Politechnika Wrocławska

Massive Data Processing

Laboratory 0

Piotr Bielak, Roman Bartusiak

October 4, 2020

HR EXCELLENCE IN RESEARCH

Politechnika Wrocławska

Overview

- Staff
- Materials
- AWS
- Grading
- Project
  - Goals
  - Plan
  - Extra
- Calendar

Staff

- Piotr Bielak
  `piotr.bielak@pwr.edu.pl`
- Roman Bartusiak
  `roman.bartusiak@pwr.edu.pl`

Exact office hours will be announced, but you can find us in **room 441, building A-1**. Please send an email beforehand.

Materials

- https://lsdp.ml
- http://docs.python.org

Politechnika Wrocławska

AWS

- 200$ per student
- Possibility to get more in special cases
- Use it reasonably
  - Remove unused resources
  - Look at the pricing
  - Try the spot instances

Grading

- Every part is graded separately
- Every part is not equal (different number of points)
- Every part must get $> 50\%$

Points range	Grade
$< 50\%$	2
$[50\%, 60\%)$	3
$[60\%, 70\%)$	3.5
$[70\%, 80\%)$	4
$[80\%, 90\%)$	4.5
$[90\%, 100\%]$	5
$> 100\%$	5.5

Grading

- 1 absence
- 90% of points for lists submitted on office hours/next week group
- 80% of points for lists submitted on next classes
- 0% of points for list after 2 weeks
- you can only delay one list (delaying any further list means not passing the course)

- all assignments are due to the Thursday at 22:00 before the next lab; in the meantime they will be checked and grades will be announced at the beginning of the next lab;
- plagiarism is not allowed and will be reported to department entities

Project

Goals

- Data acquisition
  - Monitoring
  - Use task queue
- Data transformation and unification
- Data cleaning, persistence
- Statistical analysis
- Machine learning
- Deployment

Project

Plan

Lab 0. Introduction, grading, plan

Lab 1. Bash, Docker, Python parallelization

Lab 2. Reddit posts scraping and process monitoring (Celery, InfluxDB, Prometheus, Grafana)

Lab 3. Post embedding, data persistency, statistics visualization (MongoDB, Redash)

Lab 4. Machine learning on huge data volumes using Spark

Lab 5. Application deployment (Kubernetes, Helm)

Lab 6. SPA Application (Flask, model serving)

Project

Extra

To get **grade 5.5** you must perform extra work:

1. LSH for top *k* subreddits
2. Subreddit similarity graph
3. Community detection
4. etc.

Calendar

Lab	TP		TN	
	Intro	Due	Intro	Due
0	05.10	–	05.10	–
1	05.10	**22.10**	05.10	**15.10**
2	26.10	**05.11**	19.10	**12.11**
3	09.11	**19.11**	16.11	**26.11**
4	23.11	**03.12**	30.11	**10.12**
5	07.12	**17.12**	14.12	**07.01**
6	21.12	**14.01**	11.01	**21.01**
Summary	1.02			

All due dates are set on Thursdays at 22:00 before the next lab

https://pwr.edu.pl/studenci/kalendarz-akademicki

Politechnika Wrocławska

Massive Data Processing

Laboratory 0

Piotr Bielak, Roman Bartusiak

October 4, 2020