**Re-training CA academic staff on IBDA by EU experts in CA partner HEIs - TAJIKISTAN – MAY/JUNE 2022**

V1 / WP2 / Task 2.4

Document Title

May 25, 2022



Establishment of training and research centers and courses development on intelligent big data analysis in Central Asia-ELBA

**610170-EPP-1-2019-1-ES-EPPKA2-CBHE-JP**

Table of Contents

[Table of Contents 1](#_Toc104376462)

[Logistics 2](#_Toc104376463)

[Wireless Network 2](#_Toc104376464)

[Zoom Links 2](#_Toc104376465)

[Installation instructions 3](#_Toc104376466)

[Professors 3](#_Toc104376467)

[Programme 4](#_Toc104376468)

# Logistics

To be conducted at Technological University of Tajikistan.

From May 30th, 2022 to June 3rd, 2022

**Location:**

Technological University of Tajikistan, 63/3 N. Karabaev Str, Dushanbe

**Rooms & Hours:**

Auditorium 1/206 (Building #1, 2nd floor)

# Wireless Network

**To be announced**

# Zoom Links

For those ELBA participants who cannot join the retraining event, we have prepared the following Zoom connection. All CA partners will be able to join the workshops and sessions that are scheduled for 30th May – 3 June, 2022 at the TUT:

TUT is inviting you to a scheduled Zoom meeting.

Topic: ELBA – Retraining seminars, IBDA

Time: May 30, 2022 09:00 Dushanbe

 Every day, until Jun 3, 2022, 5 occurrence(s)

 May 30, 2022 09:00

 May 31, 2022 09:00

 Jun 1, 2022 09:00

 Jun 2, 2022 09:00

 Jun 3, 2022 09:00

Join Zoom Meeting:

<https://upr-si.zoom.us/j/81442620864?pwd=YURGOTdWYWZidS9DZUVQdzRZTTdPUT09>

Meeting ID: 814 4262 0864

Passcode: 123456

# Installation instructions

Install **VirtualBox** in your laptops (<https://www.virtualbox.org/>). Ideally, you should have at least 6Gb RAM (10GB recommended for Windows 10/11 settings) and 20GB space in disk.

Install also VirtualBox Oracle VM **VirtualBox Extension Pack** (which supports USB devices and other services): <https://download.virtualbox.org/virtualbox/6.1.34/Oracle_VM_VirtualBox_Extension_Pack-6.1.34.vbox-extpack>

Install a vanilla Ubuntu installation from the official server: <https://ubuntu.com/tutorials/install-ubuntu-desktop#1-overview>

Users using any flavor of Linux should be OK with no prerequisite additional installations.

Install the **WEKA Data Mining Toolbox** on your machine. You may find the installation instructions here:
<https://waikato.github.io/weka-wiki/downloading_weka/#developer-version>
(we will be using the Developer Version of WEKA).

# Professors

The retraining organized by the University of Primorska will be run by two professors:

**Assist. prof. Branko Kavšek, PhD**. Assistant Professor in Computer Science at The Faculty of Mathematics, Natural Sciences and Information Technologies (UP FAMNIT) – University of Primorska in Koper and Researcher at the Artificial Intelligence Laboratory (AI LAB) – “Jožef Stefan” Institute in Ljubljana.

His research includes machine learning, data mining, association rules, graphs, time-series and data analysis in general. He closely collaborates with researchers from the Artificial Intelligence Laboratory and the Department of Information Sciences and Technologies.

He teaches courses Formal Languages and Computability, Introduction to Machine Learning and Data Mining, and Intelligent Systems to both undergraduate and master’s students at the university where he is the coordinator of the undergraduate study program Computer Science and vice head of the Department of Information Sciences and Technologies.

**Assoc. prof. Jernej Vičič, PhD**. Associate Professor in Computer Science at The Faculty of Mathematics, Natural Sciences and Information Technologies (UP FAMNIT) – University of Primorska, and Senior Researcher at the Research Centre of the Slovenian Academy of Sciences and Arts. He received PhD in Computer Science at University of Ljubljana, Faculty of Computer Science and Informatics with thesis "A Fast Implementation of Rules Based Machine Translation Systems for Similar Natural Languages" in August 2012.

He is the head of the Blockchain and Language Technologies Lab at UP FAMNIT.

His major research interests involve language technologies in general, specifically he focuses on the field of Natural Language Translation and Distributed Ledger Technologies.

Among his major achievements in the presented fields (apart from his Phd thesis) is the machine translation system for the related languages: Slovenian – Croatian based on the open-source Apertium translation system toolkit. Jernej is one of the developers of the Apertium toolkit.

# Programme

**DAY 1: May 30th, 2022**

**The WEKA Workbench, finding and understanding the data**

Morning session #1 (2 hours = 2 x 45 minutes):

 Introducing the WEKA DM Workbench

 Checking the installations and running the workbench

 How a typical DM pipeline looks in WEKA?

Short break (30 minutes)

Morning session #2 (2 hours = 2 x 45 minutes):

 Finding appropriate data for analysis (well-known DM Repositories)

 Importing and transforming the data

 Visualizing the data

Lunch break (1 hour and 30 minutes)

**Preparing the data for the analysis**

Afternoon session #1 (2 hours = 2 x 45 minutes):

 Types of data and type transformations

 Missing values and the date format

 Sampling and imputation

Short break (30 minutes)

Afternoon session #2 (2 hours = 2 x 45 minutes):

 Data preparation in action – lab (“cleaning” sample datasets)

**DAY 2: May 31st, 2022**

**Big Text Data I**

Morning session #1 (2 hours = 2 x 45 minutes):

 Introduction to Corpora

 Corpora usage

Short break (30 minutes)

Morning session #2 (2 hours = 2 x 45 minutes):

 Corpora construction and tagging

 Corpora creation using SketchEngine

Lunch break (1 hour and 30 minutes)

**Big Text Data II**

Afternoon session #1 (2 hours = 2 x 45 minutes):

 Terminology (definition, usage)

 Terminology extraction

Short break (30 minutes)

Afternoon session #2 (2 hours = 2 x 45 minutes):

 To be defined soon

**DAY 3: June 1st, 2022**

Morning session #1 (2 hours = 2 x 45 minutes):

 Presentation of the ACM Curriculum guidelines for Data Science Programs

Short break (30 minutes)

Morning session #2 (2 hours = 2 x 45 minutes):

 **Revision and Evaluation of Intelligent Big Data Analytics Academic Programmes in Central Asia Institutions**

Lunch break (1 hour and 30 minutes)

Afternoon session #1 (2 hours = 2 x 45 minutes):

 **Meeting with industrial and academic stakeholders to discuss ELBA’s project**

Short break (30 minutes)

Afternoon session #2 (2 hours = 2 x 45 minutes):

 **IBDA centers infrastructures**: Visit to local IBDA centers and discussion on plans to further improve IBDA-based capabilities of the CA partners (computational infrastructures, technologies to support research and industrial projects, etc.).

**DAY 4: June 2nd, 2022**

**Supervised machine learning algorithms I**

Morning session #1 (2 hours = 2 x 45 minutes):

 The majority class classifier

 One attribute does it all – the OneR classifier

 Simple decision trees – the ID3 algorithm

Short break (30 minutes)

Morning session #2 (2 hours = 2 x 45 minutes):

 Decision trees for real-life data – the C4.5 algorithm

* Handling numeric attributes
* Handling missing data
* Handling noise = pruning the trees

Lunch break (1 hour and 30 minutes)

**Supervised machine learning algorithms II**

Afternoon session #1 (2 hours = 2 x 45 minutes):

 Decision rules

 Ensemble classifiers

Short break (30 minutes)

Afternoon session #2 (2 hours = 2 x 45 minutes):

 Supervised machine learning algorithms in action – lab
(classifying sample datasets)

**DAY 5: June 3rd, 2022**

**Unsupervised machine learning algorithms**

Morning session #1 (2 hours = 2 x 45 minutes):

 Clustering

 Frequent pattern analysis

 Nearest neighbors

Short break (30 minutes)

Morning session #2 (2 hours = 2 x 45 minutes):

 Unsupervised machine learning algorithms in action – lab
(applying selected unsupervised ML algorithms on sample datasets)

Lunch break (1 hour and 30 minutes)

**Evaluation of the ML models**

Afternoon session #1 (2 hours = 2 x 45 minutes):

 Training, Validation and Test sets

 The (random) split method

 The cross-validation method

 Evaluation measures

 The unbalanced class problem

Short break (30 minutes)

Afternoon session #2 (2 hours = 2 x 45 minutes):

 Statistical significance of experimental results

 Using the WEKA Experimenter tool