upload the zip file to your exercise group's Course page on Canvas. See Figure.2 for the list of files that should be included in the zip file.

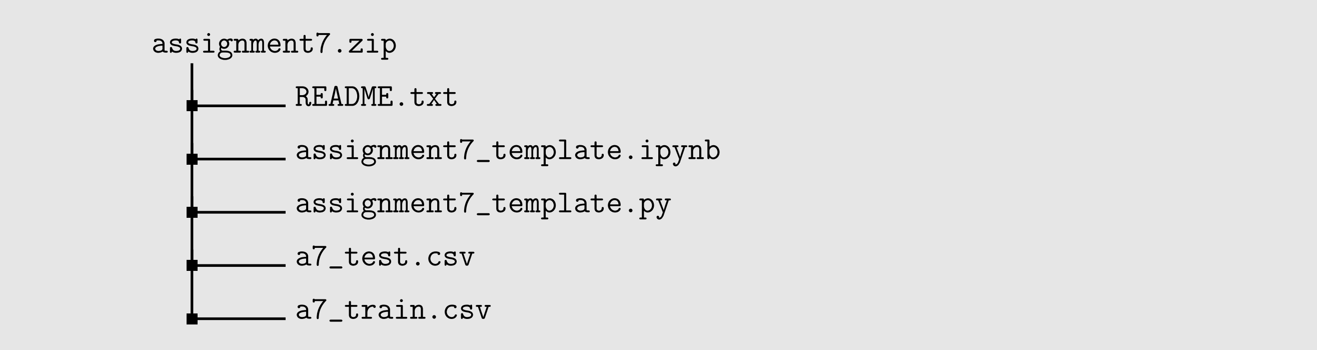


Figure 2: The required files for submission.