Politechnika Wrocławska

Massive Data Processing

Laboratory 1

Piotr Bielak, Roman Bartusiak

September 27, 2020

HR EXCELLENCE IN RESEARCH

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

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

Politechnika Wrocławska

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 two lists (delaying any further list means not passing the course)

Politechnika Wrocławska

Project

Goals

1. Data acquisition
   - Monitoring
   - Use task queue
2. Data transformation and unification
3. Data cleaning, persistence
4. Statistical analysis
5. Machine learning
6. Deployment

Project

Plan

1. Reddit posts scraping and process monitoring
2. Post embedding, data persistency
3. Statistics visualization
4. Linear regression (number of upvotes)
5. Classification (subreddit)
6. SPA Application

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

Part due	TP	TN
Intro	12.10	05.10
1.	26.10	19.10
2.	09.11	16.11
3.	23.11	30.11
4.	07.12	14.12
5.	21.12	11.01
6.	18.01	25.11
Summary / grades	–	1.02

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

Massive Data Processing

Laboratory 1

Piotr Bielak, Roman Bartusiak

September 27, 2020