Mining Massive Datasets
Administrative Issues

WS 2017 / 2018

Artur Andrzejak

Contact

- Lecturer: Prof. Dr. Artur Andrzejak
  - Group: Parallele und Verteilte Systeme (PVS)
  - Email: artur.andrzejak@informatik.uni-heidelberg.de
  - Web: http://pvs.ifi.uni-heidelberg.de/ oder goo.gl/c50qD
  - In INF 205 (Theoretikum), room 2/214
- Assistants
  - Diego Elias Costa(exercises)
    - Email: diego.costa@informatik.uni-heidelberg.de
  - Kevin Kiefer: (tutorial)
    - Email: sl416@ix.urz.uni-heidelberg.de

Time and Location

- Lecture
  - Monday, 16:15 until about 18:00 CE(S)T
  - INF 306, Seminar Room 14

- Tutorial
  - Wednesday, 16:15 until about 18:00 CE(S)T
  - INF 205, SR A

Moodle

- Please register!
  - Contains all slides and exercises
  - Mailing lists
  - Used for submitting solutions and grading

- Access info
  - https://elearning2.uni-heidelberg.de/course/view.php?id=15577
  - ID: 15577 or „IMMD2017“
  - Registration key: **mmd2017modle**

Müsli

- Please register!
  - Your data is used to enter grades to LSF
  - We also need this data to justify the number of tutors

- Access info
  - https://muesli.mathi.uni-heidelberg.de/lecture/view/752
  - Lecture: **Mining Massive Datasets**
  - To login, use your own password for Muesli

Slides

- Slides are uploaded shortly after the lecture to Moodle
  - Why *after*? Because answers to questions posed in a lecture are in the slides
- Cover 95% of material for the final exam (Klausur)
- Hints about mistakes, suggestions for improvements etc. are very welcome!
- Also the introduction slides will be uploaded to Moodle

Weekly Exercises

- Groups of maximum 3 persons are allowed
  - Please form your group in the first two weeks
- Admission criteria for the final exam
  - At least 50% of points from all weekly exercises (per group)
- First tutorial already this Wed (on 18.10.2017)
  - Used to create groups and repeat lecture
  - A VM for exercises will be distributed on a USB stick

Weekly Exercises /2

- Will be issued by Tuesday evening / night
- Submission of the solutions until Monday of the following week at 23:59 CE(S)T
  - E.g. issued on 24.10., submission until 30.10. at 23:59
  - Submit via Moodle
  - In addition, you can print out and give it to Kevin Kiefer
- Contents of the tutorials
  - Discussions on the next problem sets
  - Discussion of the solutions

Final Exam (Klausur)

- Date: 5. February 2018 (Monday)
  - Last week of semester
- Time: 16:00 - 18:00 CET (same as lecture)
- Location
  - Default is the lecture room (INF 306, room 14)
  - Since this will not be sufficient, we will request an additional one
- No books, scripts, computer, smartphone etc. („Es sind keine Hilfsmittel zugelassen“)
- Please bring your photo ID („Bitte Personalausweis / Pass mitbringen“)

Statistics on Participants

- Bachelor Informatik: 8
- Master Informatik: 24
- Physics: 9
- Computer linguistics: 2
- Geography: 1
- Bio-* sciences:
- Scientific Computing: 14
- Others:

- Total:

Books

- Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, **Mining of Massive Datasets**, Cambridge University Press, Version 2.1 von 2014 (online)
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**, Springer, 2009 (online)
- Ron Bekkerman, Misha Bilenko, John Langford, **Scaling Up Machine Learning**, Cambridge University Press, 2012
- Jiawei Han, Micheline Kamber, Jian Pei, **Data Mining: Concepts and Techniques**, Morgan Kaufmann, (third edition), 2012
- Bücher aus dem **O'Reilly Data Science Starter Kit**, 2014, http://shop.oreilly.com/category/get/data-science-kit.do

Software (selection)

- Apache Spark
  - https://spark.apache.org/
  - A virtual machine (linux) with Spark, Python, Java, and Intellij IDEA (IDE) is on the USB-stick
- Related software
  - GraphX, MLlib, Spark Streaming, Spark SQL (Shark)
- Apache Hadoop
  - http://hadoop.apache.org/
  - Hadoop Distributed File System (HDFS)
  - Hadoop MapReduce
- Related to Hadoop
  - Pig, Mahout, Hive