

А.БАЙТҰРСЫНОВ АТЫНДАҒЫ ҚОСТАНАЙ Өңірлік университеті
КОСТАНАЙСКИЙ РЕГИОНАЛЬНЫЙ УНИВЕРСИТЕТ
ИМЕНИ А.БАЙТҰРСЫНОВА
KOSTANAY REGIONAL UNIVERSITY NAMED AFTER A.BAITURSYNOV



ҚОСЫМША БІЛІМ БЕРУ
БАҒДАРЛАМАСЫ
(MINOR)
BIG DATA ЖӘНЕ КОМПЬЮТЕРЛІК ЛИНГВИСТИКА
НЕГІЗДЕРІ

ДОПОЛНИТЕЛЬНАЯ
ОБРАЗОВАТЕЛЬНАЯ ПРОГРАММА
(MINOR)
BIG DATA И ОСНОВЫ КОМПЬЮТЕРНОЙ
ЛИНГВИСТИКИ

ADDITIONAL
EDUCATIONAL PROGRAM
(MINOR)
BIG DATA AND THE BASICS OF COMPUTATIONAL
LINGUISTICS

Қостанай, 2023

ӘЗІРЛЕУШІЛЕР/ РАЗРАБОТЧИКИ/ DEVELOPERS:

Салыкова Ольга Сергеевна, бағдарламалық қамтамасыз ету кафедрасының аға оқытушысы / Салыкова Ольга Сергеевна, к.т.н., ассоциированный профессор (доцент) кафедры программного обеспечения / Olga Salykova, Candidate of Technical Sciences, Associate Professor of the Department of Software

Сатмаганбетова Жанар Зарлыкановна, бағдарламалық қамтамасыз ету кафедрасының аға оқытушысы / Сатмаганбетова Жанар Зарлыкановна, старший преподаватель кафедры программного обеспечения / Zhanar Satmaganbetova, senior lecturer of the Department of Software

А.Байтұрсынов атындағы ҚӨУ-дың оқу-әдістемелік кеңес отырысында бекітілді, 11.04.2023 ж. № 4 хаттама

Утвержден на заседании учебно-методического совета КРУ имени А.Байтұрсынова, протокол от 11.04.2023 г. № 4

Published by decision educational and methodical council of the Kostanay regional university named after A.Baitursynov (Protocol № 4 from 11.04.2023)

Курстың қысқаша сипаттамасы / Краткое описание курса / Brief description of the course:

Big data (биг дата немесе аудармада деректердің үлкен көлемі) – бұл ақпарат ағынының, технологияның, оны өңдеу әдістерінің және талдау жүйесінің жалпылама атауы. IT-саласы адамдардың айналасындағы кеңістікті сенімді түрде толтырады. Алайда, алынған білім "ешқайда" кете алмайды, ал үлкен өлшемді ескере отырсақ, сақтау көлемдері де ауқымды болуы керек. Адамзат ұзақ уақыт бойы сандық медиаға көшті, ал олардың барлығы мөлшері бойынша ерекшеленеді.

Үлкен ақпарат массивімен жұмыс істеу үшін нақты тапсырмаларды шешу үшін арнайы құралдар мен әдістер жиынтығы қажет. Шын мәнінде, әртүрлі мәліметтер жиынтығы және олармен жұмыс істеу құралдары Big Data терминін анықтайды. Бұл әлеуметтік-экономикалық құбылыс ақпараттың үлкен мөлшерімен жұмыс істеуге мүмкіндік беретін кеңейтілген технологиялардың пайда болуымен тікелей байланысты.

Үлкен деректерді талдау мамандары-бұл көзге көрінбейтін нәрсені қарастыруға көмектесетін мамандардың жаңа түрі. Олардың көмегімен деректермен жұмыс істеу жаңа сапалы деңгейге шығады. Деректер талдаушысы өзінің бизнес тарихында көптеген талап етілмеген деректерді жинаған компанияларда қажет.

Тіл технологиясының дамып келе жатқан саласында кәсіби маман болғысы келетіндер үшін. Бұл цифрлық әлемді адамдар үшін ыңғайлы ететін тілдік технологиялар.

Компьютерлік лингвистика (КЛ, Computational Linguistics). - бұл қолданбалы аймақ, яғни бұл компьютерді қолдана отырып, белгілі бір тілде жасалған тілдерді, мәтіндерді өңдеу арқылы шешуге болатын тапсырмалар жиынтығы.

Бұл бағытты игеру лингвистикалық софт (сөйлеуді тану, мәтінді автоматты өңдеу, машиналық аударма бағдарламалары) әзірлейтін компанияларда жұмыс істеуге, электрондық сөздіктерді, анықтамалықтарды, энциклопедияларды, интернеттегі сайттарды құруға қатысуға мүмкіндік береді

Big data (биг дата или в переводе большой объем данных) – это обобщающее название для информационного потока, технологии, методов его обработки и системы анализа. IT-сфера уверенно заполняет пространство вокруг людей. Однако получаемые знания не могут уходить «вникуда», а учитывая колоссальный размер, хранилище должно быть объемным. Человечество уже давно перешло на цифровые носители, при этом все они отличаются по размеру.

Для работы с большими массивами информации нужен специальный набор инструментов и методик, чтобы с их помощью решать конкретные поставленные задачи. По сути, совокупность различных данных и инструментарий работы с ними и определяет термин Big Data. Этот социально-экономический феномен напрямую связан с появлением масштабируемых технологий, которые позволяют работать с огромным количеством информации.

Специалисты в области анализа больших данных - это новый тип профессионалов, которые помогают рассмотреть то, что нельзя увидеть невооруженным глазом. С их помощью работа с данными выходит на новый качественный уровень. Аналитик данных нужен именно в тех компаниях, которые накопили за свою бизнес-историю огромное количество невостребованных данных.

Для тех, кто хочет стать профессионалом в развивающейся области языковых технологий. Именно языковые технологии сделают цифровой мир комфортным для человека.

Компьютерная лингвистика (КЛ, Computational Linguistics) – это прикладная область, то есть это набор задач, которые мы можем решить, обрабатывая языки, тексты, созданные на определенном языке, при помощи компьютера.

Освоение данного направления может позволить работать в компаниях, разрабатывающих лингвистический софт (программы распознавания речи, автоматической

обработки текста, машинного перевода), принимать участие в создании электронных словарей, справочников, энциклопедий, сайтов в Интернете.

Big data (big date or in translation a large amount of data) is a general name for an information flow, technology, methods of its processing and analysis system. The IT sphere is confidently filling the space around people. However, the knowledge gained cannot go "nowhere", and given the colossal size, the storage should be voluminous. Humanity has long since switched to digital media, with all of them differing in size.

To work with large amounts of information, you need a special set of tools and techniques in order to use them to solve specific tasks. In fact, the totality of various data and tools for working with them define the term Big Data. This socio-economic phenomenon is directly related to the emergence of scalable technologies that allow you to work with a huge amount of information.

Big data analysts are a new type of professional who helps to see things that cannot be seen with the naked eye. With their help, work with data reaches a new qualitative level. A data analyst is needed precisely in those companies that have accumulated a huge amount of unclaimed data over their business history.

For those who want to become a professional in the emerging field of language technology. It is language technologies that will make the digital world comfortable for humans.

Computational Linguistics (CL, Computational Linguistics). Is an application area, that is, a set of tasks that we can solve by processing languages, texts created in a particular language using a computer.

Mastering this direction can allow you to work in companies that develop linguistic software (speech recognition programs, automatic text processing, machine translation), to take part in the creation of electronic dictionaries, reference books, encyclopedias, Internet sites

Мақсаты/Цель/Aim: үлкен көлемдегі ақпаратты дұрыс талдау үшін үлкен деректерді талдау әдістемесін меңгеру, мәтінді автоматты өңдеу технологиясын, сондай-ақ мәтінді өңдеуге байланысты әртүрлі мәселелерді шешу үшін компьютерлік лингвистикада модельдеу әдістерін игеру / освоить методику анализа больших данных, для правильного анализа огромного количества информации; освоение технологии автоматической обработки текста, а также способов моделирования в компьютерной лингвистике для решения различных задач, связанных с обработкой текстов / to master the technique of analyzing big data, for the correct analysis of a huge amount of information, mastering the technology of automatic text processing, as well as methods of modeling in computational linguistics for solving various problems related to text processing.

Оқу міндеттері / Учебные задачи / Learning Objectives: Big Data, компьютерлік лингвистикадағы негізгі терминдер мен ұғымдарды білу; осы бағытта жұмыс істеу үшін негізгі бағдарламаларды қолдана білу / знать основные термины и понятия в Big Data, компьютерной лингвистике; уметь использовать основные программы для работы в данном направлении / know the basic terms and concepts in Big Data, computational linguistics; be able to use basic programs to work in this direction.

Қалыптастырылатын негізгі құзыреттер /Базовые компетенции, которые предстоит сформировать/ Basic competencies to be formed: үлкен деректерді талдау және өңдеу принциптерін меңгерген; қойылған міндеттерге қатысты деректерді өңдеудің заманауи алгоритмдерін таңдай және қолдана алады, алынған нәтижелерді талдайды және негізделген қорытындылар жасайды; әлеуметтік - экономикалық міндеттерді шешу үшін сапалы және сандық деректерді өңдеу, визуализациялау және талдау әдістерін меңгерген дағдылары бар; нақты жобалар мысалында деректерді басқару. Мәтінді автоматты өңдеу технологиясын тану, сондай-ақ мәтінді өңдеуге байланысты түрлі міндеттерді шешу үшін бағдарламалық құралдарды пайдалану дағдылары / Владеет принципами анализа и обработки больших данных; умеет выбирать и применять современные алгоритмы обработки данных относительно поставленных задач, анализирует полученные результаты и делает обоснованные выводы; имеет навыки овладения методами обработки, визуализации и

анализа качественных и количественных данных для решения социально-экономических задач; управления данными на примере конкретных проектов. Навыки в распознавании технологии автоматической обработки текста, а также в использовании программных средств для решения различных задач, связанных с обработкой текста / Owns the principles of analysis and processing of big data; knows how to choose and apply modern algorithms for data processing in relation to the assigned tasks, analyzes the results obtained and makes reasonable conclusions; has the skills to master the methods of processing, visualization and analysis of qualitative and quantitative data for solving social and economic problems; data management on the example of specific projects. Skills in recognition of automatic text processing technology, as well as in the use of software tools to solve various problems related to text processing.

Оқытудың нәтижелері / Результаты обучения / Learning outcomes:

Статистикалық есептеулер мен графика үшін R тілін қолдану. Үлкен деректерді талдау әдістерін қолдану. Автоматты мәтінді өңдеу технологиясын және компьютерлік лингвистиканың басқа ғылымдармен байланысын тану. Мәтіндерді өңдеумен байланысты түрлі мәселелерді шешу үшін бағдарламалық құралдармен жұмыс істеу дағдыларын көрсету. Жасанды және табиғи тілдердің элементтерін модельдеуге байланысты заманауи әдістер мен құралдарды талдау.

Использование языка R для статистических вычислений и графики. Применение методов анализа больших данных. Распознавание технологии автоматической обработки текста и взаимосвязи компьютерной лингвистики с другими науками. Демонстрация навыков работы с программными инструментальными средствами для решения различных задач, связанных с обработкой текстов. Анализ современных методов и инструментария, связанного с моделированием элементов искусственных и естественных языков.

Use of the R language for statistical calculations and graphics. Application of methods of analysis of big data. Recognition of automatic text processing technology and the relationship of computational linguistics with other sciences. Demonstration of skills in working with software tools for solving various problems related to word processing. Analysis of modern methods and tools related to modeling elements of artificial and natural languages.

Пәндер / Дисциплины / Disciplines:

1. R Негіздері (5 кредит, ауызша емтихан) / Основы R (5 кредитов, устный экзамен) / R Basics (5 credits, oral examination)

2. Big Data Сараптамасы (5 кредит, ауызша емтихан) / Аналитика Big Data (5 кредитов, УЭ) / Big Data analytics (5 credits, oral examination)

3. Мәтінді автоматты өңдеу технологиялары (5 кредит, ауызша емтихан) / Технологии автоматической обработки текста (5 кредитов, УЭ) / Technologies of automatic text processing (5 credits, oral examination)

4. Компьютерлік лингвистикадағы модельдеу (5 кредит, ауызша емтихан) / Моделирование в компьютерной лингвистике (5 кредитов, УЭ) / Modeling in computational linguistics (5 credits, oral examination)

Еңбек сыйымдылығы / Трудоемкость / Labor intensity: 20 кредитов.

Пререквизиттер / Пререквизиты / Prerequisite: Ақпараттық-коммуникациялық технологиялар / Информационно-коммуникационные технологии / Information and Communication Technologies

Minimum number of listeners / Минимальное количество слушателей / Тыңдаушылардың ең аз саны: 10

Maximum number of listeners / Максимальное количество слушателей / Тыңдаушылардың максималды саны: 50

Пән 1 / Дисциплина 1 / Discipline 1

R Негіздері / Основы R / R Basics

Пәннің қысқаша сипаттамасы / Краткое описание дисциплины / Discipline

Summary: R-бұл статистикалық есептеулер мен графикаға арналған қуатты тіл, ол деректерді өңдеу саласындағы кез-келген тапсырманы жеңе алады. Ол барлық маңызды операциялық жүйелерде жұмыс істейді және мыңдаған мамандандырылған модульдер мен қызметтік бағдарламаларды қолдайды. Мұның бәрі R-ді шикі деректер тауларынан пайдалы ақпарат алудың керемет құралы етеді. "R негіздері" - бұл тілді практикалық міндеттерге ерекше назар аудара отырып оқытуға шақыратын пән. Онда статистикалық мәліметтерді өңдеудің пайдалы мысалдары келтірілген және күрделі және толық емес мәліметтермен жұмыс істеудің талғампаз әдістері сипатталған. Пәнді оқығаннан кейін оқушы үлкен деректер көздерінің орналасуын және оларды алу әдістерін, өңдеудің негізгі принциптерін біледі; қойылған міндеттерге қатысты деректерді өңдеудің заманауи алгоритмдерін таңдай және қолдана алады, алынған нәтижелерді талдайды және негізделген қорытындылар жасайды; әлеуметтік - экономикалық мәселелерді шешу үшін сапалы және сандық деректерді өңдеу, визуализациялау және талдау әдістерін меңгеру; нақты жобалар мысалында деректерді басқару дағдылары бар/ R - это мощный язык для статистических вычислений и графики, который может справиться поистине с любой задачей в области обработки данных. Он работает во всех важных операционных системах и поддерживает тысячи специализированных модулей и утилит. Все это делает R замечательным средством для извлечения полезной информации из гор сырых данных. «Основы R» - это дисциплина, которая призывает обучению этому языку с особым вниманием к практическим задачам. В ней представлены полезные примеры статистической обработки данных и описаны изящные методы работы с запутанными и неполными данными. После изучения дисциплины обучающийся знает расположение источников больших данных и методы их извлечения, основные принципы обработки; умеет выбирать и применять современные алгоритмы обработки данных относительно поставленных задач, анализирует полученные результаты и делает обоснованные выводы; имеет навыки овладения методами обработки, визуализации и анализа качественных и количественных данных для решения социально-экономических задач; управления данными на примере конкретных проектов. / R is a powerful statistical computing and graphics language that can handle virtually any data processing task. It runs on all important operating systems and supports thousands of specialized modules and utilities. All of this makes R a great tool for extracting useful information from mountains of raw data. "R Fundamentals" is a discipline that encourages teaching this language with a focus on practical problems. It provides useful examples of statistical data manipulation and describes graceful techniques for dealing with confusing and incomplete data. After studying the discipline, the student knows the location of big data sources and methods for their extraction, the basic principles of processing; knows how to choose and apply modern algorithms for data processing in relation to the assigned tasks, analyzes the results obtained and makes reasonable conclusions; has the skills to master the methods of processing, visualization and analysis of qualitative and quantitative data for solving social and economic problems; data management on the example of specific projects

Сабақ түрлері / Виды занятий / Types of lesson: дәріс/ лекция/ lecture; зертханалық жұмыс/ лабораторная работа/ laboratory work

Оқытушы / Преподаватель / Teacher Салыкова Ольга Сергеевна, бағдарламалық қамтамасыз ету кафедрасының аға оқытушысы / Салыкова Ольга Сергеевна, к.т.н., ассоциированный профессор (доцент) кафедры программного обеспечения / Olga Salykova, Candidate of Technical Sciences, Associate Professor of the Department of Software

Пән 2 / Дисциплина 2 / Discipline 2

Big Data Сараптамасы / Аналитика Big Data / Big Data Analytics

Пәннің қысқаша сипаттамасы / Краткое описание дисциплины / Discipline

Summary: Үлкен деректерді талдау мамандары-бұл көзге көрінбейтін нәрсені қарастыруға көмектесетін мамандардың жаңа түрі. Олардың көмегімен деректермен жұмыс істеу жаңа сапалы деңгейге шығады. Деректер талдаушысы өзінің бизнес тарихында көптеген талап етілмеген деректерді жинаған компанияларда қажет. Сарапшы-аналитик статистика және SQL сияқты стандартты құралдарды ғана меңгеріп қоймай, үлкен деректермен жұмыс істеудің негізгі қағидаттарын білуі, Big data бойынша шешімдерді іске асыру үшін Hadoop жүйесінің компоненттері және 22 бұлтты платформалар туралы түсінікке ие болуы қажет. Бұл курста әр түрлі деңгейдегі мәліметтермен жұмыс қарастырылады. Үлкен деректерді өңдеу әдістері мен технологиялары. Құрылымдық және құрылымданбаған деректерді іздеу. Негізгі бизнес мәселелерін шешу үшін статистикалық әдістерді қолдану, Машиналық оқыту және предиктивті аналитиканың озық әдістері. Пәнді оқығаннан кейін оқушы үлкен деректерді талдау негіздерін, қойылған міндеттерді шешудің негізгі алгоритмдерін біледі, алгоритмдердің, олардың артықшылықтары мен кемшіліктерінің арасындағы байланысты түсінеді; Data Mining есептерін жіктей алады, деректерді интеллектуалды талдау әдістерін қолдана алады; деректерді визуалды зерттеу және ұсыну үшін кең графикалық мүмкіндіктерді қолдана алады; өзінің кәсіби қызметінде заманауи ақпараттық технологияларды қолдана алады; ақпараттың үлкен ауқымын (Big data) өңдеу дағдыларына ие; әртүрлі сипаттағы деректерді; шешімдер қабылдауды қолдау жүйелері саласында машиналық оқытудың қазіргі заманғы әдістерін және деректерді көп өлшемді талдау саласында міндеттерді шешу әдіснамасын қолдану; қазіргі заманғы бағдарламалық құралдарды қолдану / Специалисты в области анализа больших данных - это новый тип профессионалов, которые помогают рассмотреть то, что нельзя увидеть невооружённым глазом. С их помощью работа с данными выходит на новый качественный уровень. Аналитик данных нужен именно в тех компаниях, которые накопили за свою бизнес-историю огромное количество не востребовавшихся данных. Специалисту-аналитику нужно владеть не только стандартными инструментами вроде статистики и SQL, но и знать основные принципы работы с большими данными, иметь представление о компонентах системы Hadoop и 22 облачных платформах для реализации решений по big data. В данном курсе рассматривается работа с данными в различной степени структуризации. Методы и технологии обработки больших данных. Интеллектуальный анализ структурированных и неструктурированных данных. Использование методов статистики, машинное обучение и продвинутые методы предиктивной аналитики для решения ключевых бизнес-задач. После изучения дисциплины обучающийся знает основы анализа больших данных, основные алгоритмы решения поставленных задач, понимает зависимости между алгоритмами, их преимуществами и недостатками; умеет классифицировать задачи Data Mining, применять методы интеллектуального анализа данных; применять обширные графические возможности для визуального исследования и представления данных; использовать современные информационные технологии в своей профессиональной деятельности; имеет навыки обработки больших массивов информации (Big data); анализа данных различной природы; применения современных методов машинного обучения в области систем поддержки принятия решений и методологии решения задач в области многомерного анализа данных; применения современного программного инструментария/ Big data analysts are a new type of professional who helps to see things that cannot be seen with the naked eye. With their help, work with data reaches a new qualitative level. A data analyst is needed precisely in those companies that have accumulated a huge amount of unclaimed data over their business history. An analyst needs to know not only standard tools like statistics and SQL, but also know the basic principles of working with big data, have an understanding of the components of the Hadoop system and 22 cloud platforms for implementing big data solutions. Using statistical techniques, machine learning and

advanced predictive analytics techniques to solve key business problems. After studying the discipline, the student knows the basics of big data analysis, the main algorithms for solving the tasks, understands the relationship between algorithms, their advantages and disadvantages; knows how to classify Data Mining tasks, apply data mining methods; apply extensive graphical capabilities for visual exploration and data presentation; use modern information technologies in their professional activities; has the skills to process large amounts of information (Big data); analysis of data of various nature; application of modern machine learning methods in the field of decision support systems and methodology for solving problems in the field of multivariate data analysis; application of modern software tools.

Сабақ түрлері / Виды занятий / Types of lesson: дәріс/ лекция/ lecture; зертханалық жұмыс/ лабораторная работа/ laboratory work

Оқытушы /Преподаватель / Teacher Салыкова Ольга Сергеевна, бағдарламалық қамтамасыз ету кафедрасының аға оқытушысы / Салыкова Ольга Сергеевна, к.т.н., ассоциированный профессор (доцент) кафедры программного обеспечения / Olga Salykova, Candidate of Technical Sciences, Associate Professor of the Department of Software

Пән 3 / Дисциплина 3 / Discipline 3

Мәтінді автоматты өңдеу технологиялары /
Технологии автоматической обработки текста /
Technologies of automatic text processing

Пәннің қысқаша сипаттамасы / Краткое описание дисциплины / Discipline

Summary: Тіл технологиясының дамып келе жатқан саласында кәсіби маман болғысы келетіндер үшін. Бұл цифрлық әлемді адамдар үшін ыңғайлы ететін тілдік технологиялар. Компьютерлік лингвистика (КЛ, Computational Linguistics). Компьютерлік лингвистика және жасанды интеллект. Компьютерлік лингвистика бөлімдері. Компьютерлік лингвистиканың білімнің басқа салаларымен байланысы. Ақпараттық іздеу ақпараттық іздеу сипаттамасы. Индекстеу тілдері сұрақ - жауап жүйелері. Жақын мәтіндерді (күжаттарды) іздеу. Мәтіндерді кластерлеу және жіктеу. Ақпараттық іздеу жүйесі. Компьютерлік лексикография. Машиналық аударма. Мәтіндерден ақпарат алу. Автоматты түйіндеме. Мазмұнды талдау. Мәтін синтезі. Сөйлеу синтезін тану. Дыбысты белгілеу және лингвистикалық талдауға арналған компьютерлік бағдарламалар. Корпустық лингвистика. Мәтіндерді лингвистикалық талдау мен өңдеуге арналған бағдарламалық құралдар. АӨТ талдағышы. Лингвистикалық (Тілдік) ресурстар. Лингвистикалық процессорларды дамытуға арналған платформалар/ Для тех, кто хочет стать профессионалом в развивающейся области языковых технологий. Именно языковые технологии сделают цифровой мир комфортным для человека. Компьютерная лингвистика (КЛ, Computational Linguistics). Компьютерная лингвистика и искусственный интеллект. Разделы компьютерной лингвистики. Связь компьютерной лингвистики с другими областями знаний. Информационный поиск Характеристики информационного поиска. Языки запросов. Языки индексирования Вопросно-ответные системы. Поиск близких текстов (документов). Кластеризация и классификация текстов. Информационнопоисковые система. Компьютерная лексикография. Машинный перевод. Извлечение информации из текстов. Автоматическое резюмирование. Контентный анализ. Синтез текстов. Распознавание синтез речи. Компьютерные программы для разметки и лингвистического анализа звука. Корпусная лингвистика. Программные инструменты для лингвистического анализа и обработки текстов. Анализатор АӨТ. Лингвистические ресурсы. Платформы для разработки лингвистических процессоров / For those who want to become a professional in the emerging field of language technology. It is language technologies that will make the digital world comfortable for humans. Computational Linguistics (CL, Computational Linguistics). Computational linguistics and artificial intelligence. Sections of computational linguistics. Communication of computational linguistics with other areas of knowledge. Information retrieval Characteristics of information retrieval. Query languages. Indexing languages. Question-

answer systems. Search for related texts (documents). Clustering and classification of texts. Information retrieval system. Computer lexicography. Machine translate. Extracting information from texts. Automatic summarization. Content analysis. Synthesis of texts. Speech synthesis recognition. Computer programs for marking and linguistic analysis of sound. Corpus linguistics. Software tools for linguistic analysis and word processing. AOT analyzer. Linguistic resources. Linguistic Processor Development Platforms

Сабақ түрлері / Виды занятий / Types of lesson: дәріс/ лекция/ lecture; зертханалық жұмыс/ лабораторная работа/ laboratory work

Оқытушы /Преподаватель / Teacher Сатмаганбетова Жанар Зарлыкановна, бағдарламалық қамтамасыз ету кафедрасының аға оқытушысы / Сатмаганбетова Жанар Зарлыкановна, старший преподаватель кафедры программного обеспечения / Zhanar Satmaganbetova, senior lecturer of the Department of Software

Пән 4 / Дисциплина 4 / Discipline 4

Компьютерлік лингвистикадағы модельдеу /
Моделирование в компьютерной лингвистике /
Modeling in computational linguistics

Пәннің қысқаша сипаттамасы / Краткое описание дисциплины / Discipline

Summary: Жасанды тілдер. Жасанды тілдердің жіктелуі. Бағдарламалау тілдері. Жасанды тілдерді сипаттаудың ресми әдістері. Ресми грамматика. Мета-тіл түсінігі. Табиғи тілдер және есептеу. Тілдік механизмнің модельдері. Лингвистикадағы математикалық модельдер. Мәтіндегі ақпаратты таратудың ақпараттық модельдері. Компьютерлік өңдеуге арналған білімді ұсыну (логикалық, желілік, фреймдік, онтологиялық модельдер). Лингвистикалық ресурстар мен жүйелерді жобалау. Мәтінді автоматты өңдеу жүйесін құрудың негізгі кезеңдері. Анализаторларды құру принциптері. Мәтін корпусына қойылатын талаптар. Корпус түрлері, ұйымдастыру принциптері, басқару жүйелері, қолданбалы мақсаттарда пайдалану. Деректер корпусын басқару. Оқу корпустарының нысандары. Корпусты алдынала өңдеу және түрлендіру. Мәтіннің ықтималды сипаттамалары. Жиіліктік сөздіктер. Лингвистикалық модельдер. Көп факторлы лингвомодельдер. Тілдің нейрондық модельдері/ Искусственные языки. Классификация искусственных языков. Языки программирования. Формальные методы описания искусственных языков. Формальная грамматика. Понятие метаязыка. Естественные языки и вычисления. Модели языкового механизма. Математические модели в лингвистике. Информационные модели распределения информации в тексте. Представление знаний для компьютерной обработки (логические, сетевые, фреймвые, онтологические модели). Проектирование лингвистических ресурсов и систем. Основные этапы построения систем автоматической обработки текстов. Принципы построения анализаторов. Требования к корпусу текстов. Виды корпусов, принципы организации, системы управления, использование в прикладных целях. Управление корпусом данных. Объекты чтения корпусов. Предварительная обработка и преобразование корпуса. Вероятностные характеристики текста. Частотные словари. Лингвистические модели. Многофакторность лингвомоделей. Нейронные модели языка / Artificial languages. Classification of artificial languages. Programming languages. Formal methods for describing artificial languages. Formal grammar. The concept of metalanguage. Natural languages and computing. Models of the language mechanism. Mathematical models in linguistics. Information models for the distribution of information in the text. Knowledge representation for computer processing (logical, network, frame, ontological models). Design of linguistic resources and systems. The main stages of building systems for automatic text processing. Principles of constructing analyzers. Requirements for the text corpus. Types of buildings, principles of organization, control systems, use for applied purposes. Data corpus management. Corpus reading objects. Pre-processing and transformation of the hull. Probability characteristics of the text.

Frequency dictionaries. Linguistic models. Multifactoriality of linguistic models. Neural language models

Сабақ түрлері / Виды занятий / Types of lesson: дәріс/ лекция/ lecture; зертханалық жұмыс/ лабораторная работа/ laboratory work

Оқытушы / Преподаватель / Teacher Сатмаганбетова Жанар Зарлыкановна, бағдарламалық қамтамасыз ету кафедрасының аға оқытушысы / Сатмаганбетова Жанар Зарлыкановна, старший преподаватель кафедры программного обеспечения / Zhanar Satmaganbetova, senior lecturer of the Department of Software