


Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

САЕ «Институт человек цифровой эпохи».

УТВЕРЖДАЮ:
Руководитель ОПОП:


З.И. Резанова
« 31 » августа 20 22 г.

Рабочая программа дисциплины

Введение в анализ естественного языка (NLP)

по направлению подготовки

САЕ «Институт человек цифровой эпохи».

гика

Компьютерная и когнитивная лингвистика

Форма обучения

Очная

Квалификация

Магистр

Год приема

2022

Код дисциплины в учебном плане: Б1.В.ДВ.1.1.1

СОГЛАСОВАНО:

Руководитель ОП

З.И. Резанова

Председатель УМК

Ю.А. Тихомирова

Томск – 2022

1. Цель и планируемые результаты освоения дисциплины

Целью освоения дисциплины является формирование следующих компетенций:

ОПК-3 Способен выбирать оптимальные подходы и методы решения конкретных научных и прикладных задач в области лингвистики и информационных технологий

ОПК-4 Способен расширять сферу научной деятельности, участвовать в междисциплинарных исследованиях на стыке наук

ОПК-6 Способен осуществлять эффективное управление разработкой программных средств и информационных проектов в сфере своей профессиональной деятельности

ПК-1 Способен проводить самостоятельные исследования и получать новые научные результаты в области междисциплинарных лингвистических исследований

ПК-4 Способность разрабатывать проекты прикладной направленности в области когнитивной и компьютерной лингвистики с применением современных технических средств и информационных технологий, в том числе в области искусственного интеллекта

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИПК-4.2 Разрабатывает программу действий по решению задач проекта в области когнитивной и компьютерной лингвистики с учетом имеющихся технических средств и информационных технологий, в том числе в области искусственного интеллекта..

ИПК-4.1 Формулирует цель проекта прикладной направленности в области когнитивной и компьютерной лингвистики, обосновывает необходимость применения современных технических средств и информационных технологий, в том числе в области искусственного интеллекта.

ИПК-1.1 Обнаруживает знания об актуальных направлениях междисциплинарных лингвистических исследований в избранной научной сфере.

ИОПК-6.1 Аргументированно выбирает математические и лингвистические методы решения профессиональных задач с применением языков программирования.

ИОПК-4.1 Демонстрирует знание новых теорий в сфере междисциплинарного взаимодействия лингвистики и наук гуманитарного, математического и естественно-научного циклов.

ИОПК-3.2 Критически сопоставляет и оценивает существующие подходы и методы решения конкретных научных и прикладных задач в области лингвистики и информационных технологий.

2. Задачи освоения дисциплины

– получение студентом знаний об основных теоретических и прикладных направлениях науки в области компьютерной лингвистики, основные понятия в области компьютерной обработки естественного языка и методов машинного обучения, применяемых в данной области.

3. Место дисциплины в структуре образовательной программы

Дисциплина относится к части образовательной программы, формируемой участниками образовательных отношений, предлагается обучающимся на выбор. Дисциплина входит в модуль Компьютерная лингвистика.

4. Семестр(ы) освоения и форма(ы) промежуточной аттестации по дисциплине

Первый семестр, экзамен

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются компетенции, сформированные в ходе освоения образовательных программ предшествующего уровня образования.

6. Язык реализации

Русский

7. Объем дисциплины

Общая трудоемкость дисциплины составляет 3 з.е., 108 часов, из которых:

-лекции: 6 ч.

-практические занятия: 20 ч.

Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины, структурированное по темам

Тема 1. Введение в проблематику обработки языка. Теоретические проблемы, соотношение NLP и лингвистики, основные источники. N-граммы и моделирование языка. Векторная семантика

Тема 2. Современное состояние рынка вакансий в области NLP. Вакансии в области обработки языковых данных: требования и зарплаты

Тема 3. Машинное обучение в применении к обработке естественного языка. Naïve Bayes - наивный байесовский классификатор. k-nearest neighbor, k-means clustering - k ближайших соседей, кластеризация по k-средним. Hidden Markov Models - скрытые марковские модели. Decision trees, Random Forest - дерево решений, случайный лес. Support Vector Machine - метод опорных векторов. Principal Component Analysis - метод главных компонент. Artificial Neural Networks - нейронные сети. Logistic Regression, можно больше о регрессии. Gradient Boosting и подобные

Тема 4. Сентимент-анализ. Современное состояние поля, реализации Bag-of-Words и более продвинутых подходов. Решения в области, проприетарные и открытые; продукты для мониторинга соцсетей; использование метаданных

Тема 5. Виртуальные ассистенты и чат-боты

Тема 6. Машинный перевод: от истоков к трансформерам, не забывая про конечного пользователя

Тема 7. Speech Recognition: фонетическая структура языка, акустика речи, восприятие и производство речи. Машинное обучение и распознавание речи.

Тема 8. Этика искусственного интеллекта. Этика искусственного интеллекта: проблемы, подходы, решения. Скандальный случай из области больших данных: краткий доклад

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине проводится путем контроля посещаемости, проведения контрольных работ, выполнения домашних заданий и фиксируется в форме контрольной точки не менее одного раза в семестр.

Анализ и оценивание существующих исследований на предмет охвата материала, качества анализа и теоретических оснований. Представление критического анализа в устном выступлении на семинаре.

Написание эссе на темы, предложенные преподавателем

Планирование исследования: студенты получают практическое задание в области компьютерной лингвистики, например, «Чат-бот «Психолог» для студентов университета»

и должны разработать план выполнения с конкретными источниками литературы, аналогичными проектами и шагами реализации.

Подготовка аннотированной биографии (10-15 источников) к одной из тем семинаров.

Индивидуальный/парный/групповой семинар: студентам предъявляется тема для мозгового штурма, например, «Какое приложение может иметь теория функциональной грамматики в области семантических репрезентаций?». Дается пять минут на индивидуальную подготовку, затем студенты объединяются в пары и выбирают три лучших идеи (следующие пять минут). После этого группа выбирает 3 лучших идеи и представляет их преподавателю.

Задания на определение ключевых понятий различных областей NLP.

Лабораторная работа по заполнению словаря, программированию словарных статей, корпусов различных типов, правке онто разметки, извлечению знаний из интеллектуальной информационной системы, распознаванию речи с помощью доступных программных платформ.

Написанию отчета по сравнению результатов работы с различными системами, написанию заявки в техническую поддержку интеллектуальной информационной системы; написанию предложений по улучшению и расширению функционала существующих систем.

10. Порядок проведения и критерии оценивания промежуточной аттестации

Экзамен в первом семестре проводится устной форме. Экзаменационное задание состоит из трех частей. Продолжительность экзамена 1,5 часа.

Промежуточная аттестация и ее оценивание осуществляются следующим образом.

Экзамен состоит в выполнении практического задания по позиционированию своего магистерского исследования в рамках поля NLP. Работа выполняется письменно и оформляется согласно требованиям, предъявляемым к курсовым работам. Объем работы – до 15 стр. А4. Защита работы происходит на экзамене в форме презентации и ответа на вопросы экзаменатора и присутствующих.

Работа и презентация может быть оценена, если она соответствует следующим критериям.

Работа демонстрирует наличие у студента знаний о поле NLP, основных и специальных методах и приемах анализа естественного языка. Магистрант должен проявить способность позиционировать свою работу в контексте современной практики обработки естественного языка. Работа должна излагать суть планируемого проекта магистранта и его место в современной парадигме NLP. Приветствуется наличие детализированного плана разработки в соотношении с необходимыми для его реализации технологиями.

Работа должна отвечать специфике научного стиля, быть ясно изложена. Задание, выполненное студентом, должно демонстрировать знание принципов создания и презентации научного текста в форме плана проекта, реферата, плана исследования, научного эссе, их структурирования и оформления. Особое внимание уделяется корректности оформления текста, иллюстраций и ссылок по существующему ГОСТу.

Требования к презентации:

1. успешное удержание внимания на речи и презентации докладчика;
2. адекватное оформление презентационных материалов по времени, дизайну и структуре;
3. разделение поданной информации на главную и второстепенную;
4. умение адаптировать презентацию к нуждам аудитории;
5. умение корректно (в рамках научной дискуссии) отвечать на поставленные вопросы аудитории и сделанные замечания.

Оценка «удовлетворительно» может быть поставлена при выполнении 60 процентов самостоятельной работы в семестре, выполнении проектной работы, фрагментарно описывающей магистерскую проектную работу, отражающей недостаточное знание поля NLP, технологий обработки языка и современного состояния субдисциплин. Презентация работы выполнена с нарушением критериев, ответы на вопросы даны некорректно или не по существу.

Оценка «хорошо» может быть поставлена при выполнении 80 процентов самостоятельной работы в семестре, выполнении проектной работы, в целом описывающей магистерскую проектную работу, отражающей знание поля NLP, основных технологий обработки языка и современного состояния субдисциплин. Презентация работы выполнена согласно критериям, возможны недочеты в изложении содержания работы, затруднения при ответах на вопросы.

Оценка «отлично» может быть поставлена при выполнении 95-100 процентов самостоятельной работы в семестре, выполнении проектной работы, исчерпывающе описывающей магистерскую проектную работу, отражающей полноценное знание поля NLP, технологий обработки языка и современного состояния субдисциплин. Презентация полностью отвечает всем критериям, ответы на вопросы экзаменатора и присутствующих демонстрируют знание специфики релевантной области NLP.

Первая часть работы проверяет компетенции ИПК-4.2, ИПК-4.1, ИПК-1.1. Вторая часть работы проверяет компетенции ИОПК-6.1, ИОПК-4.1, ИОПК-3.2.

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в электронном университете «Moodle» - <https://moodle.tsu.ru/course/view.php?id=14711>

б) Оценочные материалы текущего контроля и промежуточной аттестации по дисциплине представлены в курсе «Moodle»

в) План практических занятий по дисциплине соответствует п. 8 (лекции не предусмотрены).

г) Методические указания по организации самостоятельной работы студентов.

Самостоятельная работа студентов включает чтение текстов из списка литературы, подготовку к лекциям по технологии перевернутого класса. Подробные методические указания представлены в курсе «Moodle» - <https://moodle.tsu.ru/mod/page/view.php?id=199720>.

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

1. Handbook of Natural Language Processing. / Eds. Nitin Indurkha, Fred J. Damerau. – 2nd ed. — Chapman & Hall/CRC, 2010. – 692 p.

2. Jurafsky D., Martin J.H. Speech and Language Processing (3rd ed. draft). – URL: <https://web.stanford.edu/~jurafsky/slp3/>

3. The handbook of computational linguistics and natural language processing / Eds Alexander Clark, Chris Fox, Shalom Lappin. – Wiley-Blackwell, 2010. – 801 p.

4. Rogers, Simon, and Mark Girolami. A first course in machine learning. Chapman and Hall/CRC, 2016.

5. Гласснер Э. Глубокое обучение без математики. Т. 1: Основы / пер. с англ. В. А. Яроцкого. – М.: ДМК Пресс, 2019. – 584 с.: ил

6. Liu, Bing. Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge university press, 2020.

7. Danneman, Nathan, and Richard Heimann. Social media mining with R. Packt Publishing Ltd, 2014.

8. 97 Things Every Data Engineer Should Know. Tobias Macey 2021 O'Reilly Media, Inc.. 256 p.

б) дополнительная литература:

1. Molnar, Christoph. Interpretable machine learning. Lulu. com, 2020.
2. Coeckelbergh, Mark. AI ethics. MIT Press, 2020.
3. Franks, Bill. 97 Things About Ethics Everyone in Data Science Should Know. O'Reilly Media, 2020.

в) ресурсы сети Интернет:

1. ACL Anthology. URL: <http://aclweb.org/anthology/>, <http://aclanthology.info/>
2. Behavior Research Methods. URL: <https://link.springer.com/journal/13428>
3. Computational Linguistics. URL: <http://www.mitpressjournals.org/loi/coli>
4. Computer Speech and Language. URL: <https://www.journals.elsevier.com/computer-speech-and-language>
5. International Journal of Corpus Linguistics. URL: <https://benjamins.com/#catalog/journals/ijcl/main>
6. Journal of Information Retrieval. URL: <http://www.springer.com/computer/database+management+%26+information+retrieval/journal/10791>
7. Journal of Machine Learning. URL: <http://www.springer.com/computer/ai/journal/10994>
8. Language and Linguistics Compass. URL: [http://onlinelibrary.wiley.com/journal/10.1111/\(ISSN\)1749-818X](http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1749-818X)
9. Language Resources and Evaluation. URL: <https://link.springer.com/journal/10579>
10. Machine Translation. URL: <https://link.springer.com/journal/10590>
11. Natural Language Semantics. URL: <https://link.springer.com/journal/11050>
12. Transactions of the Association for Computational Linguistics. URL: <https://www.transacl.org/ojs/index.php/tacl/issue/view/13>
13. Диалог. URL: <http://www.dialog-21.ru/digest/>

13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:

– Microsoft Office 2019 Russian: пакет программ. Включает приложения: MS Office Word, MS Office Excel, MS Office PowerPoint, MS Office OneNote, MS Office Publisher, MS Outlook, MS Office Web Apps (Word Excel MS PowerPoint Outlook);

– публично доступные облачные технологии (Google Docs, Яндекс диск и т.п.).

б) информационные справочные системы:

– Электронный каталог Научной библиотеки ТГУ – <http://chamo.lib.tsu.ru/search/query?locale=ru&theme=system>

– Электронная библиотека (репозиторий) ТГУ – <http://vital.lib.tsu.ru/vital/access/manager/Index>

– ЭБС Лань – <http://e.lanbook.com/>

– ЭБС Консультант студента – <http://www.studentlibrary.ru/>

– Образовательная платформа Юрайт – <https://urait.ru/>

– ЭБС ZNANIUM.com – <https://znanium.com/>

– ЭБС IPRbooks – <http://www.iprbookshop.ru/>

14. Материально-техническое обеспечение

Аудитории для проведения занятий лекционного типа.

Аудитории для проведения занятий семинарского типа, индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации.

Помещения для самостоятельной работы, оснащенные компьютерной техникой и доступом к сети Интернет, в электронную информационно-образовательную среду и к информационным справочным системам.

15. Информация о разработчиках

Шиляев Константин Сергеевич, к. филол. н., доцент, кафедра общей, компьютерной и когнитивной лингвистики