МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
"Казанский (Приволжский) федеральный университет"
Институт филологии и межкультурной коммуникации
Высшая школа зарубежной филологии и межкультурной коммуникации им. И.А. Бодуэна де Куртенэ

УТВЕРЖДАЮ
Проректор по образовательной деятельности КФУ
Е.А. Турилова
26 февраля 2024 г.

подписано электронно-цифровой подписью

Программа дисциплины

Обработка естественных языков (NLP)

Направление подготовки: 45.03.01 - Филология
Профиль подготовки: Прикладная филология: иностранный (английский) язык в международной коммуникации
Квалификация выпускника: бакалавр
Форма обучения: очное
Язык обучения: русский
Год начала обучения по образовательной программе: 2024

Содержание

1. Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения ОПОП ВО

2. Место дисциплины (модуля) в структуре ОПОП ВО

3. Объем дисциплины (модуля) в зачетных единицах с указанием количества часов, выделенных на контактную работу обучающихся с преподавателем (по видам учебных занятий) и на самостоятельную работу обучающихся

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Структура и тематический план контактной и самостоятельной работы по дисциплине (модулю)

4.2. Содержание дисциплины (модуля)

5. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине (модулю)

6. Фонд оценочных средств по дисциплине (модулю)

7. Перечень литературы, необходимой для освоения дисциплины (модуля)

8. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

9. Методические указания для обучающихся по освоению дисциплины (модуля)

10. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень программного обеспечения и информационных справочных систем (при необходимости)

11. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

12. Средства адаптации преподавания дисциплины (модуля) к потребностям обучающихся инвалидов и лиц с ограниченными возможностями здоровья

13. Приложение №1. Фонд оценочных средств

14. Приложение №2. Перечень литературы, необходимой для освоения дисциплины (модуля)

15. Приложение №3. Перечень информационных технологий, используемых для освоения дисциплины (модуля), включая перечень программного обеспечения и информационных справочных систем

Программу дисциплины разработал(а)(и): доцент, к.н. (доцент) Варламова Е.В. (кафедра романо-германской филологии, Высшая школа зарубежной филологии и межкультурной коммуникации им И А Бодуэна де Куртенэ), Elena.Varlamova@kpfu.ru

1. Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения ОПОП ВО

Обучающийся, освоивший дисциплину (модуль), должен обладать следующими компетенциями:

Шифр компетенции	Расшифровка приобретаемой компетенции
ПК-2	Способен проводить научные исследования в конкретной области филологического знания с формулировкой аргументированных умозаключений и выводов
ПК-7	Способность эффективно применять базовые математические знания и информационные технологии при решении прикладных задач
УК-4	Способен осуществлять деловую коммуникацию в устной и письменной формах на государственном языке Российской федерации и иностранном(ых) языке(ах)

Обучающийся, освоивший дисциплину (модуль):

Должен знать:

- современную теоретическую концепцию культуры речи, орфоэпические, акцентологические, грамматические, лексические нормы русского литературного языка: коммуникативные методы и технологии на государственном и иностранном языках;

- информационно-коммуникационные технологии для поиска необходимой информации ;пути обмена информацией, их значение в работе образовательного учреждения, значение планирования в профессиональной деятельности;

- основы стилистики официальных и неофициальных писем на государственном и иностранном(ых) языках; правила и закономерности деловой устной и письменной коммуникации;

- существующие методы и приемы анализа и интерпретации языковых (художественных) текстов, перспективы его использования в различных областях науки и культуры;

- методику информационно-словарного описания;

- современные информационные технологии и программные средства для осуществления профессиональной деятельности, принципы их работы;

- принципы, методы и средства решения стандартных задач профессиональной деятельности на основе информационной и библиографической культуры с применением информационно- коммуникационных технологий и с учетом основных требований информационной безопасности.

Должен уметь:

- эффективно находить, воспринимать и использовать информацию на государственном и иностранном языках, полученную из печатных и электронных источников для решения стандартных коммуникативных задач в профессиональной деятельности; использовать государственный и иностранный язык в профессиональной деятельности;

- самостоятельно решать творческие задачи профессиональной деятельности с использованием информационных, библиографических ресурсов;

- использовать не менее 900 терминологических единиц и терминоэлементов на государственном и иностранном(ых) языках для ведения деловой переписки;

- представлять результаты собственного исследования в виде письменных жанров научной коммуникации, размещать материалы собственных исследований в информационных сетях;

- самостоятельно использовать созданные тексты официально-делового, научного, художественного стилей в соответствии с ситуацией общения и поставленными задачами;

- применять современные информационные технологии и программные средства при решении задач профессиональной деятельности,

программные средства обеспечения безопасности данных на автономном ПК и в интерактивной среде;

- решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности.

Должен владеть:

- способами критической оценки эффективности различных коммуникативных методов и технологии на государственном и иностранном языках;

техникой деловой речевой коммуникации, опираясь на современное состояние языковой культуры;

- навыками извлечения необходимой информации из оригинального текста на иностранном языке по профессиональной проблематике с помощью информационно-коммуникационных технологий;

- навыками ведения деловой переписки с учетом особенностей стилистики официальных и неофициальных писем, социокультурных различий в формате корреспонденции на государственном и иностранном(ых) языках; навыками устного делового разговора на государственном и иностранном(-ых) языках;

- навыками обобщения материала научного исследования;

- навыками ведения письменной и устной коммуникации посредством создания текстов различных типов, жанров и стилей: официально-делового, научного, художественного;

- широким диапазоном различных информационно-коммуникационных технологий при решении задач профессиональной деятельности;

- навыками применения различных ИКТ для решения задач профессиональной деятельности.

2. Место дисциплины (модуля) в структуре ОПОП ВО

Данная дисциплина (модуль) включена в раздел "Б1.В.21 Дисциплины (модули)" основной профессиональной образовательной программы 45.03.01 "Филология (Прикладная филология: иностранный (английский) язык в международной коммуникации)" и относится к части ОПОП ВО, формируемой участниками образовательных отношений.

Осваивается на 4 курсе в 7, 8 семестрах.

3. Объем дисциплины (модуля) в зачетных единицах с указанием количества часов, выделенных на контактную работу обучающихся с преподавателем (по видам учебных занятий) и на самостоятельную работу обучающихся

Общая трудоемкость дисциплины составляет 3 зачетных(ые) единиц(ы) на 108 часа(ов).

Контактная работа - 57 часа(ов), в том числе лекции - 20 часа(ов), практические занятия - 36 часа(ов), лабораторные работы - 0 часа(ов), контроль самостоятельной работы - 1 часа(ов).

Самостоятельная работа - 24 часа(ов).

Контроль (зачёт / экзамен) - 27 часа(ов).

Форма промежуточного контроля дисциплины: отсутствует в 7 семестре; экзамен в 8 семестре.

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1 Структура и тематический план контактной и самостоятельной работы по дисциплине (модулю)

N	Разделы дисциплины / модуля	Семестр	Виды и часы контактной работы, их трудоемкость (в часах)						Самостоятельная работа
			Лекции, всего	Лекции в эл. форме	Практические занятия, всего	Практические в эл. форме	Лабораторные работы, всего	Лабораторные в эл. форме	
1.	Тема 1. Основные задачи и методы компьютерной лингвистики	7	4	0	1	0	0	0	4
2.	Тема 2. Предварительная обработка текста (токенизация, регулярные выражения)	7	2	0	2	0	0	0	2
3.	Тема 3. Частотность слов, мешок слов, tf.idf	7	2	0	1	0	0	0	2

N	Разделы дисциплины / модуля	Се-местр	Виды и часы контактной работы, их трудоемкость (в часах)						Само-стоя-тель-ная ра-бота
			Лекции, всего	Лекции в эл. форме	Практи-ческие занятия, всего	Практи-ческие в эл. форме	Лабора-торные работы, всего	Лабора-торные в эл. форме	
4.	Тема 4. Статистичсекие языковые модели и марковские цепи	7	2	0	2	0	0	0	2
5.	Тема 5. Автоматический морфологический анализ	7	2	0	1	0	0	0	2
6.	Тема 6. Введение в синтаксический анализ	7	2	0	2	0	0	0	2
7.	Тема 7. Выделение ключевых слов и устойчивых словосочетаний	7	2	0	1	0	0	0	2
8.	Тема 8. Извлечение именованных сущностей	7	2	0	2	0	0	0	2
9.	Тема 9. Инструменты извлечения информации из текста	7	2	0	2	0	0	0	2
10.	Тема 10. Основы машинного обучения	8	0	0	2	0	0	0	1
11.	Тема 11. Обучение с учителем: классификация	8	0	0	2	0	0	0	1
12.	Тема 12. Обучение с учителем: регрессия	8	0	0	2	0	0	0	1
13.	Тема 13. Оценка модели и валидация	8	0	0	2	0	0	0	1
14.	Тема 14. Обучение без учителя: кластеризация	8	0	0	2	0	0	0	
15.	Тема 15. Обучение без учителя: PCA, SVD, t-SNE, UMAP	8	0	0	2	0	0	0	
16.	Тема 16. Ансамбли	8	0	0	2	0	0	0	
17.	Тема 17. Введение в нейронные сети	8	0	0	2	0	0	0	
18.	Тема 18. Нейронные архитектуры CNN, RNN, LSTM	8	0	0	2	0	0	0	
19.	Тема 19. Векторная семантика	8	0	0	2	0	0	0	
20.	Тема 20. Модели-трансформеры	8	0	0	2	0	0	0	
	Итого		20	0	36	0	0	0	24

4.2 Содержание дисциплины (модуля)

Тема 1. Основные задачи и методы компьютерной лингвистики

Направления компьютерной лингвистики

Обработка естественного языка (англ. natural language processing). Уровни обработки и анализа текста: синтаксический, морфологический, семантический.

К задачам и направлениям компьютерной лингвистики относят:

Корпусную лингвистику, создание и использование электронных корпусов текстов.

Создание электронных словарей, тезаурусов, онтологий. Например, Lingvo. Словари используют, например, для автоматического перевода, проверки орфографии.

Автоматический перевод текстов. Среди русских переводчиков популярным является Промт. Среди бесплатных известен переводчик Google Translate.

Автоматическое извлечение фактов из текста (извлечение информации) (англ. fact extraction, text mining)

Автореферирование (англ. automatic text summarization). Эта функция включена, например, в Microsoft Word.

Построение систем управления знаниями. См. Экспертные системы.

Создание вопросно-ответных систем (англ. question answering systems).

Оптическое распознавание символов (англ. OCR). Например, с помощью программы FineReader

Автоматическое распознавание речи (англ. ASR).

Автоматический синтез речи.

Тема 2. Предварительная обработка текста (токенизация, регулярные выражения)

Предобработка текста переводит текст на естественном языке в формат удобный для дальнейшей работы. Предобработка состоит из различных этапов, которые могут отличаться в зависимости от задачи и реализации. Далее приведен один из возможных набор этапов:

Перевод всех букв в тексте в нижний или верхний регистры;

Удаление цифр (чисел) или замена на текстовый эквивалент (обычно используются регулярные выражения);

Удаление пунктуации. Обычно реализуется как удаление из текста символов из заранее заданного набора;

Удаление пробельных символов (whitespaces);

Токенизация (обычно реализуется на основе регулярных выражений);

Удаление стоп слов;

Стемминг;

Лемматизация;

Векторизация.

Тема 3. Частотность слов, мешок слов, tf.idf

Принцип метода мешка слов (Bag of Words, BoW) чрезвычайно прост. Мы считаем как часто встречается каждое слово в тексте.

Несмотря на простоту, при правильной предобработке текста (в первую очередь удалении стоп-слов, которые и будут наиболее частотными) этот метод показывает неплохие результаты.

Чуть более сложный и продвинутый метод определения значимости слов в тексте называется TF-IDF (term frequency - inverse document frequency).

Основная идея

Если слово часто встречается во всех документах (это в первую очередь касается предлогов, союзов и других стоп-слов), то вряд ли эти слова имеют большое значение. И наоборот, если слово встречаться только в одном документе, вероятно оно в большей степени определяет его содержание.

Другими словами, определяется не только значимость слова в тексте, но и значимость слова с учётом всех текстов.

Тема 4. Статистичсекие языковые модели и марковские цепи

N-граммы - это статистические модели, которые предсказывают следующее слово после N-1 слов на основе вероятности их сочетания. Например, сочетание I want to в английском языке имеет высокую вероятностью, a want I to - низкую. Говоря простым языком, N-грамма - это последовательность n слов. Например, биграммы - это последовательности из двух слов (I want, want to, to, go, go to, to the...), триграммы - последовательности из трех слов (I want to, want to go, to go to...) и так далее.

Такие распределения вероятностей имеют широкое применение в машинном переводе, автоматической проверке орфографии, распознавании речи и умном вводе.

Тема 5. Автоматический морфологический анализ

Существуют программы, которые могут проводить морфологический анализ автоматически, - они называются морфологическими анализаторами. Внутреннее устройство анализатора зависит от языка, для которого он предназначен, так как набор частей речи и их грамматических категорий отличается от языка к языку. Наиболее популярными анализаторами для русского языка являются библиотека pymorphy2 для Python и программа mystem, работающие на основе словаря и системы правил.

Тема 6. Введение в синтаксический анализ

Синтакси́ческий ана́лиз (или разбор, жарг. па́рсинг ← англ. parsing) в лингвистике и информатике - процесс сопоставления линейной последовательности лексем (слов, токенов) естественного или формального языка с его формальной грамматикой. Результатом обычно является дерево разбора (синтаксическое дерево). Обычно применяется совместно с лексическим анализом.

Синтаксический анализатор (жарг. па́рсер ← англ. parser) - это программа или часть программы, выполняющая синтаксический анализ.

Тема 7. Выделение ключевых слов и устойчивых словосочетаний

Предобработка данных осуществлялась с помощью языка программирования python. На этапе предобработки с помощью регулярных выражений данные почистили от пунктуации html-тэгов, а также удалили слова, состоящие из одного буквы. Токенизация и удаление стоп-слов было осуществлено с помощью библиотека nltk. Для приведения слов к единой форме была использована лемматизация из nltk.stem.WordNetLemmatizer() и pymorphy2.MorphAnalyzer(). По итогу обработки получили релевантные данные для каждого документа. Во-первых, данных стало гораздо меньше, что способствует ускорению работы алгоритмов, во-вторых, сами слова представлены в удобном виде, подлежащем анализу.

Следующим этапом применялся статистический метод IDF для отбора кандидатов, наиболее информативных слов. Получив результаты, слова с наибольшими и наименьшими значениями IDF были удалены, так как первые можно отнести в группу стоп-слов, которые слишком часто встречаются, а вторая группа наоборот - редкие единичные слова, которые также не подлежат анализу.

Итоговые данные были переведены в формат BOW (Bag-of-words) и разбиты на темы с помощью тематического моделирования. Для реализации была использована модель LDA с аддитивной регуляризация из библиотеки BigARTM. Bag-of-Words или мешок слов - это модель, часто используемая при обработке текстов, представляющая собой неупорядоченный набор слов, входящих в обрабатываемый текст.

Тема 8. Извлечение именованных сущностей

Обычно выделяют три категории именованных сущностей - люди, места и организации, но в специализированных решениях часто рассматриваются и другие категории, такие как названия продуктов и произведений искусства. Также к именованным сущностям часто относят и численные выражения (цены, даты, время).

Задача извлечения именованных сущностей сводится к выявлению отрезков текста, которые являются именами собственными, и их категоризации по типу сущности. Результаты решения можно в дальнейшем использовать для снижения разреженности данных при классификации текста, для выявления объекта в рамках анализа тональности и в вопросно-ответных системах.

Тема 9. Инструменты извлечения информации из текста

Извлечение информации является разновидностью информационного поиска, связанного с обработкой текста на естественном языке. Примером извлечения информации может быть поиск деловых визитов - формально это записывается так: НанеслиВизит(Компания-Кто, Компания-Кому, ДатаВизита) - из новостных лент, таких как: "Вчера, 1 апреля 2007 года, представители корпорации Пепелац Интернэшнл посетили офис компании Гравицап Продакшнз". Главная цель такого преобразования - возможность анализа изначально "хаотичной" информации с помощью стандартных методов обработки данных. Более узкой целью может служить, например, задача выявить логические закономерности в описанных в тексте событиях.

В современных информационных технологиях роль такой процедуры, как извлечение информации, всё больше возрастает - из-за стремительного увеличения количества неструктурированной (без метаданных) информации, в частности, в Интернете. Эта информация может быть сделана более структурированной посредством преобразования в реляционную форму или добавлением XML разметки. При мониторинге новостных лент с помощью интеллектуальных агентов как раз и потребуются методы извлечения информации и преобразования её в такую форму, с которой будет удобнее работать позже.

Тема 10. Основы машинного обучения

Сегодня объемы информации и данных быстро растут. Они содержат потенциал для извлечения ценной информации и принятия осмысленных решений. Однако, для использования этих данных необходимы инструменты, способные извлекать скрытые закономерности и прогнозировать будущие события. Здесь на сцену выходит машинное обучение (ML). Суть его заключается в создании алгоритмов и моделей, которые способны автоматически извлекать знания из данных и решать задачи или предсказывать результаты на их основе.

Машинное обучение - это область искусственного интеллекта (AI), занимающаяся разработкой алгоритмов и моделей, которые способны обучаться, используя данные, составлять прогнозы, а также принимать решения без программирования.

Тема 11. Обучение с учителем: классификация

Обуче́ние с учи́телем (англ. Supervised learning) - один из способов машинного обучения, в ходе которого испытуемая система принудительно обучается с помощью примеров "стимул-реакция". С точки зрения кибернетики, является одним из видов кибернетического эксперимента. Между входами и эталонными выходами (стимул-реакция) может существовать некоторая зависимость, но она неизвестна. Известна только конечная совокупность прецедентов - пар "стимул-реакция", называемая обучающей выборкой. На основе этих данных требуется восстановить зависимость (построить модель отношений стимул-реакция, пригодных для прогнозирования), то есть построить алгоритм, способный для любого объекта выдать достаточно точный ответ. Для измерения точности ответов, так же как и в обучении на примерах, может вводиться функционал качества.

Тема 12. Обучение с учителем: регрессия

Обуче́ние с учи́телем (англ. Supervised learning) - один из способов машинного обучения, в ходе которого испытуемая система принудительно обучается с помощью примеров "стимул-реакция". С точки зрения кибернетики, является одним из видов кибернетического эксперимента. Между входами и эталонными выходами (стимул-реакция) может существовать некоторая зависимость, но она неизвестна. Известна только конечная совокупность прецедентов - пар "стимул-реакция", называемая обучающей выборкой. На основе этих данных требуется восстановить зависимость (построить модель отношений стимул-реакция, пригодных для прогнозирования), то есть построить алгоритм, способный для любого объекта выдать достаточно точный ответ. Для измерения точности ответов, так же как и в обучении на примерах, может вводиться функционал качества.

Тема 13. Оценка модели и валидация

Валидация - проверка правильности работы (предсказательной способности) аналитической модели, построенной на основе машинного обучения, а также удостоверение, что она соответствует требованиям решаемой задачи.

Проводится на независимом (т.е. не использовавшемся для обучения и тестирования) валидационном множестве после обучения и тестирования модели.

Тема 14. Обучение без учителя: кластеризация

Обучение без учителя (самообучение, спонтанное обучение, англ. Unsupervised learning) - один из способов машинного обучения, при котором испытуемая система спонтанно обучается выполнять поставленную задачу без вмешательства со стороны экспериментатора. С точки зрения кибернетики, это является одним из видов кибернетического эксперимента. Как правило, это пригодно только для задач, в которых известны описания множества объектов (обучающей выборки), и требуется обнаружить внутренние взаимосвязи, зависимости, закономерности, существующие между объектами.

Обучение без учителя часто противопоставляется обучению с учителем, когда для каждого обучающего объекта принудительно задаётся "правильный ответ", и требуется найти зависимость между стимулами и реакциями системы.

Тема 15. Обучение без учителя: PCA, SVD, t-SNE, UMAP

Метод главных компонент (Principal Component Analysis или же PCA) - алгоритм обучения без учителя, используемый для понижения размерности и выявления наиболее информативных признаков в данных. Его суть заключается в предположении о линейности отношений данных и их проекции на подпространство ортогональных векторов, в которых дисперсия будет максимальной.

Тема 16. Ансамбли

Метод машинного обучения, где несколько моделей обучаются для решения одной и той же проблемы и объединяются для получения лучших результатов называется ансамблевым методом. Основная предпосылка заключается в том, что результат работы нескольких моделей будет более точен, чем результат только одной модели.

Когда говорится об ансамблях, то вводится понятие слабого ученика(обычные модели вроде линейной регрессии или дерева решений). Множество слабых учеников являются строительными блоками для более сложных моделей. Объединение слабых учеников для улучшения качества модели, уменьшения смещения или разброса, называется сильным учеником.

Тема 17. Введение в нейронные сети

Нейросети - это предпочтительный инструмент во многих предсказательных прикладных программах исследования данных из-за их мощности, гибкости и простоты использования. Предсказательные нейросети полезны в частности в прикладных программах со сложными базовыми процессами, такими как:

Предсказание пользовательских запросов на рационализацию производства и стоимость доставки.

Предсказание вероятности отклика на прямой почтовый маркетинг для определения, по каким адресам из списка рассылки нужно направлять предложения.

Оценка клиента для определения риска предоставления ему кредита.

Обнаружение мошеннических транзакций в базе данных страховых случаев.

Тема 18. Нейронные архитектуры CNN, RNN, LSTM

Нейронные сети (Neural Networks) представляют собой мощный инструмент в машинном обучении и искусственном интеллекте. Они спроектированы по аналогии с биологическими нейронами и способны обучаться и решать разнообразные задачи. 1) Свёрточные нейронные сети (CNN): Свёрточные нейронные сети являются идеальным выбором для обработки данных, имеющих пространственную структуру, таких как изображения и видео. Они используют свёрточные слои для выделения важных признаков из входных данных и пулинговые слои для уменьшения размерности. CNN широко применяются в задачах классификации изображений, распознавания объектов и детектирования лиц.

Тема 19. Векторная семантика

Векторная семантика - это метод анализа, обработки естественного языка, который позволяет компьютерам понимать смысл запросов, их взаимосвязь в текстах. Эта технология может стать мощным инструментом для улучшения работы сайтов. Это метод, используется в обработке естественного языка для представления слов в виде числового вектора, где каждая координата отвечает за свой семантический признак. Такое представление обработки естественного языка, позволяет лучше понимать смысл запросов и их отношения друг к другу.

Тема 20. Модели-трансформеры

Трансфо́рмер (англ. Transformer) - архитектура глубоких нейронных сетей, представленная в 2017 году исследователями из Google Brain.

По аналогии с рекуррентными нейронными сетями (РНС) трансформеры предназначены для обработки последовательностей, таких как текст на естественном языке, и решения таких задач как машинный перевод и автоматическое реферирование. В отличие от РНС, трансформеры не требуют обработки последовательностей по порядку. Например, если входные данные - это текст, то трансформеру не требуется обрабатывать конец текста после обработки его начала. Благодаря этому трансформеры распараллеливаются легче чем РНС и могут быть быстрее обучены.

5. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине (модулю)

Самостоятельная работа обучающихся выполняется по заданию и при методическом руководстве преподавателя, но без его непосредственного участия. Самостоятельная работа подразделяется на самостоятельную работу на аудиторных занятиях и на внеаудиторную самостоятельную работу. Самостоятельная работа обучающихся включает как полностью самостоятельное освоение отдельных тем (разделов) дисциплины, так и проработку тем (разделов), осваиваемых во время аудиторной работы. Во время самостоятельной работы обучающиеся читают и конспектируют учебную, научную и справочную литературу, выполняют задания, направленные на закрепление знаний и отработку умений и навыков, готовятся к текущему и промежуточному контролю по дисциплине.

Организация самостоятельной работы обучающихся регламентируется нормативными документами, учебно-методической литературой и электронными образовательными ресурсами, включая:

Порядок организации и осуществления образовательной деятельности по образовательным программам высшего образования - программам бакалавриата, программам специалитета, программам магистратуры (утвержден приказом Министерства науки и высшего образования Российской Федерации от 6 апреля 2021 года №245)

Письмо Министерства образования Российской Федерации №14-55-996ин/15 от 27 ноября 2002 г. "Об активизации самостоятельной работы студентов высших учебных заведений"

Устав федерального государственного автономного образовательного учреждения "Казанский (Приволжский) федеральный университет"

Правила внутреннего распорядка федерального государственного автономного образовательного учреждения высшего профессионального образования "Казанский (Приволжский) федеральный университет"

Локальные нормативные акты Казанского (Приволжского) федерального университета

6. Фонд оценочных средств по дисциплине (модулю)

Фонд оценочных средств по дисциплине (модулю) включает оценочные материалы, направленные на проверку освоения компетенций, в том числе знаний, умений и навыков. Фонд оценочных средств включает оценочные средства текущего контроля и оценочные средства промежуточной аттестации.

В фонде оценочных средств содержится следующая информация:

- соответствие компетенций планируемым результатам обучения по дисциплине (модулю);

- критерии оценивания сформированности компетенций;

- механизм формирования оценки по дисциплине (модулю);

- описание порядка применения и процедуры оценивания для каждого оценочного средства;

- критерии оценивания для каждого оценочного средства;

- содержание оценочных средств, включая требования, предъявляемые к действиям обучающихся, демонстрируемым результатам, задания различных типов.

Фонд оценочных средств по дисциплине находится в Приложении 1 к программе дисциплины (модулю).

7. Перечень литературы, необходимой для освоения дисциплины (модуля)

Освоение дисциплины (модуля) предполагает изучение основной и дополнительной учебной литературы. Литература может быть доступна обучающимся в одном из двух вариантов (либо в обоих из них):

- в электронном виде - через электронные библиотечные системы на основании заключенных КФУ договоров с правообладателями;

- в печатном виде - в Научной библиотеке им. Н.И. Лобачевского. Обучающиеся получают учебную литературу на абонементе по читательским билетам в соответствии с правилами пользования Научной библиотекой.

Электронные издания доступны дистанционно из любой точки при введении обучающимся своего логина и пароля от личного кабинета в системе "Электронный университет". При использовании печатных изданий библиотечный фонд должен быть укомплектован ими из расчета не менее 0,5 экземпляра (для обучающихся по ФГОС 3++ - не менее 0,25 экземпляра) каждого из изданий основной литературы и не менее 0,25 экземпляра дополнительной литературы на каждого обучающегося из числа лиц, одновременно осваивающих данную дисциплину.

Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины (модуля), находится в Приложении 2 к рабочей программе дисциплины. Он подлежит обновлению при изменении условий договоров КФУ с правообладателями электронных изданий и при изменении комплектования фондов Научной библиотеки КФУ.

8. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

Видеолекции по курсу Natural Language Processing - https://www.coursera.org/course/nlp
Семинар Обработка естественного языка - http://nlpseminar.ru/archive/
Школа анализа данных Яндекс - http://shad.yandex.ru/

9. Методические указания для обучающихся по освоению дисциплины (модуля)

Вид работ	Методические рекомендации
лекции	В ходе лекционных занятий вести конспектирование учебного материала. Обращать внимание на категории, формулировки, раскрывающие содержание тех или иных явлений и процессов, научные выводы и практические рекомендации, положительный опыт в ораторском искусстве. Желательно оставить в рабочих конспектах поля, на которых делать пометки из рекомендованной литературы, дополняющие материал прослушанной лекции, а также подчеркивающие особую важность тех или иных теоретических положений. Задавать преподавателю уточняющие вопросы с целью уяснения теоретических положений, разрешения спорных ситуаций.
практические занятия	В ходе лекционных занятий вести конспектирование учебного материала. Обращать внимание на категории, формулировки, раскрывающие содержание тех или иных явлений и процессов, научные выводы и практические рекомендации, положительный опыт в ораторском искусстве. Желательно оставить в рабочих конспектах поля, на которых делать пометки из рекомендованной литературы, дополняющие материал прослушанной лекции, а также подчеркивающие особую важность тех или иных теоретических положений. Задавать преподавателю уточняющие вопросы с целью уяснения теоретических положений, разрешения спорных ситуаций.
самостоя-тельная работа	Процесс самостоятельной работы можно разделить на базовую и дополнительную части. Базовая СРС обеспечивает подготовку студента к текущим аудиторным занятиям и контрольным мероприятиям для всех дисциплин учебного плана. Результаты этой подготовки проявляются в активности студента на занятиях и в качестве выполненных контрольных работ, тестовых заданий, сделанных докладов и других форм текущего контроля. Базовая СРС может включать следующие формы работ: изучение лекционного материала, предусматривающие проработку конспекта лекций и учебной литературы; поиск (подбор) и обзор литературы и электронных источников информации по индивидуально заданной проблеме курса; выполнение домашнего задания или домашней контрольной работы, выдаваемых на практических занятиях; изучение материала, вынесенного на самостоятельное изучение; подготовка к практическим занятиям; подготовка к контрольной работе или коллоквиуму; подготовка к зачету, аттестациям; написание реферата (эссе) по заданной проблеме. Дополнительная СРС направлена на углубление и закрепление знаний студента, развитие аналитических навыков по проблематике учебной дисциплины. К ней относятся: подготовка к экзамену; выполнение расчетно-графической работы; выполнение курсовой работы или проекта; исследовательская работа и участие в научных студенческих конференциях, семинарах и олимпиадах; анализ научной публикации по заранее определенной преподавателем теме; анализ статистических и фактических материалов по заданной теме, проведение расчетов, составление схем и моделей на основе статистических материалов и др.
экзамен	Изучение дисциплины завершается зачетом. Подготовка к зачету способствует закреплению, углублению и обобщению знаний, получаемых, в процессе обучения, а также применению их к решению практических задач. Готовясь к зачету, студент ликвидирует имеющиеся пробелы в знаниях, углубляет, систематизирует и упорядочивает свои знания. На зачете студент демонстрирует то, что он приобрел в процессе обучения по конкретной учебной дисциплине. За 3-4 дня нужно систематизировать уже имеющиеся знания. На консультации перед зачетом студентов познакомят с основными требованиями, ответят на возникшие у них вопросы. Поэтому посещение консультаций обязательно. Требования к организации подготовки к зачетам те же, что и при занятиях в течение семестра, но соблюдаться они должны более строго. При подготовке к зачетам у студента должен быть хороший учебник или конспект литературы, прочитанной по указанию преподавателя в течение семестра. Здесь можно эффективно использовать листы опорных сигналов. Вначале следует просмотреть весь материал по сдаваемой дисциплине, отметить для себя трудные вопросы. Обязательно в них разобраться. В заключение еще раз целесообразно повторить основные положения, используя при этом листы опорных сигналов.

10. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень программного обеспечения и информационных справочных систем (при необходимости)

Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень программного обеспечения и информационных справочных систем, представлен в Приложении 3 к рабочей программе дисциплины (модуля).

11. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Материально-техническое обеспечение образовательного процесса по дисциплине (модулю) включает в себя следующие компоненты:

Помещения для самостоятельной работы обучающихся, укомплектованные специализированной мебелью (столы и стулья) и оснащенные компьютерной техникой с возможностью подключения к сети "Интернет" и обеспечением доступа в электронную информационно-образовательную среду КФУ.

Учебные аудитории для контактной работы с преподавателем, укомплектованные специализированной мебелью (столы и стулья).

Компьютер и принтер для распечатки раздаточных материалов.

12. Средства адаптации преподавания дисциплины к потребностям обучающихся инвалидов и лиц с ограниченными возможностями здоровья

При необходимости в образовательном процессе применяются следующие методы и технологии, облегчающие восприятие информации обучающимися инвалидами и лицами с ограниченными возможностями здоровья:

- создание текстовой версии любого нетекстового контента для его возможного преобразования в альтернативные формы, удобные для различных пользователей;

- создание контента, который можно представить в различных видах без потери данных или структуры, предусмотреть возможность масштабирования текста и изображений без потери качества, предусмотреть доступность управления контентом с клавиатуры;

- создание возможностей для обучающихся воспринимать одну и ту же информацию из разных источников - например, так, чтобы лица с нарушениями слуха получали информацию визуально, с нарушениями зрения - аудиально;

- применение программных средств, обеспечивающих возможность освоения навыков и умений, формируемых дисциплиной, за счёт альтернативных способов, в том числе виртуальных лабораторий и симуляционных технологий;

- применение дистанционных образовательных технологий для передачи информации, организации различных форм интерактивной контактной работы обучающегося с преподавателем, в том числе вебинаров, которые могут быть использованы для проведения виртуальных лекций с возможностью взаимодействия всех участников дистанционного обучения, проведения семинаров, выступления с докладами и защиты выполненных работ, проведения тренингов, организации коллективной работы;

- применение дистанционных образовательных технологий для организации форм текущего и промежуточного контроля;

- увеличение продолжительности сдачи обучающимся инвалидом или лицом с ограниченными возможностями здоровья форм промежуточной аттестации по отношению к установленной продолжительности их сдачи:

- продолжительности сдачи зачёта или экзамена, проводимого в письменной форме, - не более чем на 90 минут;

- продолжительности подготовки обучающегося к ответу на зачёте или экзамене, проводимом в устной форме, - не более чем на 20 минут;

- продолжительности выступления обучающегося при защите курсовой работы - не более чем на 15 минут.

Программа составлена в соответствии с требованиями ФГОС ВО и учебным планом по направлению 45.03.01 "Филология" и профилю подготовки "Прикладная филология: иностранный (английский) язык в международной коммуникации".

Приложение 2
к рабочей программе дисциплины (модуля)
Б1.В.21 Обработка естественных языков (NLP)

Перечень литературы, необходимой для освоения дисциплины (модуля)

Направление подготовки: 45.03.01 - Филология
Профиль подготовки: Прикладная филология: иностранный (английский) язык в международной коммуникации
Квалификация выпускника: бакалавр
Форма обучения: очное
Язык обучения: русский
Год начала обучения по образовательной программе: 2024

Основная литература:

Хиценко В.П., Основы программирования [Электронный ресурс]: учебное пособие / Хиценко В.П. - Новосибирск : Изд-во НГТУ, 2015. - 83 с. - ISBN 978-5-7782-2706-4 - Режим доступа: http://www.studentlibrary.ru/book/ISBN9785778227064.html
Грант С.И., Обработка неструктурированных текстов. Поиск, организация и манипулирование [Электронный ресурс] / Грант С. Ингерсолл, Томас С. Мортон, Эндрю Л. Фэррис - М.: ДМК Пресс, 2015. - 414 с. - ISBN 978-5-97060-144-0 - Режим доступа: http://www.studentlibrary.ru/book/ISBN9785970601440.html
Матвеев М.Г., Модели и методы искусственного интеллекта. Применение в экономике [Электронный ресурс] /: учеб. пособие / М.Г. Матвеев, А.С. Свиридов, Н.А. Алейникова. - М.: Финансы и статистика, 2014. - 448 с. - ISBN 978-5-279-03279-2 - Режим доступа: http://www.studentlibrary.ru/book/ISBN9785279032792.html

Дополнительная литература:

Лукашевич Н.В., Тезаурусы в задачах информационного поиска [Электронный ресурс] / Лукашевич Н.В. - М.: Издательство Московского государственного университета, 2011. - 512 с. - ISBN 978-5-211-05926-9 - Режим доступа: http://www.studentlibrary.ru/book/ISBN9785211059269.html
Хожемпо В.В., Азбука научно-исследовательской работы студента [Электронный ресурс] : учеб. пособие / В.В. Хожемпо, К.С. Тарасов, М.Е. Пухлянко. - изд. 2-е, испр. и доп. - М. : Издательство РУДН, 2010. - 107 с. - ISBN 978-5-209-03527-5 - Режим доступа: http://www.studentlibrary.ru/book/ISBN9785209035275.html

Приложение 3
к рабочей программе дисциплины (модуля)
Б1.В.21 Обработка естественных языков (NLP)

Перечень информационных технологий, используемых для освоения дисциплины (модуля), включая перечень программного обеспечения и информационных справочных систем

Направление подготовки: 45.03.01 - Филология

Профиль подготовки: Прикладная филология: иностранный (английский) язык в международной коммуникации

Квалификация выпускника: бакалавр

Форма обучения: очное

Язык обучения: русский

Год начала обучения по образовательной программе: 2024

Освоение дисциплины (модуля) предполагает использование следующего программного обеспечения и информационно-справочных систем:

Операционная система Microsoft Windows 7 Профессиональная или Windows XP (Volume License)

Пакет офисного программного обеспечения Microsoft Office 365 или Microsoft Office Professional plus 2010

Браузер Mozilla Firefox

Браузер Google Chrome

Adobe Reader XI или Adobe Acrobat Reader DC

Kaspersky Endpoint Security для Windows

Учебно-методическая литература для данной дисциплины имеется в наличии в электронно-библиотечной системе "ZNANIUM.COM", доступ к которой предоставлен обучающимся. ЭБС "ZNANIUM.COM" содержит произведения крупнейших российских учёных, руководителей государственных органов, преподавателей ведущих вузов страны, высококвалифицированных специалистов в различных сферах бизнеса. Фонд библиотеки сформирован с учетом всех изменений образовательных стандартов и включает учебники, учебные пособия, учебно-методические комплексы, монографии, авторефераты, диссертации, энциклопедии, словари и справочники, законодательно-нормативные документы, специальные периодические издания и издания, выпускаемые издательствами вузов. В настоящее время ЭБС ZNANIUM.COM соответствует всем требованиям федеральных государственных образовательных стандартов высшего образования (ФГОС ВО) нового поколения.

Учебно-методическая литература для данной дисциплины имеется в наличии в электронно-библиотечной системе Издательства "Лань" , доступ к которой предоставлен обучающимся. ЭБС Издательства "Лань" включает в себя электронные версии книг издательства "Лань" и других ведущих издательств учебной литературы, а также электронные версии периодических изданий по естественным, техническим и гуманитарным наукам. ЭБС Издательства "Лань" обеспечивает доступ к научной, учебной литературе и научным периодическим изданиям по максимальному количеству профильных направлений с соблюдением всех авторских и смежных прав.

Учебно-методическая литература для данной дисциплины имеется в наличии в электронно-библиотечной системе "Консультант студента", доступ к которой предоставлен обучающимся. Многопрофильный образовательный ресурс "Консультант студента" является электронной библиотечной системой (ЭБС), предоставляющей доступ через сеть Интернет к учебной литературе и дополнительным материалам, приобретенным на основании прямых договоров с правообладателями. Полностью соответствует требованиям федеральных государственных образовательных стандартов высшего образования к комплектованию библиотек, в том числе электронных, в части формирования фондов основной и дополнительной литературы.