Politechnika
Wrocławska

Massive Data Processing

Laboratory 1

Roman Bartusiak, Piotr Bielak

October 2, 2019

HR EXCELLENCE IN RESEARCH

Overview

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

Staff

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

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

o

Project

Goals

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

Project

Plan

1. Reddit posts scraping and process monitoring
2. Post embedding, data persistency
3. Statistics visualization
4. Linear regression
5. Classification (subreddit, NSFW)
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

Calendar

Part	TP	TN
1.	17.10	10.10
2.	14.11	24.10
3.	28.11	07.11
4.	12.12	21.11
5.	09.01	05.12
6.	16.01	19.12
Final	30.01	23.01

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

Politechnika Wrocławska

Massive Data Processing

Laboratory 1

Roman Bartusiak, Piotr Bielak

October 2, 2019