

STUDIJŲ DALYKO APRAŠAS

Dalyko kodas	Dalyko apimtis kreditais	Institucija	Fakultetas	Katedra
INF8003	9	VDU	IF	Taikomosios informatikos

Pavadinimas

KALBOS TECHNOLOGIJOS IR INFORMACIJOS MODELIAVIMAS

Pavadinimas anglų kalba

LANGUAGE TECHNOLOGIES AND INFORMATION RETRIEVAL

Studijų būdas	Kreditų skaičius
Paskaitos	2
Konsultacijos	2
Individualus darbas	5

Dalyko anotacija lietuvių kalba (iki 500 simbolių)

Kalbos technologijos (KT) – sritis, kuri apima metodų (kurių pagalba programos geba suprasti, analizuoti, modifikuoti, generuoti natūralią žmogaus kalbą) kūrimą ir taikymą. Žmogaus kalba yra daugiausiai iššūkių keliantis tyrimo objektas, reikalaujantis didelių duomenų ir ištobulintų dirbtinio intelekto modelių. Dėl šių priežasčių KT šiuo metu sparčiausiai besivystanti dirbtinio intelekto šaka. Studijų metu doktorantai bus supažindinti su KT problematika, pažangiausiais metodais, kuriuos praktiškai išbandys sprendami įvairius kalbos uždavinius.

Informacijos paieška (angl. Information Retrieval, IR) apima visus informacijos paieškos proceso aspektus: informacijos surinkimą, analizę, indeksavimą, tinkamiausios informacijos radimą, bei jos pateikimą. Studijų metu doktorantai bus supažindinti su IR problematika, pažangiausiais metodais ir įrankiais, kuriuos išbandys sprendami įvairius uždavinius.

Dalyko anotacija anglų kalba (iki 500 simbolių)

Language technologies (LTs) is a field that includes the development and application of methods (with the help of which programs can understand, analyze, modify, and generate natural human language). Human language is the most challenging research subject, requiring big data and advanced artificial intelligence models. Due to these reasons, LTs are currently the fastest-growing branch of artificial intelligence. During the studies, Ph.D. students will be introduced to LTs issues, the most advanced methods, which they will test in practice while solving various language problems.

Information Retrieval (IR) covers information collection, analysis, indexing, retrieval and representation, and presentation. Doctoral students will be introduced to different IR topics, and the most advanced methods and tools, which will be used in practice.

Dalyko poreikis ir aktualumas

KT IR poreikis iš vartotojo perspektyvos milžiniškas: tai paskatino atsirasti ChatGPT, Quillbot, Grammarly, Google paiešką ir kt. sistemas. Natūralios kalbos suvokimas, analizavimas, generavimas

yra vienos sudėtingiausių dirbtinio intelekto problemų, todėl inovacijos dažnai pirmiausiai atsiranda KT srityje ir vėliau yra perkeliamos į kitas sritis.

KT problemų sprendimas reikalauja mokslinio tyrimo: 1) detalios problemos analizės ir formalizavimo; 2) tinkamai parinktos dirbtinio intelekto metodikos; 3) eksperimentinių tyrimų; 4) teisingo rezultatų interpretavimo ir jų statistinio vertinimo. Naujausių dirbtinio intelekto metodikų žinojimas ir moksliniai tyrimai yra būtent tai ko reikia doktorantūroje.

Dalyko tikslas

Dalyko tikslas: 1) supažindinti doktorantus su KT ir IR problematika, iššūkiais, metodikomis, technologijomis, įrankiais; 2) išugdyti gebėjimą analizuoti ir formalizuoti sprendžiamą kalbos, informacijos modeliavimo uždavinį, teisingai parinkti metodiką, atlikti eksperimentinius tyrimus, vertinti gautus rezultatus, numatyti priemones metodikos tobulinimui.

Dalyko turinys, temos ir studijų metodai

Dalyko turinys realizuojamas per šias temas:

1. KT problematika, iššūkiai
2. KT sprendžiami uždaviniai (uždavinių tipai) ir esami įrankiai
3. Žodžių įterpiniai
4. Teksto klasifikavimas ir generavimas klasikiniiais giliaisiais neuroniniais tinklais (konvoliuciniu, rekurentiniu, ilgos trumpalaikės atminties, dvikrypčiu ilgos trumpalaikės atminties)
5. Atminties mechanizmas ir jo įtaka
6. Transformaciniai modeliai
7. BERT transformacinis kalbos modelis ir jo taikymai teksto klasifikavimui
8. Sakinių transformaciniai modeliai.
9. GPT metodai ir jų taikymas teksto generavimui, ChatGPT
10. Daugiakalbiškumas ir tarpkalbiškumas: kaip iš to laimi mažiau resursų turinčios kalbos
11. Eksperimentinių tyrimų eiga, rezultatų interpretavimas (vertinimo metrikos, rezultatų statistinė analizė, palyginamieji tyrimai).
12. Informacijos modeliavimo problematika, iššūkiai.
13. Informacijos parengimas, indeksavimas, saugojimas.
14. Pagrindiniai informacijos paieškos modeliai: loginis, vektorinis.
15. Paieškos modelių kokybės įvertinimas.
16. Paieška internete, PageRank ir kiti modeliai.

Studijų metodai – savarankiškas studijavimas (pateiktos literatūros analizė, pasirinkto mini mokslinio tyrimo atlikimas) ir individualios konsultacijos.

Išklause dalyką, studentai gebės:

1. tinkamai suformuluoti sprendžiamą kalbos ir informacijos modeliavimo uždavinį.
2. analitiniu būdu parinkti galimas uždavinio sprendimo metodikas (atsižvelgiant į uždavinio tipą, duomenų aibę, kalbą ir kt.).
3. programiškai realizuoti pasirinktas metodikas.
4. atlikti eksperimentinius tyrimus (mokyti modelį, optimizuoti hiper-parametrus, testuoti).
5. įvertinti ir lyginti rezultatus, surasti geriausią metodą.
6. suformuluoti rekomendacijas panašioms kalbos uždaviniams.
7. atlikti klaidų analizę ir pasiūlyti tolimesnius metodikos tobulinimus.

Studijų pasiekimų vertinimas

Teorinių žinių įsisavinimas vertinamas egzaminuojant. Egzamino įvertinimas sudaro 50% galutinio pažymio.

Gebėjimas sprendžiamai kalbos problemai parinkti ir taikyti KT ir IR metodus vertinamas individualiomis užduotimis ir tai sudaro 50% galutinio pažymio.

Literatūra

1. Stanfordo universiteto kursas NLP with Deep Learning (2021 m) youtube platformoje: <https://www.youtube.com/playlist?list=PLoROMvodv4rOSH4v6133s9LFPRHjEmbmJ>
2. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. <https://nlp.stanford.edu/IR-book/information-retrieval-book.html>
3. Croft, W.B., Metzler, D., & Strohman, T. (2009). Search Engines - Information Retrieval in Practice. <https://ciir.cs.umass.edu/irbook/>
4. Publikacijų kalbos technologijų srityje archyvas: <https://aclanthology.org/>
5. Dirbtinio intelekto platforma (modeliai, duomenų aibės): <https://huggingface.co/>
6. Mokslinių publikacijų ir programinio kodo archyvas: <https://paperswithcode.com/>

Dalyko programos rengėjas/jai

Vardas, pavardė	Institucija	Pedagoginis vardas, mokslo laipsnis	Elektroninio pašto adresas
Tomas Krilavičius	VDU	prof. dr.	tomas.krilavicius@vdu.lt
Jurgita Kapočiūtė-Dzikiene	VDU	prof. dr.	jurgita.kapociute-dzikiene@vdu.lt