Федеральное государственное бюджетное образовательное

учреждение высшего образования

Московский государственный университет имени М.В. Ломоносова

Факультет вычислительной математики и кибернетики

**УТВЕРЖДАЮ**

декан факультета вычислительной математики и кибернетики

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_/И.А. Соколов /**

**«\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_\_г.**

**РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ**

**Наименование дисциплины:**

**Анализ неструктурированных данных**

**Уровень высшего образования:**

**бакалавриат**

**Направление подготовки / специальность:**

**01.03.02 «Прикладная математика и информатика» (3++)**

**Направленность (профиль):**

**Системное программирование и компьютерные науки**

**Форма обучения:**

**очная**

**Москва 2023**

Рабочая программа дисциплины (модуля) разработана в соответствии с самостоятельно установленным МГУ образовательным стандартом (ОС МГУ) для реализуемых основных профессиональных образовательных программ высшего образования по направлению подготовки 01.03.02, 01.04.02 "Прикладная математика и информатика" программы бакалавриата Утвержден приказом МГУ от 30 августа 2019 года № 1041 (в редакции приказов МГУ от 11 сентября 2019 года № 1109, от 10 июня 2021 года № 609, от 7 октября 2021 года № 1048, от 21 декабря 2021 года № 1404, от 2 ноября 2022 года № 1299)

**1. МЕСТО ДИСЦИПЛИНЫ (МОДУЛЯ) В СТРУКТУРЕ ОПОП ВО:**

Дисциплина относится к вариативной части ОПОП ВО.

**2. ЦЕЛИ И ЗАДАЧИ ДИСЦИПЛИНЫ:**

Главная цель изучения учебной дисциплины «Анализ неструктурированных данных» – освоение основ автоматической обработки неструктурированных данных (текстов и изображений), что предполагает овладение базовыми навыками работы с соответствующими программными средствами и наборами данных. Дисциплина организована как практико-ориентированный курс в поддержку потокового курса по машинному обучению, из всех видов задач обработки данных преимущественно рассматриваются задачи анализа текстовых данных.

Изучаются основные этапы и средства обработки текстов на естественном языке, включая статистический анализ и языковые модели. Обзорно рассматриваются базовые и современные архитектуры нейронных сетей. Показываются пути применения методов машинного обучения, включая нейронные сети, для решения ряда прикладных задач анализа данных (в том числе классификация, кластеризация, генерация текстов и др.). Теоретический материал подкрепляется практическими семинарскими занятиями и домашними работами.

**3. РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ (МОДУЛЮ):**

|  |  |  |
| --- | --- | --- |
| Планируемые результаты обучения по дисциплине (модулю) | |  |
| Содержание и код компетенции. | Индикатор (показатель) достижения компетенции | Планируемые результаты обучения по дисциплине, сопряженные с индикаторами  достижения компетенций |
| УК-2. Способен в контексте профессиональной деятельности использовать знания об основных понятиях и методах естествознания. | УК-2.3 – Имеет навыки выбора методов решения задач профессиональной деятельности на основе теоретических знаний | Понимать основные особенности неструктурированных данных. Знать подходы к построению компьютерных моделей анализа данных, принципы поуровневого и статистического анализа текстов на естественном языке. Иметь представление о методах машинного обучения, включая нейронные сети, а также о современных языковых моделях, применяемых для анализа текстов. |
| ОПК-2. Способен применять и адаптировать существующие математические и компьютерные методы для разработки и реализации алгоритмов решения прикладных задач. | ОПК-2.1. Выбирает компьютерные/суперкомпьютерные методы для решения задач профессиональной деятельности;  ОПК-2.2. Использует современное программное обеспечение, в том числе отечественного происхождения, для решения задач профессиональной деятельности | Понимать особенности современных архитектур нейронных сетей и области их применения. Знать типичные прикладные задачи анализа текстовых данных и принципы построения прикладных систем на базе машинного обучения, иметь представление о необходимых для этого лингвистических ресурсах. Уметь определять метод решения задачи анализа данных и способ его программной реализации. |
| **СПК–9.** Способен применять современные системы программирования для решения прикладных задач. |  | Иметь представление о возможностях современных программных библиотек для анализа данных, в том числе на основе нейронных сетей. Владеть навыками применения готовых библиотечных модулей и открытых ресурсов (датасетов, корпусов) для программирования и обучения моделей анализа данных и проведения экспериментов с ними. |

**4. ОБЪЕМ ДИСЦИПЛИНЫ**

составляет 3 з.е., в том числе 54 академических часа, отведенных на контактную работу обучающихся с преподавателем, 54 академических часа на самостоятельную работу обучающихся.

**5. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ), СТРУКТУРИРОВАННОЕ ПО ТЕМАМ (РАЗДЕЛАМ) С УКАЗАНИЕМ ОТВЕДЕННОГО НА НИХ КОЛИЧЕСТВА АКАДЕМИЧЕСКИХ ЧАСОВ И ВИДЫ УЧЕБНЫХ ЗАНЯТИЙ:**

**5.1. Структура дисциплины (модуля) по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий (в строгом соответствии с учебным планом)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Наименование разделов и тем дисциплины (модуля),  Форма промежуточной аттестации по дисциплине (модулю) | Номинальные трудозатраты обучающегося | | | Всего академических часов |
| Контактная работа  (во взаимодействии с преподавателем)  Виды работы, академические часы | | Самостоятельная работа обучающегося,  академические часы |
| Занятия лекционного типа | Занятия семинарского типа |
| 1. Введение | 4 | 0 | 2 | 4 |
| 2. Базовые этапы анализа текста | 4 | 2 | 8 | 14 |
| 3. Статистический анализ текстовых данных | 6 | 2 | 8 | 16 |
| 4. Машинное обучение в задачах классификации и кластеризации | 4 | 2 | 6 | 12 |
| 5. Нейронные сети для анализа данных | 6 | 6 | 8 | 22 |
| 6. Нейросетевые языковые модели | 4 | 2 | 6 | 12 |
| 7. Современные прикладные задачи анализа данных | 8 | 4 | 8 | 20 |
| Промежуточная аттестация (экзамен) |  |  | 8 | 8 |
| **Итого** | 36 | 18 | 54 | 108 |

**5.2. Содержание разделов (тем) дисциплины**

|  |  |  |
| --- | --- | --- |
| № п/п | Наименование разделов (тем) дисциплины | Содержание разделов (тем) дисциплин |
| 1. | Введение | Виды неструктурированных данных и актуальность их обработки. Обзор прикладных задач анализа неструктурированных данных, подходы к их решению, понятие набора данных (датасета). Язык Питон как средство программирования систем анализа данных. |
| 2. | Базовые этапы анализа текста | Единицы текста на естественном языке и этапы его анализа. Неоднозначность интерпретации единиц текста, полисемия и омонимия.  Сегментация и токенизация теста. Методы и инструментальные средства сегментации.  Морфологический и лексический анализ текста. Современные морфологические процессоры для русского языка, их функции. |
| 3. | Статистический анализ текстовых данных | Абсолютные и относительные частоты текстовых объектов. Понятие N-граммы. Закон Ципфа-Мандельброта. Статистика встречаемости букв и  N-грамм и ее применение.  Текстовые коллекции и корпуса как источник статистики и средство машинного обучения. Примеры корпусов, виды лингвистической разметки.  Принципы построения статистической языковой модели и ее применение. Сглаживание в языковой модели. Понятие перплексии и ее смысл. |
| 4. | Машинное обучение в задачах классификации и кластеризации | Векторные модели документов текстовой коллекции, модель "мешка слов", косинусная мера близости текстов. Обзор методов машинного обучения для задачи классификации текстов коллекции. Меры оценки качества классификации.  Методы кластеризации, основные особенности (обзор). Способы оценки качества кластеризации. Примеры применения кластеризации. |
| 5. | Нейронные сети для анализа данных | Нейронная сеть как модель человеческой памяти. Персептрон: архитектура, ограничения. Метод обратного распространения ошибки для обучения нейронной сети (обзорно). Применение персептрона в задаче распознавания символов.  Современные архитектуры нейронных сетей. Сверточные нейронные сети, роли сверток и пулинга (обзорно). Понятие глубокого обучения. Рекуррентные нейронные сети, их особенности. Архитектура Трансформер. Примеры применения сверточных и рекуррентных сетей, трансформерных моделей для задач анализа данных. |
| 6. | Нейросетевые языковые модели | Понятие дистрибутивной семантики, счетные дистрибутивные модели, векторные представления слов (эмбеддинги). Предсказательные нейросетевые языковые модели Word2Vec и FastText, их особенности, геометрические свойства векторного пространства. Предобученные модели для русского языка ресурса RusVectores. Маскированная языковая модель BERT, особенности ее обучения и применения, контекстуализированные эмбеддинги. |
| 7. | Современные прикладные задачи анализа текстов | Подходы к разработке систем анализа данных: на правилах (инженерный) и машинное обучение. Обзор прикладных задач автоматической обработки текстов: машинный перевод, информационный поиск, извлечение информации, оценка тональности, коррекция ошибок в текстах. Генерация текстов: виды и подходы к решению. Виртуальные помощники и чат-боты (обзорно). |

**6. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ (ФОС, ОЦЕНОЧНЫЕ И МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ) ДЛЯ ОЦЕНИВАНИЯ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ (МОДУЛЮ).**

**6.1. Типовые контрольные задания или иные материалы для проведения текущего контроля успеваемости**

***Текущий контроль*** включает самостоятельные аудиторные работы по текущим темам дисциплины (тесты, практические работы); коллоквиум для проверки освоения основных понятий по трем первым темам, а также домашние задания на изучение открытых систем и ресурсов сети Интернет и практическую разработку программ анализа данных.

Формы текущего контроля успеваемости, соотнесенные со структурой дисциплины (темами)

|  |  |
| --- | --- |
| Наименование разделов и тем дисциплины (модуля) | Форма текущего контроля успеваемости (наименование) |
| 1. Введение |  |
| 2. Базовые этапы анализа текста | Тест, Домашнее задание |
| 3. Статистический анализ текстовых данных | Домашнее задание, Коллоквиум |
| 4. Машинное обучение в задачах классификации и кластеризации | Домашнее задание |
| 5. Нейронные сети для анализа данных | Тест, Домашнее задание |
| 6. Нейросетевые языковые модели | Домашнее задание, Тест |
| 7. Современные прикладные задачи анализа данных | Домашнее задание |

***Примеры вопросов тестов и коллоквиума:***

1. В чем заключается этап графематического анализа текста? Что такое токен?
2. Назовите виды автоматического морфологического анализа.
3. В чем суть явления омонимии? Приведите пример морфологической омонимии.
4. Для заданного предложения текста указать количество словоупотреблений, число различных словоформ, число различных лемм.
5. Что такое N-грамма? Перечислите все символьные триграммы в словосочетании *на бал*.
6. В чем отличие коллекции текстов от корпуса? Какие бывают типы разметки в корпусе?
7. Что такое "мешок слов"? Как его создать и "наполнить"? Приведите пример.
8. Поясните смысл показателей *idf* и *tf.idf*.
9. Что лучше позволит разделить тексты на группы: классификация или кластеризация?
10. Какие задачи может решать персептрон, а какие принципиально не может?
11. Что характерно для векторов слов в дистрибутивных моделях семантики?
12. В чем заключается преимущество контекстуализированных эмбеддингов?
13. Какие метрики используются для оценки качества автоматического анализа данных?

### *Примеры домашних практических заданий*

1. Для заданного слова русского языка исследовать временные изменения частоты его употребления слова и смысла, рассмотрев его значения и толкования в различных толковых словарях, в Национальном корпусе русского языка (НКРЯ), а также в яндекс-новостях.
2. Составить программу, выполняющую морфологический анализ словоформ заданного русскоязычного текста и вычисление 12-15 статистических характеристик разного типа:

* морфологические: процент разных частей речи, наиболее частотные падежи у существительных и прилагательных, относительная частота падежей, наиболее частотные морфологические формы глаголов (время /лицо/ число) и т.п.
* общестатистические: общее число словоупотреблений, число различных (уникальных) словоформ, длины словоформ и их средние значения и т.п.

Следует выводить подсчитанные характеристики в удобной, обозримой форме. Предлагается рассмотреть тексты двух разных авторов/стилей/жанров и сравнить полученные результаты.

1. На основе предобученной статической модели векторного представления словдля русского языка (*Word2Vec*, *FastText*), взятой с ресурса *RusVectores:*  [*https://rusvectores.org/ru/models/*](https://rusvectores.org/ru/models/%20)  
    провести исследование семантики нескольких (5-7) самостоятельно выбранных слов (достаточно частотных, разных частей речи). Воспользовавшись предоставляемым веб-интерфейсом, найти семантически  близкие  слова, определить близость пар слов, а также исследовать другие операции, допускаемые моделью. Рассмотреть несколько разных  моделей (2-3) и сравнить их результаты для выбранных слов.

**6.2. Типовые контрольные задания или иные материалы для проведения промежуточной аттестации по дисциплине, критерии и шкалы оценивания**

Промежуточная аттестация осуществляется в форме устного экзамена, а итоговая оценка выставляется с учетом:

* суммарной оценки (в баллах) всех выполненных домашних и аудиторных работ;
* оценки итоговой письменной контрольной, состоящей из вопросов по всем темам дисциплины;
* результатов устного опроса на экзамене.

***Примеры вопросов итоговой контрольной:***

1. Укажите типичные приложения методов анализа неструктурированных данных.
2. Перечислите основные этапы обработки текста в системах АОТ.
3. Что такое лемматизация? Приведите пример.
4. Объясните смысл закона Ципфа-Мальдельброта.
5. Что такое статистическая языковая модель?
6. В чем заключается задача классификации текстов и в чем ее отличие от кластеризации?
7. Разъясните термин "переобучение" (в чем оно выражается, как обнаруживается, как преодолевается).
8. Назовите известные архитектуры нейронных сетей, укажите их главные особенности.
9. Какие операции с векторами доступны в языковых моделях Word2Vec и FastText?
10. Назовите несколько прикладных задач, решаемых с помощью нейронных  
    сетей разной архитектуры. Почему используется именно такая архитектура?

***Примеры вопросов на устном экзамене:***

1. Укажите основные особенности и сложности обработки неструктурированных данных.
2. Назовите основные подходы к решению задач автоматического анализа данных.
3. Какие лингвистические ресурсы используются для задач анализа текстов?
4. Покажите на примере, чем словоупотребление отличается от словоформы.
5. Как определяется статистика словоупотреблений в текстах?
6. Чем классификация текстов отличается от кластеризации?
7. Что такое валидационная выборка и в чем ее отличие от тестовой?
8. Какую роль выполняет функция активации в нейронных сетях? Приведите пример такой функции.
9. Что характерно для векторов слов в дистрибутивных моделях семантики?
10. В чем отличие обучения модели FastText от обучения Word2Vec ?

**6.3 Критерии и шкалы оценивания**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | **ШКАЛА И КРИТЕРИИ ОЦЕНИВАНИЯ результатов обучения (РО) по дисциплине** | | | | | | Оценка  виды оценочных средств | 2 (не зачтено) | 3 (зачтено) | 4 (зачтено) | 5 (зачтено) | | **Знания** | Отсутствие знаний | Фрагментарные знания | Общие, но не структурированные знания | Сформированные систематические знания | | **Умения** | Отсутствие умений | В целом успешное, но не систематическое умение | В целом успешное, но содержащее отдельные пробелы умение (неточности непринципиального характера) | Успешное и систематическое умение | | **Навыки  (владения, опыт деятельности)** | Отсутствие навыков (владений, опыта) | Наличие отдельных навыков (наличие фрагментарного опыта) | В целом, сформированные навыки (владения), но используемые не в активной форме | Сформированные навыки (владения), применяемые при решении задач | |

**7. РЕСУРСНОЕ ОБЕСПЕЧЕНИЕ:**

**7.1. Перечень основной и дополнительной литературы**

**Основная литература**

1. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Большакова Е.И. и др. – М.: Изд-во НИУ ВШЭ, 2017.

<https://miem.hse.ru/clschool/the_book>

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др. – М.: МИЭМ, 2011.

<http://clschool.miem.edu.ru/uploads/swfupload/files/98e8cdfb0288b275a3197626ffe06e277a03d43d.pdf>

1. Келлехер Д. Глубокое обучение. Самый краткий и понятный курс – М.: Эксмо, 2022.
2. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. – Вильямс, 2011.
3. Jurafsky D., Martin J. Speech and Language Processing. An Introduction to Natural Language Processing, ***C***omputational Linguistics and Speech Recognition***,*** Second Edition, Prentice Hall, 2009. Third Edition Draft, 2023 <https://web.stanford.edu/~jurafsky/slp3/>

**Дополнительная литература**

1. Гэддис Т. Начинаем программировать на Python. 2019.
2. Ингерсолл Г.С. и др. Обработка неструктурированных текстов. Поиск, организация и манипулирование / Пер. с англ. – М.: ДМК Пресс, 2015.
3. Коэльо Л.П., Ричарт В. Построение систем машинного обучения на языке Python. – 2016.
4. Лутц М. Изучаем Python, том 1. – 2019.
5. Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. Погружение в мир нейронных сетей. – Питер СПб, 2020.
6. Прикладная и компьютерная лингвистика / Под ред. Николаева И.С. и др. – М.: ЛЕНАНД, 2016.
7. Таулли Т. Основы искусственного интеллекта: нетехническое введение. – СПб.: Издательство БХВ, 2012 - 288 с.
8. Manning, Ch., H. Schütze. Foundations of Statistical Natural Language Processing. MIT Press, 1999.
   1. **Перечень лицензионного программного обеспечения, в том числе отечественного производства**

При реализации дисциплины может быть использовано следующее программное обеспечение:

* Операционная система Ubuntu 18.04.
* Программный продукт Python 3.5.1 (64-bit) Python Software Foundation
* Операционная система Microsoft Windows 7 - корпоративная академическая лицензия
* Операционная система Microsoft Windows 10 Education - академическая лицензия
  1. **Перечень ресурсов информационно-телекоммуникационной сети «Интернет»**
* Дистанционное обучение (ВМК) [Электронный ресурс] URL: <https://moodle.cs.msu.ru/>
* Национальный Корпус Русского Языка. [Электронный ресурс] URL: <http://ruscorpora.ru>
* АОТ: автоматическая обработка текста. [Электронный ресурс] URL: <http://www.aot.ru>
* RusVectōrēs: семантические модели для русского языка [Электронный ресурс] URL:

[https://rusvectores.org/ru/models/](https://rusvectores.org/ru/models/%20)

* 1. **Описание материально-технического обеспечения.**

Для преподавания дисциплины используется компьютерный класс, оборудованный проектором, а также маркерной или меловой доской. Компьютеры должны иметь подключение к сети Интернет, а также доступ к системе [https://moodle.cs.msu.ru/](https://moodle.cs.msu.ru/%20) .

**8. МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ПО ОРГАНИЗАЦИИ ИЗУЧЕНИЯ ДИСЦИПЛИНЫ**

**8.1. Формы и методы преподавания дисциплины**

В преподавании данной дисциплины сочетаются:

* лекции в форме презентаций, доступных студентам для их самостоятельной работы и подготовки к тестам и практическим работам;
* самостоятельные аудиторные работы по текущим темам дисциплины (тесты, практические работы по изучению интернет-ресурсов, а также прикладного программного обеспечения);
* коллоквиум для проверки освоения базовых понятий дисциплины;
* домашние задания на изучение и применение лингвистических ресурсов (корпусов и датасетов), методов автоматического анализа данных и реализующих их библиотек.

Все учебные материалы, включая презентации лекций, описания домашних заданий, варианты заданий и тесты, а также выполненные практические работы (программы и др.) выкладываются в систему [https://moodle.cs.msu.ru/](https://moodle.cs.msu.ru/%20) (факультета ВМК МГУ) и доступны студентам. Все студенты подключаются к этой системе в начале изучения дисциплины и выполняют аудиторные самостоятельные работы и тесты на базе этой системы.

Для текущего контроля успеваемости студентов применяется балльно-рейтинговая система: все выполненные студентом работы, включая аудиторные и домашние, оцениваются в баллах, которые суммируются по всем видам работ и учитываются на промежуточной аттестации. В конце изучения дисциплины проводится итоговая письменная контрольная по всем темам дисциплины, по которой выставляется оценка по пятибалльной системе.

Промежуточная аттестация проводится в форме устного экзамена (как устное собеседование), а итоговая оценка за дисциплину выставляется исходя из набранной суммы баллов за текущий контроль, полученной оценки за контрольную и результатов устного собеседования.

**8.2. Методические рекомендации преподавателю**

Лекционный материал занимает бо'льшую часть аудиторного времени, и целесообразно в ходе лекции активизировать аудиторию, задавая вопросы с целью уточнения деталей обсуждаемых понятий и методов, а также дополнняя их.

В случае лекций длительностью более 1 часа целесообразно разделять их на две логические части, проводя между ними самостоятельную работу студентов или тест.

**8.3. Методические рекомендации студентам по организации самостоятельной работы.**

Перед началом изучения дисциплины студентам рекомендуется ознакомиться с ее учебной программой (она заранее выкладывается в системе moodle.cs.msu.ru). Для успешного выполнения аудиторных самостоятельных работ и тестов следует просматривать пройденный на последнем занятии материал, пользуясь презентациями лекций. Также необходимо внимательно знакомиться с описаниями домашних заданий, для выполнения всех поставленных в них требований.

**9. Разработчик программы**

к.ф.- м.н., доцент Большакова Е.И. ([bolsh@cs.msu.ru](mailto:sol@cs.msu.ru))