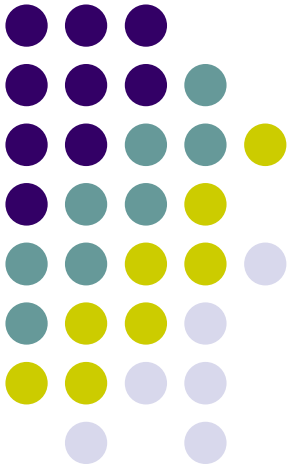
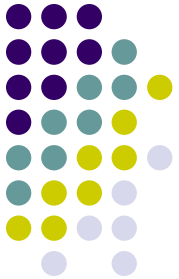
Big Data Analytics

Ninth Assignment



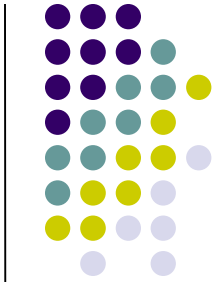
Deadline: 15 March 2021



First part

- Replicate Assignment 7 focusing on the «polite» variable as an outcome by applying both the Regularized Regression as well as the Gradient Boosting algorithms

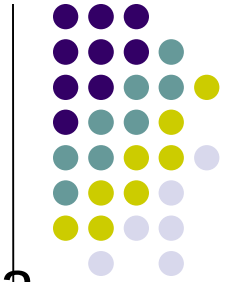
Deadline: 15 March 2021



Second part

- Form 3 groups: 1) Sofia, Federica and Anna; 2) Fabiana, Valentina, Riccardo; 3) Carola, Chunmin, Jacopo
- Run a search on Twitter as you like and download between 5,000 and 10,000 tweets in any language you want
- Define a set of categories for the tweets you have downloaded (it can be a 2-set categories such as positive/negative, or a 3-set categories such as positive/negative/neutral, or anything else you want)
- Discuss within the group the meaning of each class you have selected

Deadline: 15 March 2021



- Define a training-set (around 200 tweets if you have a 2-set categories; 300 if you have a 3-set categories, etc.) and a test-set. Each of the student in the group must codify the same tweets
- Check your inter-coder reliability
- Then run on the training-set both the Regularized Regression as well as the Gradient Boosting algorithms
- Cross-validated your results and pick up the best algorithm
- Then classify the test-set