Белорусский государственный университет

УТВЕРЖДАЮ
Проректор по учебной работе

_____ А.Л. Толстик
25 мая 2016

Регистрационный № УД-3968 / уч.

ВЫЧИСЛИТЕЛЬНАЯ ЛИНГВИСТИКА

Учебная программа Белорусского государственного университета по учебной дисциплине для специальности:

Направление специальности 1-31 03 07-01 Прикладная информатика (программное обеспечение компьютерных систем)

2016 г.

Учебная программа составлена на основе образовательного стандарта высшего образования ОСВО 1-31 03 07-2013 и учебного плана G31-167/уч., G31и-194/уч., 30.05.2013 1-31 03 07-01 Прикладная информатика (программное обеспечение компьютерных систем)

СОСТАВИТЕЛЬ:

Н.К. Рубашко, старший преподаватель кафедры информационных систем управления Белорусского государственного университета

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой информационных систем управления Белорусского государственного университета (протокол № 12 от 12 мая 2016 г.);

Методической комиссией факультета прикладной математики и информатики Белорусского государственного университета (протокол № 6 от 24 мая 2016 г.).

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Учебная дисциплина специализации «Вычислительная лингвистика» знакомит студентов с основами вычислительной лингвистики и алгоритмами решения классических задач машинной обработки текста. В программу дисциплины включены как классические подходы к компьютерному анализу текста, так и ряд методов, разработанных автором данного курса.

Основной целью дисциплины «Вычислительная лингвистика» является подготовка студентов по следующим направлениям данной предметной области:

разработка лингвистических баз знаний;

технология создания лингвистических процессоров;

применение инструментария вычислительной лингвистики к автоматизации лингвистических исследований.

Целью преподавания дисциплины является научить студентов математическим и алгоритмическим основам обработки естественного языка, научить разрабатывать лингвистические процессоры и решать прикладные задачи, связанные с обработкой естественного языка.

Основные задачи данной дисциплины состоят в подготовке специалистов, умеющих разработать и формализовать постановку задачи, выбрать и оценить алгоритмы, создать программное обеспечение, эксплуатировать и сопровождать разработанные системы, оценивать результаты, определять новые области применения компьютерных систем обработки естественного языка.

Настоящая дисциплина дает студентам базу, необходимую для самостоятельного углубленного изучения вычислительной лингвистики, а также возможность получить знания, необходимые им в дальнейшей профессиональной деятельности в областях, не связанных непосредственно с машинной обработкой текста.

Содержание дисциплины ориентированно на теоретическую и техническую подготовку студентов в рамках создания аппаратно-программных комплексов на базе ЭВМ для решения задач обработки естественного языка в различных областях народного хозяйства Республики Беларусь, связанных с поиском и обработкой информации.

Дисциплина базируется на современных достижениях в области искусственного интеллекта и обработки естественного языка и ориентирована на решение прикладных проблем на основе современных информационных технологий.

Основой для изучения дисциплины «Вычислительная лингвистика» являются дисциплины – дискретная математика, теория вероятностей, теория формальных грамматик, языки программирования, искусственный интеллект.

Требования к профессиональным компетенциям специалиста

Специалист должен быть способен:

Научно-исследовательская деятельность

ПК-1. Работать с научно-технической, нормативно-справочной и

специальной литературой.

ПК-2. Заниматься аналитической и научно-исследовательской деятельностью в области прикладной математики.

ПК-4. Профессионально ставить задачи, вырабатывать идеи и принимать решения.

ПК-5. Владеть современными методами математического моделирования систем и процессов, участвовать в исследованиях новых методов и технологий.

ПК-6. Владеть методами автоматизации научных исследований и применять их в своей работе.

ПК-7. Разрабатывать, анализировать и оптимизировать алгоритмы исследования математических моделей естественнонаучных, производственных и социально-экономических задач.

ПК-8. Разрабатывать, эксплуатировать и сопровождать соответствующие программные компьютерные системы.

Проектно-конструкторская деятельность

ПК-10. Обрабатывать полученные результаты, анализировать их с учетом имеющихся научно-технологических достижений.

ПК-12. Анализировать варианты и находить оптимальные проектные решения.

ПК-13. Обосновывать предложенные решения на современном научно-техническом и профессиональном уровне.

Организационно-управленческая деятельность

ПК-18. Владеть методами и средствами организации работ малых коллективов исполнителей для достижения поставленных целей.

ПК-21. Разрабатывать, представлять и согласовывать необходимые материалы.

ПК-23. Владеть современными средствами телекоммуникаций.

Инновационная деятельность

ПК-27. Разрабатывать бизнес-планы создания новых информационных технологий.

ПК-30. Применять методы анализа и организации внедрения инноваций.

В результате изучения дисциплины обучаемый должен

знать:

– основные задачи, решаемые вычислительной лингвистикой;

– основные составляющие лингвистических баз знаний, принципы их формирования;

– базовые подходы к автоматизации лингвистических исследований в вычислительной лингвистике;

уметь:

– пользоваться основными понятиями прикладной математики и лингвистики, применяемыми в области вычислительной лингвистики;

– реализовывать различные подходы к формированию лингвистических баз знаний;

владеть:

– терминологией вычислительной лингвистики;

– необходимыми знаниями для самостоятельного построения системы автоматизации лингвистических исследований.

В соответствии с учебным планом по специальности 1-31 03 07-01 «Прикладная информатика (программное обеспечение компьютерных систем)» учебная программа предусматривает для изучения дисциплины 68 аудиторных часов, в том числе лекционных – 34 часа, лабораторных – 30 часов и 4 часа управляемой самостоятельной работы. Общее количество часов – 150 часов.

Форма получения высшего образования – дневная (очная). Формы текущей аттестации по учебной дисциплине – экзамен в 5 семестре.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Введение

Введение в вычислительную лингвистику. ИИ- и ИЛ-подходы к решению задач вычислительной лингвистики. Естественный язык как универсальное средство описания действительности и коммуникации с вычислительной системой.

Раздел I. Основные понятия вычислительной лингвистики

Тема 1.1. Особенности естественного языка как объекта моделирования.

Тема 1.2. Формальные языки. Суть, особенности, отличия, примеры.

Раздел II. Инструментарий вычислительной лингвистики

Тема 2.1 Лингвистические базы знаний: состав, технологии.

Раздел III. Базовые элементы лингвистических баз знаний

Тема 3.1. Классификаторы свойств естественного языка.

Тема 3.2. Словари естественного языка. Назначение, состав, правила составления.

Раздел IV. Корпусная лингвистика как составная часть вычислительной лингвистики

Тема 4.1. Понятие текста. Текст как основной источник знаний. Основные единицы текста.

Тема 4.2. Понятие корпуса текстов. Типы, назначение, принципы составления.

Тема 4.3. Аннотированный корпус текстов. АРМ разработчика корпуса текстов.

Раздел V. Лингвистический процессор

Тема 5.1. Понятие базового лингвистического процессора. Функциональность. Этапы анализа текста. Методы и алгоритмы.

Тема 5.2. Марковские модели и машинное обучение. Тестирование базового лингвистического процессора.

Раздел VI. Прикладные задачи вычислительной лингвистики

Тема 6.1. Структурно-функциональная схема системы информационного поиска. Автоматизация индексирования текстовых документов и запросов пользователя.

Тема 6.2. ЕЯ-интерфейс пользователя. Оценка эффективности систем информационного поиска, оптимизация их алгоритмов.

Тема 6.3. Машинный перевод. Основные подходы. Оценка эффективности.

Тема 6.4. Автоматическое реферирование текстов.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

№п/п	Название раздела, темы	Количество часов				Количество часов УСР	Форма контроля знаний
		Аудиторные					
		Лекции	Практ. и сем. занятия	Лаб. занятия	Иное		
	Введение	**2**					
	Введение в вычислительную лингвистику. ИИ- и ИЛ-подходы к решению задач вычислительной лингвистики Естественный язык как универсальное средство описания действительности и коммуникации с вычислительной системой	2					Устная форма
1	**Основные понятия вычислительной лингвистики**	**4**		**2**			
1.1	Особенности естественного языка как объекта моделирования.	2					Устная форма
1.2	Формальные языки. Суть, особенности, отличия, примеры.	2		2			Письменная форма
2	**Инструментарий вычислительной лингвистики**	**2**		**2**		**2**	
2.1	Лингвистические базы знаний: состав, технологии.	2		2		2	Выполнение лабораторных работ на компьютере с последующей устной защитой
3	**Базовые элементы лингвистических баз знаний**	**4**		**6**			
3.1	Классификаторы свойств естественного языка.	2		2			Письменная форма
3.2	Словари естественного языка. Назначение, состав, правила составления.	2		4			Выполнение лабораторных работ на компьютере с последую-

№п/п	Название раздела, темы	Количество часов				Количество часов УСР	Форма контроля знаний
		Аудиторные					
		Лекции	Практ. и сем. занятия	Лаб. занятия	Иное		
							щей устной защитой
4	**Корпусная лингвистика как составная часть вычислительной лингвистики**	**8**		**12**			
4.1	Понятие текста. Текст как основной источник знаний. Основные единицы текста..	2		2			Устная форма
4.2	Понятие корпуса текстов. Типы, назначение, принципы составления..	4		4			Устная форма
4.3	Аннотированный корпус текстов. АРМ разработчика корпуса текстов	2		6			Письменная форма
5	**Лингвистический процессор**	**6**		**4**		**2**	
5.1	Понятие базового лингвистического процессора. Функциональность. Этапы анализа текста. Методы и алгоритмы.	2		2			Устная форма
5.2	Марковские модели и машинное обучение. Тестирование базового лингвистического процессора.	4		2			Письменная форма
6.	Прикладные задачи вычислительной лингвистики	8		4		2	Выполнение лабораторных работ на компьютере с последующей устной защитой
6.1	Структурно-функциональная схема системы информационного поиска. Автоматизация индексирования текстовых документов и запросов пользователя.	2		2			Устная форма

№п/п	Название раздела, темы	Количество часов				Количество часов УСР	Форма контроля знаний
		Аудиторные					
		Лекции	Практ. и сем. занятия	Лаб. занятия	Иное		
6.2	ЕЯ-интерфейс пользователя. Оценка эффективности систем информационного поиска, оптимизация их алгоритмов.	2		2			Выполнение лабораторных работ на компьютере с последующей устной защитой
6.3	Машинный перевод. Основные подходы. Оценка эффективности.	2					Устная форма
6.4	Автоматическое реферирование текстов.	2					Устная форма
ИТОГО		**34**		**30**		**4**	

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Рекомендуемая литература

Основная

1. Ахо, А. Теория синтаксического анализа перевода и компиляции / А. Ахо, Д. Ульман. – М.: Мир, 1978. – Т. 1. – 613 с.
2. Совпель, И.В. Инженерно-лингвистические принципы, методы и алгоритмы автоматическойпереработки текста / И.В. Совпель. –Мн., Вышэйшая школа, 1991.
3. Солтон, Дж.. Динамические библиотечно-информационные системы / Дж. Солтон. –М., Мир, 1997.
4. Joung, S, Bloothooft, G. Corpus-based methods in language and speech processing / S. Joung, G. Bloothooft.– Kluwer academic publishers, 1997.
5. Jurafsky, D. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition / D. Jurafsky, J. H. Martin. – New Jersey: Prentice Hall PTR, 2000. – 934 p.
6. Kornai, A. Extended Finite State Models Of Language / A. Kornai. – UK: Cambridge University Press, 1999. – 291 p.
7. Rozenberg, G., Thomas, W. Developments in language theory / G. Rozenberg, W. Thomas. – World scientific publishing, 2000.

Дополнительная

1. Рубашко, Н. К. Разработка лексико-грамматических классификаторов русского и белорусского языков и их применение / Н. К. Рубашко // Вестник БГУ, серия "Физика, математика, информатика". – 2007. – No 5. – С. 14–18.
2. Fastus: A cascaded finite-state transducer for extracting information from natural-language text / D. Israel [et al.] // Finite State Devices for Natural Language Processing / ed. by Roche, Schabes. – Cambridge, MA, USA: MIT Press, 1996. – P. 383–406.
3. Mohri, M. Finite-state transducers in language and speech processing / M. Mohri // Computational Linguistics. – 1997. – Vol. 23, no. 2. – P. 269–311.
4. Roche, E. Introduction to finite-state devices in natural language processing. – Technical Report, Mitsubishi Electric Research Laboratories. – 1996.
5. Johansson, S. – The Tagged LOB Corpus: Users’ manual. – The Norwegian Computing Centre for the Humanities, Bergen University, Norway, 1986.
6. Чеусов, А. В. Разработка лингвистических процессоров промышленной обработки текстовых документов / А. В. Чеусов // Искусственный интеллект. – 2006. – No 4. – С. 635–639.

Организация управляемой самостоятельной работы (УСР) студентов

Самостоятельная работа студентов – это любая деятельность, связанная с воспитанием мышления будущего профессионала. В широком смысле под самостоятельной работой следует понимать совокупность всей самостоятельной деятельности студентов как в учебной аудитории, так и за её пределами, в контакте с преподавателем и в его отсутствии.

Самостоятельная работа реализуется:

1. Непосредственно в процессе аудиторных занятий – на лекциях, практических и семинарских занятиях, при выполнении контрольных работ.

2. В контакте с преподавателем вне рамок расписания – на консультациях по учебным вопросам, в ходе творческих контактов, при ликвидации задолженностей, при выполнении индивидуальных заданий и т.д.

3. В библиотеке, дома, в общежитии, на кафедре при решении студентом учебных и творческих задач.

При изучении дисциплины организация самостоятельной работы студентов должна представлять единство трех взаимосвязанных форм:

1. Внеаудиторная самостоятельная работа;

2. Аудиторная самостоятельная работа, которая осуществляется под непосредственным руководством преподавателя;

3. Творческая, в том числе научно-исследовательская работа.

Виды внеаудиторной самостоятельной работы студентов разнообразны: подготовка и написание рефератов, докладов, очерков и других письменных работ на заданные темы.

Аудиторная самостоятельная работа может реализовываться при проведении практических занятий, семинаров, выполнении лабораторного практикума и во время чтения лекций.

При чтении лекционного курса непосредственно в аудитории необходимо контролировать усвоение материала основной массой студентов путем проведения экспресс-опросов по конкретным темам.

На практических занятиях различные виды самостоятельной работы студентов позволяют сделать процесс обучения более интересным и поднять активность значительной части студентов в группе.

На лабораторных занятиях нужно не менее 1 часа из двух (50% времени) отводить на самостоятельное решение задач. Практические занятия целесообразно строить следующим образом: 1. Вводное слово преподавателя (цели занятия, основные вопросы, которые должны быть рассмотрены). 2. Беглый опрос. 3. Решение 1-2 типовых задач. 4. Самостоятельное решение задач. 5. Разбор типовых ошибок при решении (в конце текущего занятия или в начале следующего).

Результативность самостоятельной работы студентов во многом определяется наличием активных методов ее контроля. Существуют следующие виды контроля:

– входной контроль знаний и умений студентов в начале изучения очередной дисциплины;

– текущий контроль, то есть регулярное отслеживание уровня усвоения материала на лекциях, практических и лабораторных занятиях;

– промежуточный контроль по окончании изучения раздела или модуля курса;

– самоконтроль, осуществляемый студентом в процессе изучения дисциплины при подготовке к контрольным мероприятиям;

– итоговый контроль по дисциплине в виде зачета или экзамена;

– контроль остаточных знаний и умений спустя определенное время после завершения изучения дисциплины.

Примерный перечень заданий УСР

Формирование исходных корпусов текстов естественного языка (по выбору).

Получение базовых словарей.

Разработка классификаторов свойств естественного языка.

Аннотирование базовых словарей.

Аннотирование корпусов текстов.

Снятие статистики.

Разработка инструментария для аннотирования и постобработки

Рекомендации по контролю качества усвоения знаний и проведению аттестации

Перечень рекомендуемых форм диагностики компетенций

Для аттестации обучающихся на соответствие их персональных достижений поэтапным и конечным требованиям образовательной программы создаются фонды оценочных средств, включающие типовые задания, контрольные работы и тесты. Оценочными средствами предусматривается оценка способности обучающихся к творческой деятельности, их готовность вести поиск решения новых задач, связанных с недостаточностью конкретных специальных знаний и отсутствием общепринятых алгоритмов.

Для диагностики компетенций в рамках учебной дисциплины рекомендуется использовать следующие формы:

1. устная форма: собеседования, устные промежуточные и итоговый зачеты;
2. письменная форма: тесты, контрольные опросы, контрольная работа;
3. устно-письменная форма: отчеты по домашним практическим упражнениям с их устной защитой;
4. . Выполнение лабораторных работ на компьютере с последующей устной защитой.

Контрольные мероприятия проводятся в соответствии с учебно-методической картой дисциплины. В случае неявки на контрольное мероприя-

тие по уважительной причине студент вправе по согласованию с преподавателем выполнить его в дополнительное время. Для студентов, получивших неудовлетворительные оценки за контрольные мероприятия, либо не явившихся по неуважительной причине, по согласованию с преподавателем и с разрешения заведующего кафедрой мероприятие может быть проведено повторно.

Оценка текущей успеваемости рассчитывается как среднее оценок за каждую из письменных контрольных работ, оценки за отчеты по домашним практическим упражнениям и оценки за итоговый тест.

Итоговая аттестация предусматривает проведение экзамена. При этом рекомендуется использовать оценивание успеваемости на основе модульно-рейтинговой системы.

Рекомендации по контролю качества усвоения знаний и проведению аттестации

На лекционных занятиях по дисциплине «Вычислительная лингвистика» возможно использование элементов проблемного обучения: проблемное изложение некоторых аспектов, поисковая беседа (частично-поисковый метод).

На лабораторных занятиях по дисциплине рекомендуется использовать такие приемы преподавания, как сопоставление с новыми фактами, анализ известных фактов, управление исследовательской деятельностью, а также следующие приемы учения: исследование проблемы, самостоятельное выдвижение гипотезы по решению задачи, соотнесение полученных, результатов с выдвинутым предположением, обобщение по проблеме в целом.

Перечень рекомендуемых форм диагностики компетенций

Для аттестации обучающихся на соответствие их персональных достижений поэтапным и конечным требованиям образовательной программы создаются фонды оценочных средств, включающие типовые задания, контрольные работы и тесты. Оценочными средствами предусматривается оценка способности обучающихся к творческой деятельности, их готовность вести поиск решения новых задач, связанных с недостаточностью конкретных специальных знаний и отсутствием общепринятых алгоритмов.

Для диагностики компетенций в рамках учебной дисциплины рекомендуется использовать следующие формы:

1. Устная форма: собеседования, промежуточные и итоговые зачеты.
2. Письменная форма: тесты, контрольные опросы, контрольная работа.
3. Устно-письменная форма: отчеты по домашним практическим упражнениям с их устной защитой.
4. Выполнение лабораторных работ на компьютере с последующей устной защитой.

Контрольные мероприятия проводятся в соответствии с учебно-методической картой дисциплины. В случае неявки на контрольное мероприятие по уважительной причине студент вправе по согласованию с преподавате-

лем выполнить его в дополнительное время. Для студентов, получивших неудовлетворительные оценки за контрольные мероприятия, либо не явившихся по неуважительной причине, по согласованию с преподавателем и с разрешения заведующего кафедрой мероприятие может быть проведено повторно.

Оценка текущей успеваемости рассчитывается как среднее оценок за каждую из письменных контрольных работ, оценки за отчеты по домашним практическим упражнениям, лабораторным работам и оценки за итоговый тест.

Текущая аттестация предусматривает проведение экзамена. При этом рекомендуется использовать оценивание успеваемости на основе модульно-рейтинговой системы.

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ

Название учебной дисциплины, с которой требуется согласование	Название кафедры	Предложения об изменениях в содержании учебной программы учреждения высшего образования по учебной дисциплине	Решение, принятое кафедрой, разработавшей учебную программу (с указанием даты и номера протокола)
Программирование	Кафедра информационных систем управления	нет	Оставить содержание учебной дисциплины без изменения, протокол № 12 от 12 мая 2016 г.
Дискретная математика	Кафедра дискретной математики и алгоритмизации		Оставить содержание учебной дисциплины без изменения, протокол № 12 от 12 мая 2016 г.

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ

на _____/_____ учебный год

№№ Пп	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры информационных систем управления (протокол № ____ от _____ 201_ г.)

Заведующий кафедрой

_____ _____ _____

(ученая степень, звание) (подпись) (И.О. Фамилия)

УТВЕРЖДАЮ
Декан факультета

_____ _____ _____

(ученая степень, звание) (подпись) (И.О.Фамилия)