STUDIJŲ DALYKO (MODULIO)
APRAŠAS

Dalyko (modulio) pavadinimas	Kodas
Didžiųjų duomenų programiniai įrankiai	

Dėstytojas (-ai)	Padalinys (-iai)
Koordinuojantis: dr. Tadas Danielius **Kitas (-i):**	Statistinės analizės katedra

Studijų pakopa	Dalyko lygmuo	Dalyko (modulio) tipas
Antroji	Pradedančiųjų	Privalomasis

Įgyvendinimo forma	Vykdymo laikotarpis	Vykdymo kalba (-os)
Auditorinė	Penktas (rudens) semestras	Lietuvių

Reikalavimai studijuojančiajam	
Išankstiniai reikalavimai: Pitono pagrindai; anglų kalba B1 (arba aukštesniu) lygiu pagal įprastą Europoje galiojančią lygių sistemą.	**Gretutiniai reikalavimai (jei yra):**

Dalyko (modulio) apimtis kreditais	Visas studento darbo krūvis	Kontaktinio darbo valandos	Savarankiško darbo valandos
5	125	48	77

Dalyko (modulio) tikslas: studijų programos ugdomos kompetencijos

Studentai turėtų išugdyti šias kompetencijas (B – bendrosios, D – dalykinės kompetencijos):

- gebėjimas analizuoti, sisteminti, mokytis ir taikyti įgytas žinias praktikoje (B1);
- gebėjimas rinkti, valdyti ir tvarkyti duomenis (D5);
- gebėjimas rinktis tinkamą analizės metodologiją bei jai reikalingus įrankius (D6);
- gebėjimas interpretuoti ir reprezentuoti analizės rezultatus (D7).

Dalyko (modulio) studijų siekiniai (išklausę modulį studentai turėtų):	Studijų metodai	Vertinimo metodai
• žinoti specifines problemas, kurios iškyla analizuojant didžiuosius duomenis; • gebėti naudotis keliais programiniais įrankiais, skirtais didžiųjų duomenų analizei; • gebėtis skaityti tikslinę literatūrą bei taikyti aprašytus metodus praktiškai.	Paskaitos, pratybos naudojant programinę įrangą, savarankiškas užduočių sprendimas ir teorinės medžiagos studijavimas.	Kontroliniai darbai, savarankiškos užduotys

Temos	Kontaktinio darbo valandos							Savarankiškų studijų laikas ir užduotys	
	Paskaitos	Konsultacijos	Seminarai	Pratybos	Laboratoriniai darbai	Praktika	**Visas kontaktinis darbas**	**Savarankiškas darbas**	**Užduotys**

1. Įvadas. Pagrindines sąvokos ir problemos, susijusias su dideliais duomenimis, taip pat tipinės užduotys ir apdorojimo modeliai. Trumpa duomenų valdymo ir integravimo metodų apžvalga, aptariami efektyvūs duomenų saugojimo, paieškos ir integravimo būdai. Apžvelgiami įvairūs programinės įrangos įrankiai, dažniausiai naudojami didelių duomenų analizėje.	2						**2**	**5**	Perskaityti [2] knygos 1 skyrių. Savarankiškai surasti ir perskaityti įvadinį straipsnį apie didelių duomenų masyvų analizę.
2. Numpy modulis. Pagrindinis objektas ir operacijos. Vaidmuo python ekosistemoje didžiųjų duomenų analizės kontekste.	2				4		**6**	48	Reguliariai spręsti dėstytojo pateiktas užduotis, skirtas reikalingų įgūdžių lavinimui. Kadangi užduotys priklauso nuo naudojamos programinės versijos įrangos, o pati įranga nuolat tobulinama ir versijuojama dažnai, užduotys pateikiamos semestro metu kuomet skaitomas kursas.
3. Python moduliai lygiagretiems skaičiavimams. multiprocessing, mpi4py ir ipyparallel apžvalga, pagrindiniai funkcionavimo principai. Kelios alternatyvos.	4				4		**8**		
4. Apache Spark. pyspark.sql duomenų struktūros. pyspark.mllib vartotojo sąsaja pavyzdžiais (paprasčiausi modeliai, nereikalaujantys gilių statistikos žinių).	4				4		**8**		
5. Paskirstyti skaičiavimai kompiuterių telkiniuose. linux komandinės eilutės pradmenys. Užduočių vykdymas SLURM aplinkoje.	4				4		**8**		
6. Kontroliniai darbai, individuali užduotis ir egzaminas.					16		**16**	**24**	Pasiruošti atsiskaitymams.
Iš viso							**48**	**77**	

Vertinimo strategija	**Svoris proc.**	**Atsiskaitymo laikas**	**Vertinimo kriterijai**
1 kontrolinis darbas	20	ketvirta studijų savaitė	Kontrolinį sudaro ne daugiau 5 užduočių, skirtų patikrinti žinių lygiui. Bendra užduočių vertė – 2 balai. Kiekvienos užduoties vertė svyruoja nuo 0,1 iki 1 balo. Užduotys atliekamos raštu arba prie kompiuterio. Atskiros užduoties vertinimo principai: a) išskiriamos dalys, už kurias skiriama dalis visos užduoties taškų; b) atlikus atitinkamą dalį be klaidų už ją skiriamas maksimalus taškų skaičius, priešingu atveju taškų skaičius mažinamas atsižvelgiant į padarytas klaidas; c) klaidingas kažkurios dalies atlikimas neturi įtakos kitų dalių vertinimui.
2 kontrolinis darbas	20	aštunta studijų savaitė	
3 kontrolinis darbas	20	dvylikta studijų savaitė	
4 kontrolinis darbas	20	šešiolikta studijų savaitė	
5 Individuali užduotis	20	Ne vėliau kaip savaitė iki egzamino	Užduoties vertinimo principai tokie patys kaip ir kontrolinių darbų. Skirtumas tas, kad užduotį studentas atlieka savarankiškai semestro eigoje ir atsiskaito dėstytojui individualiai interaktyviai atsakydamas į pateiktus klausimus.
Egzaminas		Sesijos metu	Egzamino metu kiekvienam studentui pasirinktinai leidžiama perrašyti vieną kontrolinį darbą. Gavus prastesnį balą paliekamas ankstesnis įvertinimas.
Egzamino perlaikymas eksternu			Studentas privalo atsiskaityti už visas dalis (80%) ir pristatyti parengtą individualią užduotį (20%).

Autorius	**Leidi**	**Pavadinimas**	**Periodinio**	**Leidimo vieta ir leidykla ar**

	mo metai		leidinio Nr. ar leidinio tomas	internetinė nuoroda
Privaloma literatūra				
1. NumPy development team	2021	Numpy quickstart		Numpy quickstart
2. Bart Baesens	2014	Analytics in a Big Data World		John Wiley & Sons, Inc. (Nuoroda į 1-ą skyrių internete)
3. William E. Shotts, Jr.	2000-2021	LinuxCommand.org		https://linuxcommand.org/index.php
4. Apache Spark development team	2021	Documentation		https://spark.apache.org/docs/latest/
5. Slurm Team	2021	Documentation		https://slurm.schedmd.com/documentation.html
Papildoma literatūra				
6. Jupyter project development team	2021	Documentation		https://jupyter.org/documentation
7. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani	2013	An Introduction to Statistical Learning: with Applications in R		Springer
8. Bernd Klein	2011-2020	Numerical Programming with Python		https://www.python-course.eu/numerical_programming_with_python.php

COURSE UNIT (MODULE) DESCRIPTION

Course unit (module) title	Code
Big data software	

Lecturer(s)	Department(s) where the course unit (module) is delivered
Coordinator: dr. Tadas Danielius **Other(s):**	Department of Statistical Analysis

Study cycle	Level of course	Type of the course unit (module)
First	Beginners	Compulsory

Mode of delivery	Period when the course unit (module) is delivered	Language(s) of instruction
Face-to-face	Fifth (spring) semester	Lithuanian

Requirements for students	
Prerequisites: basics of Python; ability to understand English at the level of independent user (B1 according to CEFR classification).	**Additional requirements (if any):**

Course (module) volume in credits	Total student's workload	Contact hours	Self-study hours
5	125	48	77

Purpose of the course unit (module): programme competences to be developed (the number in the brackets coincides with that given in the official description of the programme)

- Gebėjimas analizuoti, sisteminti, mokytis ir taikyti įgytas žinias praktikoje (1);
- Gebėjimas rinkti, valdyti ir tvarkyti duomenis (5);
- Gebėjimas rinktis tinkamą analizės metodologiją bei jai reikalingus įrankius (6);
- Gebėjimas interpretuoti ir reprezentuoti analizės rezultatus (7);

Learning outcomes of the course unit (module); after completing the course students should:	Teaching and learning methods	Assessment methods
be familiar with specific issues encountered in big data analysis; be able to use several software tools designed for big data analysis; be able to read the literature devoted to big data software and apply the gained knowledge practically.	Lectures, problem solving and reading, assignments	Tests, individual tasks

Content: breakdown of the topics	Contact hours							Self-study work: time and assignments	
	Lectures	Tutorials	Seminars	Exercises	Laboratory work	Internship/work placement	Contact hours	Self-study hours	Assignments
1. Introduction. Introduction to big data and big data analytics. Cover main concepts and issues surrounding big data, along with typical tasks and processing models. Brief review of data management and integration techniques, discussing efficient data storage, retrieval, and integration methods. Various software tools commonly used in big data analytics are also explored.	2						**2**	**5**	Read ch. 1 of [2]. Find an introductory article on big data analysis and read it on your own.
2. NumPy. Central object, its role in python's ecosystem for Big Data Analysis, basic operations.	2				4		**6**	48	Regularly accomplish exercises designed for gaining of appropriate level of skills. Since exercises depend on and evolve with a version of particular software, they are not explicated here but sequentially given during the factual period of delivery of the course.
3. Python for parallel processing. Multiprocessing, mpi4py, ipyparallel: review and functioning basics. Some alternatives.	4				4		**8**		
4. Apache Spark. pyspark.sql for data manipulation. pyspark.mllib API by example (simplest models).	4				4		**8**		
5. Distributed computations. Linux command line basics. Running jobs on the cluster managed by SLURM.	4				4		**8**		
6. Assessments and exam.					16		**16**	**24**	Prepare for tests and assessment.
Total							**48**	**77**	

Assessment strategy	Weight,%	Deadline	Assessment criteria
Test 1	20	4th study week	The test consists of several practical tasks intended to check the level of knowledge obtained. The total weight of these tasks equals to 1 point. The weight of each task ranges from 0.1 to 1 point. Tasks are designed to be solved by making use of computer and appropriate software. Each task is evaluated as follows: a) the task is divided into
Test 2	20	8th study week	
Test 3	20	12th study week	

Test 4	20	16th study week	parts and each part is assigned an appropriate amount of points; b) if student accomplishes the part without mistakes, the whole amount of that part is attained; otherwise, the amount is reduced considering the mistakes made; c) the parts are evaluated independently.
Individual task	20	One week before exam.	Evaluation principles are the same as those of tests. The major difference is due to the form of assessment: students have to account for their work in an interactive mode.
Exam		The final examination session	During the exam, the students are optionally allowed to rewrite one of the four tests taken during the regular semester. In case of worse outcome, result of test remains unchanged.
External examination			The student must perform all parts (80%) and present a prepared individual assignment (20%).

Author	**Year of publication**	**Title**	**Issue of a periodical or volume of a publication**	**Publishing place and house or web link**
Compulsary reading				
1. NumPy development team	2021	Numpy quickstart		Numpy quickstart
2. Bart Baesens	2014	Analytics in a Big Data World		John Wiley & Sons, Inc. (Nuoroda į 1-ą skyrių internete)
3. William E. Shotts, Jr.	2000-2021	LinuxCommand.org		https://linuxcommand.org/index.php
4. Apache Spark development team	2021	Documentation		https://spark.apache.org/docs/latest/
5. Slurm Team	2021	Documentation		https://slurm.schedmd.com/documentation.html
Optional reading				
6. Jupyter project development team	2021	Documentation		https://jupyter.org/documentation
7. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani	2013	An Introduction to Statistical Learning: with Applications in R		Springer
8. Bernd Klein	2011-2020	Numerical Programming with Python		https://www.python-course.eu/numerical_programming_with_python.php