Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Сибирский государственный индустриальный университет»

Кафедра прикладных информационных технологий и программирования

УТВЕРЖДАЮ
Директор института
информационных технологий и
автоматизированных систем
_____ А.В. Зимин
подпись
«____» _____ 20___ г.

# РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Анализ текстовых данных

09.04.03 «Прикладная информатика»
(направленность (профиль): «Искусственный интеллект и машинное обучение»)

Квалификация выпускника
Магистр

Форма обучения
Очная форма

Срок обучения: 2 года

Год начала подготовки 2026

Новокузнецк
2026

СибГИУ
Подлинник электронного документа хранится в системе электронного согласования документов СибГИУ
ДОКУМЕНТ ПОДПИСАН ЭП
Сертификат: daba56d7b892efca3bf5c04ed0251c98b19b099c
Владелец: Директор ИИТиАС Зимин Алексей Валерьевич
Действителен: с 15 января 2025 г. по 10 апреля 2026 г.
Дата и время: 23-02-2026 19:28:36 ПРОВЕРИТЬ ПОДПИСЬ

## 1 Цели и задачи освоения учебной дисциплины

Целями учебной дисциплины являются:

– изучение подходов к решению основных задач автоматической обработки текстов на основе классического машинного обучения и глубоких нейронный сетей.

Задачами учебной дисциплины являются:

– освоение методов предобработки текстовых данных;

– применение на практике основных методов классификации и кластеризации текстов, методов поиска и / или генерации ответа на вопрос и базовых методов машинного перевода.

## 2 Место учебной дисциплины в структуре ООП по направлению подготовки (специальности)

Учебная дисциплина относится к учебным дисциплинам части, формируемой участниками образовательных отношений **Блока 1 «Дисциплины (модули)»** ООП по направлению подготовки (специальности) 09.04.03 «Прикладная информатика».

Учебная дисциплина базируется на предварительном усвоении обучающимися учебных дисциплин:

– Математические и инструментальные методы анализа данных;

– Машинное обучение;

– Программирование глубоких нейронных сетей на Python;

– Наука о данных и аналитика больших данных.

Учебная дисциплина дополняет знания и умения, получаемые по одновременно изучаемым и последующим учебным дисциплинам:

– Наука о данных и аналитика больших данных;

– Методы и инструменты цифровой трансформации.

## 3 Планируемые результаты обучения по учебной дисциплине

Процесс изучения учебной дисциплины направлен на формирование следующих компетенций:

### – Профессиональные компетенции

| Наименование категории (группы) ПК | Код и наименование ПК | Код и наименование индикатора достижения ПК | Планируемые результаты обучения |
|---|---|---|---|
| | ПК-1: Способен участвовать в управлении работами по получению, хранению и обработки | ПК-1.1 Принимает участие в разработке моделей данных, проводит анализ больших объемов данных, строит модели на основе данных | – знать: основные математические и алгоритмические модели систем, методы их имитационного моделирования, основы построения компьютерных |

| | | | |
|---|---|---|---|
| | больших объемов данных | | дискретно-математических моделей.<br>– уметь: проводить необходимые статистические расчёты в рамках построенной экономической модели, выбирать современные методы принятия экономических решений в информационных системах. |
| | ПК-2: Способен к проведению работ по обработке и анализу научно-технической информации и результатов исследований | ПК-2.1 Собирает и изучает научно-техническую информацию по теме исследований и разработок | – знать: способы и средства сбора научно-технической и статистической информации по тематике исследования.<br>– уметь: применять информационные технологии и статистические методы для сбора, обработки и анализа научно-технической информации по тематике исследования. |
| | | ПК-2.2 Проводит анализа научных данных, результатов экспериментов и наблюдений | – знать: основные положения теории статистики, основные методы анализа информационных процессов в сложных экономических системах.<br>– уметь: решать задачи теоретического и прикладного характера из различных разделов математики, статистики и теории систем, строить модели объектов и понятий. |

## 4 Объем и содержание учебной дисциплины

Учебные занятия по учебной дисциплине проводятся в форме контактной работы и в форме самостоятельной работы обучающихся.

Контактная работа включает в себя занятия лекционного типа (лекции), занятия семинарского типа (семинары, практические занятия,

практикумы), промежуточную аттестацию обучающихся и иные формы взаимодействия обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации ООП на иных условиях, в том числе при проведении промежуточной аттестации обучающихся. Контактная работа может проводиться с применением электронного обучения, дистанционных образовательных технологий.

## Объем учебной дисциплины

| Семестр / курс | | ***ИТОГО*** | ***3 семестр*** |
|---|---|---|---|
| Форма промежуточной аттестации | | | *экзамен* |
| Трудоёмкость | *академ. час.* | ***144*** | *144* |
| | *зачетных единиц* | ***4*** | *4* |
| Лекции, *академ. час.* | | ***16*** | *16* |
| в форме практической подготовки | | ***0*** | *0* |
| Лабораторные работы, *академ. час.* | | ***0*** | *0* |
| в форме практической подготовки | | ***0*** | *0* |
| Практические занятия, *академ. час.* | | ***32*** | *32* |
| в форме практической подготовки | | ***0*** | *0* |
| Курсовая работа / проект, *академ. час.* | | ***0*** | *0* |
| в форме практической подготовки | | ***0*** | *0* |
| Консультации, *академ. час.* | | ***0*** | *0* |
| в форме практической подготовки | | ***0*** | *0* |
| Самостоятельная работа, *академ. час.* | | ***51*** | *51* |
| в форме практической подготовки | | ***0*** | *0* |
| Контроль, *академ. час.* | | ***45*** | *45* |
| в форме практической подготовки | | ***0*** | *0* |

## Содержание учебной дисциплины

Раздел 1 Введение в анализ текстов, базовые методы предобработки и выделения признаков;

Тема 1.1 Задачи анализа текстов, спецификация, история. Базовая предобработка текстов;

Тема 1.2 Простейшие текстовые признаки: «мешок слов», TF-IDF. Предобработка текстовых данных: регулярные выражения;

Раздел 2 Неглубокие векторные представления слов;

Тема 2.1 Идея векторных представлений, one-hot-векторы, SVD;

Тема 2.2 Модель word2vec, методы её обучения. Оптимизации обучения word2vec: SGNS и иерархический softmax;

Тема 2.3 Модель GloVe. Модель и библиотека FastText, приём Hashing Trick;

Раздел 3 Классификация текстов;

Тема 3.1 Задача классификация текстов. Логистическая регрессия на счётчиках и TF-IDF;

Тема 3.2 Неглубокие векторные представления документов. Библиотека FastText для классификации текстов. CNN для классификации текстов. Работа с обучающими данными;

Раздел 4 Разметка последовательности;

Тема 4.1 Счетные языковые модели. Морфологический анализ, скрытые Марковские модели. Нейросетевые языковые модели. Рекуррентные нейронные сети (RNN);

Тема 4.2 Генерация текстов. Извлечение именованных сущностей (NER). Перекрестное обучение;

Раздел 5 Машинный перевод;

Тема 5.1 Модели класса кодировщик-декодировщик. Механизм внимания;

Тема 5.2 Модель Трансформер. Метрики качества в машинном переводе. Улучшение качества машинного перевода;

Раздел 6 Предобученные языковые модели. Улица Сезам 1;

Тема 6.1 Векторное представление предложений;

Тема 6.2 Модель ELMo. Модель BERT. Модель GPT2. Оценка качества моделей. BERTоведение;

Раздел 7 Предобученные языковые модели. Улица Сезам 2;

Тема 7.1 (Модель XLNet. Король умер, да здравствует король! Сжатие языковых моделей;

Тема 7.2 Потомки модели BERT. Мультимодальные модели. Мультиязычные модели. Zero-shot мультиязычная классификация текстов. Мультиязычное распознавание именованных сущностей;

Раздел 8 Синтаксис в рамках грамматики зависимостей;

Тема 8.1 Автоматический синтаксический анализ. Теоретические подходы;

Тема 8.2 Парсинг на основе грамматики зависимостей. Метрики и соревнования. Инструменты. Разное. Синтаксический парсинг;

Раздел 9 Тематическое моделирование;

Тема 9.1 Постановка задачи тематического моделирования;

Тема 9.2 Модель PLSA. EM-алгоритм. Модель ARTM. Модель LDA. Модель M-ARTM. Технические аспекты обучения TM;

Раздел 10 Суммаризация и симплификация текстов;

Тема 10.1 Суммаризация текстов. Абстрактивная суммаризация. Упрощение текстов;

Раздел 11 QA-системы;

Тема 11.1 Основы разработки чат-ботов. Инструменты разработки чат-ботов;

Тема 11.2 Виртуальные ассистенты. Вопросно-ответные системы. Машинное чтение. Вопросно-ответные системы в индустрии;

Раздел 12 Графы знаний;

Тема 12.1 Методы извлечения графа знаний. Машинное обучение для извлечения графа знаний;

Тема 12.2 Разрешение многозначности.

## 5 Перечень тем лекций

| № раздела / темы дисциплины | Темы лекций | Трудоемкость, *академ.час* | |
|---|---|---|---|
| | | **всего** | **в форме практической подготовки** |
| Раздел 1. | Введение в анализ текстов, базовые методы предобработки и выделения признаков | 2 | |
| Раздел 2. | Неглубокие векторные представления слов | 2 | |
| Раздел 3. | Классификация текстов | 2 | |
| Раздел 4. | Разметка последовательности | 2 | |
| Раздел 5. | Машинный перевод | 1 | |
| Раздел 6. | Предобученные языковые модели. Улица Сезам 1 | 1 | |
| Раздел 7. | Предобученные языковые модели. Улица Сезам 2 | 1 | |
| Раздел 8. | Синтаксис в рамках грамматики зависимостей | 1 | |
| Раздел 9. | Тематическое моделирование | 1 | |
| Раздел 10. | Суммаризация и симплификация текстов | 1 | |
| Раздел 11. | QA-системы | 1 | |
| Раздел 12. | Графы знаний | 1 | |
| **Итого:** | | **16** | **0** |

## 6 Перечень тем практических занятий (семинаров)

| № раздела / темы дисциплины | Темы практических занятий (семинаров) | Трудоемкость, *академ.час* | |
|---|---|---|---|
| | | **всего** | **в форме практической подготовки** |
| Раздел 1. | Предобработка текстовых данных. Выделение признаков из текстов | 3 | |
| Раздел 2. | Модели word2vec и fastText | 3 | |
| Раздел 3. | Word2vec и fastText для классификации текстов.CNN для классификации текстов | 4 | |
| Раздел 4. | Языковые модели для генерации текстов. Извлечение именованных сущностей | 3 | |
| Раздел 5. | Модель кодировщик-декодировщик. Модель Трансформер | 4 | |
| Раздел 6. | Классификация текстов. | 3 | |

| | | | |
|---|---|---|---|
| Раздел 9. | Тематическое моделирование | 4 | |
| Раздел 10. | Алгоритм TextRank | 4 | |
| Раздел 12. | Библиотека CoreNLP | 4 | |
| **Итого:** | | **32** | **0** |

## 7 Перечень тем лабораторных работ

| **№ раздела / темы дисциплины** | **Темы лабораторных работ** | **Трудоемкость, *академ.час*** | |
|---|---|---|---|
| | | **всего** | **в форме практической подготовки** |
| | *Отсутствуют* | | |
| **Итого:** | | **0** | **0** |

## 8 Перечень тем курсовых работ (проектов)

| **№ раздела / темы дисциплины** | **Темы курсовых работ (проектов)** | **Трудоемкость, *академ.час*** | |
|---|---|---|---|
| | | **всего** | **в форме практической подготовки** |
| | *Отсутствуют* | | |
| **Итого:** | | **0** | **0** |

## 9 Виды самостоятельной работы

| **№ раздела / темы дисциплины** | **Виды самостоятельной работы** | **Трудоемкость, *академ.час*** | |
|---|---|---|---|
| | | **всего** | **в форме практической подготовки** |
| Раздел 1. | 1. Изучение лекционного материала;<br>2. Подготовка к практическому занятию;<br>3. Прохождение тестирования. | 4 | |
| Раздел 2. | 1. Изучение лекционного материала;<br>2. Подготовка к практическому занятию;<br>3. Прохождение тестирования. | 4 | |
| Раздел 3. | 1. Изучение лекционного материала;<br>2. Подготовка к практическому занятию;<br>3. Прохождение тестирования. | 4 | |
| Раздел 4. | 1. Изучение лекционного материала;<br>2. Подготовка к практическому занятию; | 4 | |

| | | | |
|---|---|---|---|
| | 3. Прохождение тестирования. | | |
| Раздел 5. | 1. Изучение лекционного материала;<br>2. Подготовка к практическому занятию;<br>3. Прохождение тестирования. | 4 | |
| Раздел 6. | 1. Изучение лекционного материала;<br>2. Подготовка к практическому занятию;<br>3. Прохождение тестирования. | 4 | |
| Раздел 7. | 1. Изучение лекционного материала;<br>2. Прохождение тестирования. | 4 | |
| Раздел 8. | 1. Изучение лекционного материала;<br>2. Прохождение тестирования. | 4 | |
| Раздел 9. | 1. Изучение лекционного материала;<br>2. Подготовка к практическому занятию;<br>3. Прохождение тестирования. | 5 | |
| Раздел 10. | 1. Изучение лекционного материала;<br>2. Подготовка к практическому занятию;<br>3. Прохождение тестирования. | 5 | |
| Раздел 11. | 1. Изучение лекционного материала;<br>2. Прохождение тестирования. | 4 | |
| Раздел 12. | 1. Изучение лекционного материала;<br>2. Подготовка к практическому занятию;<br>3. Прохождение тестирования. | 5 | |
| *Контроль* | *Подготовка к экзамену* | 45 | |
| **Итого:** | | **96** | **0** |

## 10 Матрица компетентностных задач по учебной дисциплине

| Контролируемые разделы (темы) дисциплины | Контролируемые компетенции | Оценочные средства |
|---|---|---|

| | | |
|---|---|---|
| Раздел 1. | ПК-1;<br>ПК-2 | 1. Наблюдение за работой на практическом занятии;<br>2. Тестовые задания |
| Раздел 2. | ПК-1;<br>ПК-2 | 1. Наблюдение за работой на практическом занятии;<br>2. Тестовые задания |
| Раздел 3. | ПК-1;<br>ПК-2 | 1. Наблюдение за работой на практическом занятии;<br>2. Тестовые задания |
| Раздел 4. | ПК-1;<br>ПК-2 | 1. Наблюдение за работой на практическом занятии;<br>2. Тестовые задания |
| Раздел 5. | ПК-1;<br>ПК-2 | 1. Наблюдение за работой на практическом занятии;<br>2. Тестовые задания |
| Раздел 6. | ПК-1;<br>ПК-2 | 1. Наблюдение за работой на практическом занятии;<br>2. Тестовые задания |
| Раздел 7. | ПК-1;<br>ПК-2 | 1. Тестовые задания |
| Раздел 8. | ПК-1;<br>ПК-2 | 1. Тестовые задания |
| Раздел 9. | ПК-1;<br>ПК-2 | 1. Наблюдение за работой на практическом занятии;<br>2. Тестовые задания |
| Раздел 10. | ПК-1;<br>ПК-2 | 1. Наблюдение за работой на практическом занятии;<br>2. Тестовые задания |
| Раздел 11. | ПК-1;<br>ПК-2 | 1. Тестовые задания |
| Раздел 12. | ПК-1;<br>ПК-2 | 1. Наблюдение за работой на практическом занятии;<br>2. Тестовые задания |
| Экзамен. | ПК-1;<br>ПК-2 | 1. Экзамен на основе результатов компьютерного тестирования;<br>2. Экзамен на основе результатов текущего контроля |

## 11 Входной контроль

**11.1 Цель входного контроля** − определить начальный уровень подготовленности обучающихся и выстроить индивидуальную траекторию обучения. В условиях личностно-ориентированной образовательной среды результаты входного оценивания обучающегося используются как начальные значения в индивидуальном профиле академической успешности обучающегося.

### 11.2 Описание оценочных средств

Форма проведения входного контроля – тестирование.

Длительность тестирования – 30 минут.

Тестирование проводится в течение первых двух недель изучения учебной дисциплины в СУО СибГИУ.

Для проведения тестирования используются тестовые задания для входного контроля (для ответа на 1 вопрос, как правило, выделяется 1 минута).

**Шкала оценивания результатов тестирования**

За каждый правильный ответ выставляется 1 балл. Оценка выставляется в зависимости от процента правильных ответов:

Оценка формируется в соответствии с критериями таблицы:

| Максимальный балл | Оценка |
|---|---|
| 10 | зачтено |
| 4 | Не зачтено |

Основываясь на результатах тестирования, формируется индивидуальная траектория обучающегося (группы обучающихся), исходя из следующих соображений:

– 5 – 10 – обучающийся готов к изучению учебной дисциплины без дополнительного повторения изученного ранее материала;

– 4 и менее – обучающийся имеет слабую подготовку к изучению учебной дисциплины, необходимы корректирующие занятия и консультации.

**Тестовые задания для входного контроля знаний обучающихся приведены в приложении.**

**12 Текущий контроль**

**12.1 Текущий контроль** знаний используется для оперативного и регулярного управления учебной деятельностью (в том числе самостоятельной) обучающихся. Данный вид контроля стимулирует у обучающихся стремление к систематической самостоятельной работе по изучению учебной дисциплины и позволяет оценить степень освоения изученного материала.

Текущий контроль успеваемости осуществляется в течение семестра, в ходе повседневной учебной работы с ис-пользованием онлайн-курса Анализ текстовых данных, доступного через сеть Интернет, размещенного по адресу: https://openedu.ru/course/hse/TEXT/.

Формы проведения текущего контроля – наблюдение за работой на практическом занятии (оценка результатов выполнения в онлайн-курсе; тестовые задания (в онлайн-курсе Анализ текстовых данных).

В процессе обучения осуществляется тьюторское (организационно-методическое) и техническое сопровождение самостоятельной работы

слушателей в онлайн-курсе, предоставление условий для прохождения слушателем контрольного испытания с идентификации личности слушателя в процессе контроля, с целью проверки результатов освоения онлайн-курса.

Самостоятельная работа в онлайн-курсе является асинхронной и осуществляется с использованием инструментов электронно-информационной образовательной среды, интернет-сервисов и дистанционного онлайн-взаимодействия преподавателя (тьютора) он-лайн-курса. под руководством

Ход обучения и результаты текущего и промежуточного контроля знаний фиксируются в электронной ведомости обучающегося, что позволяет контролировать учебные достижения обучающегося и своевременно проводить коррекцию его обучения.

В рамках текущего контроля формируется оценка по сто-балльной шкале. При этом вклад в итоговую оценку по курсу распределяется следующим образом:

Учебные задания, выполняемые на практических занятиях – 40%;

Итоговый контроль – 60 %.

В онлайн-курсе задан вес оценок.

Взвешенная оценка рассчитывается умножением оценки на вес задания. Взвешенная оценка используется для расчета балла за курс. Сводка по оценкам

| Тип задания | Вес | Оценка | Проходной балл | Взвешенная оценка |
|---|---|---|---|---|
| Учебное задание | 40% | Y | 40% | 0.4*Y |
| Итоговый контроль | 60% | B | 60% | 0.6*B |

## 12.2 Описание оценочных средств
## Шкалы оценивания (методики оценки)

### 12.2.1 Рекомендации по оцениванию результатов тестирования обучающихся

Тестирование проводится по изученным разделам (темам) учебной дисциплины для выявления уровня знаний и умений обучающегося, мотивирования обучающегося к активизации самостоятельной работы по усвоению учебного материала, выявления и устранения пробелов в знаниях, повышения дисциплины и организации учебной деятельности обучающегося.

Тестирование проводится на платформе «Открытое образование» в онлайн-курсе Анализ текстовых данных на Python в рамках часов, отведенных на самостоятельную работу обучающихся и частично, отведенные на выполнение практических заданий. В курсе предусмотрены тесты самоконтроля, контроль по модулям курса и итоговый контрольный тест.

### 12.2.2 Рекомендации по оцениванию в процессе наблюдения за работой на практическом занятии

Наблюдение за работой на практическом занятии – важнейший компонент контроля деятельности обучающихся и качества их обучения.

**Критерии оценки в процессе наблюдения за работой на практическом занятии:**

– самостоятельность выполнения задания и работы с методическими указаниями (учитывается работа в течение занятия, быстрота и способность нахождения в методических указаниях нужной информации);

– правильность выполнения работы (учитывается последовательность выполняемых действий, умение работать с инструментами и/или средствами вычислительной техники);

– рациональность использования времени, отведенного на задание (не одобряется затянутость выполнения работы с учетом индивидуальных особенностей обучающихся);

– активность обучающегося при выполнении практической работы (учитывается индивидуальная работа и работа в группе).

Критерии оценки по результатам наблюдения за работой на практическом занятии

| Наименование параметра оценки | Максимальное количество баллов |
|---|---|
| Самостоятельно выполняет задания; умеет работать с методическими указаниями. | 2,5 |
| Указанные в методических указаниях действия выполняет в правильной последовательности; умеет работать с инструментами и/или средствами вычислительной техники. | 1,25 |
| Рационально использует время, отведенное на задание; проявляет активность при выполнении практической работы. | 1,25 |
| **Итого** | **5** |

Максимальное количество – 5 баллов соответствует оценке «отлично», 4 балла – «хорошо», 3 балла – «удовлетворительно», 2 балла – «неудовлетворительно».

**Задания и вопросы для текущего контроля знаний обучающихся по разделам (темам) учебной дисциплины приведены в приложении.**

## 13 Промежуточная аттестация

**13.1 Оценочные материалы для промежуточной аттестации** обучающихся по учебной дисциплине предназначены для оценки степени достижения запланированных результатов обучения по завершению изучения учебной дисциплины в установленной учебным

планом форме и позволяют определить качество усвоения изученного материала.

Итоговой формой контроля сформированности компетенций у обучающихся по учебной дисциплине является: экзамен.

ФОС промежуточной аттестации включает: экзамен на основе результатов компьютерного тестирования; экзамен на основе результатов текущего контроля. Промежуточная аттестация производится на основе текущего контроля успеваемости при условии выполнения всех заданий в семестре, предусмотренных фондом оценочных средств по дисциплине, и выполнения итогового контрольного теста с прокторингом на платформе «Открытое образование» в онлайн-курсе "Анализ текстовых данных". – (https://openedu.ru/course/hse/TEXT/).

### 13.2 Оценивание обучающегося на экзамене

| Оценка | Требования к знаниям |
|---|---|
| отлично | Оценка «отлично» выставляется обучающемуся, если он глубоко и прочно усвоил программный материал, исчерпывающе, последовательно, четко и логически его излагает, умеет тесно увязывать теорию с практикой, свободно справляется с задачами, вопросами и другими видами применения знаний, причем не затрудняется с ответом при видоизменении заданий, использует в ответе материал различной литературы, правильно обосновывает принятое нестандартное решение, обладает приемами выполнения практических задач по формированию компетенций. |
| хорошо | Оценка «хорошо» выставляется обучающемуся, если он твердо знает материал, грамотно и по существу излагает его, не допуская существенных неточностей в ответе на вопрос, правильно применяет теоретические положения при решении практических вопросов и задач, обладает необходимыми умениями и приемами их выполнения. |
| удовлетворительно | Оценка «удовлетворительно» выставляется обучающемуся, если он имеет знания только основного материала, но не усвоил его деталей, допускает неточности, недостаточно правильные формулировки, испытывает сложности при выполнении практических работ и затрудняется связать теорию вопроса с практикой. |
| неудовлетворительно | Оценка «неудовлетворительно» выставляется обучающемуся, который не знает значительной части программного материала, неуверенно отвечает, допускает серьезные ошибки, не имеет представлений по методике выполнения практической работы. Как правило, оценка «неудовлетворительно» ставится обучающимся, которые не могут продолжить обучение без дополнительных занятий по данной учебной дисциплине. |

**Задания и вопросы для промежуточной аттестации обучающихся приведены в приложении.**

**14 Учебно-методическое и информационное обеспечение учебной дисциплины**

**а) литература:**

1 Рабчевский, А. Н. Синтетические данные и развитие нейросетевых технологий : учебное пособие для вузов / А. Н. Рабчевский. — Москва : Издательство Юрайт, 2024. — 187 с. — (Высшее образование). — ISBN 978-5-534-17716-9. – URL: https://urait.ru/bcode/545036 (дата обращения: 17.02.2026);

2 Анализ данных : учебник для вузов / В. С. Мхитарян [и др.] ; под редакцией В. С. Мхитаряна. — Москва : Издательство Юрайт, 2024. — 490 с. — (Высшее образование). — ISBN 978-5-534-00616-2. – URL: https://urait.ru/bcode/536007 (дата обращения: 17.02.2026).

**б) ресурсы информационно-телекоммуникационной сети «Интернет»:**

1 Электронный каталог : сайт / Научно-техническая библиотека СибГИУ. – Новокузнецк, [199 – ]. – URL: http://libr.sibsiu.ru. – URL: https://libr.sibsiu.ru;

2 Электронная библиотека // Научно-техническая библиотека СибГИУ : сайт. – Новокузнецк, [200 – ]. – URL: http://library.sibsiu.ru/LibrELibraryFullText.asp. – Режим доступа: для авторизир. пользователей. – URL: https://library.sibsiu.ru/LibrELibraryFullText.asp;

3 Консультант студента : электронно-библиотечная система / ООО «КОНСУЛЬТАНТ СТУДЕНТА». – Москва, [200 – ]. – URL: http://www.studentlibrary.ru. – Режим доступа: для авторизир. пользователей;

4 ЛАНЬ : электронно-библиотечная система : [коллекция «Единая профессиональная база знаний для технических вузов - Издательство Лань" ЭБС»] / ООО «Издательство ЛАНЬ». – Санкт-Петербург, [200 – ]. – URL: http://e.lanbook.com. – Режим доступа: для авторизир. пользователей;

5 Образовательная платформа ЮРАЙТ / ООО «Электронное издательство ЮРАЙТ». – Москва, [200 – ]. – URL: https://urait.ru. – Режим доступа: для авторизир. пользователей;

6 Университетская библиотека онлайн : электронно-библиотечная система / ООО «Директ-Медиа». – Москва, [200 – ]. – URL: https://biblioclub.ru. – Режим доступа: для авторизир. пользователей. – URL: http://www.biblioclub.ru;

7 Электронная библиотека УМЦ ЖДТ : [коллекция «Эксплуатация железных дорог»] / ФГБУ ДПО «Учебно-методический центр по образованию на железнодорожном транспорте». – Москва, [2013 – ]. – URL: https://umczdt.ru/books/. – Режим доступа: для авторизир. пользователей;

8 НАУЧНАЯ ЭЛЕКТРОННАЯ БИБЛИОТЕКА eLIBRARY.RU : база данных / ООО «НЭБ». – Москва, [200 – ]. – URL: http://elibrary.ru. – Режим доступа: по подписке;

9 Электронные периодические издания ИВИС : универсальная база данных / ООО «ИВИС». – Москва, [200 – ]. – URL: http://eivis.ru. – Режим доступа: по подписке;

10 Онлайн-курс «Анализ текстовых данных». – URL: https://openedu.ru/course/hse/TEXT/.

**в) лицензионное и свободно распространяемое программное обеспечение:**

– 7-Zip;
– Adobe Acrobat Reader;
– Astra Linux Special Edition;
– Kaspersky Endpoint Security;
– Microsoft Office;
– Microsoft Windows;
– OnlyOffice;
– PyCharm;
– Р7-Офис.

**г) базы данных и информационно-справочные системы:**

1 ГАРАНТ : справочно-правовая система / ООО «Правовой центр «Гарант». – Кемерово, [200 – ]. – Режим доступа: компьютерная сеть Сиб. гос. индустр. ун-та.;

2 КонсультантПлюс : справочно-правовая система / ООО «Информационный центр АНВИК». – Новокузнецк, [199 – ]. – Режим доступа: компьютерная сеть библиотеки Сиб. гос. индустр. ун-та.;

3 Техэксперт : информационно-справочная система / ООО «Группа компаний «Кодекс». – Кемерово, [200 – ]. – Режим доступа: компьютерная сеть Сиб. гос. индустр. ун-та.

## 15 Материально-техническое обеспечение учебной дисциплины

Материально-техническое обеспечение учебной дисциплины включает учебные аудитории, оснащенные оборудованием, компьютерной техникой, и техническими средствами обучения, в том числе:

- учебную аудиторию для проведения занятий лекционного типа, оборудованную учебной доской, экраном и мультимедийным проектором;
- учебную аудиторию для проведения занятий семинарского типа (практических занятий);
- учебную аудиторию (помещения) для групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации;
- помещения для самостоятельной работы, оснащенные компьютерной техникой с возможностью подключения к сети «Интернет» и

обеспечением доступа в электронную информационно-образовательную среду, научно-техническую библиотеку СибГИУ. Обучение по дисциплине проводится с использованием онлайн-курса «Анализ текстовых данных», доступного через сеть Интернет, размещенного по адресу: https://openedu.ru/course/hse/TEXT/. В процессе обучения осуществляется тьюторское (организационно-методическое) и техническое сопровождение самостоятельной работы слушателей в онлайн-курсе, предоставление условий для прохождения слушателем контрольного испытания с идентификации личности слушателя в процессе контроля, с целью проверки результатов освоения онлайн-курса. Самостоятельная работа в онлайн-курсе является асинхронной и осуществляется с использованием инструментов электронно-информационной образовательной среды, интернет-сервисов и дистанционного онлайн-взаимодействия преподавателя (тьютора) под руководством онлайн-курса. Ход обучения и результаты промежуточного и итогового контроля знаний фиксируются в электронной ведомости обучающегося, что позволяет контролировать учебные достижения обучающегося и своевременно проводить коррекцию его обучения. Содержание программы разработано преподавателями НИУ ВШЭ: Апишев М.А., Артемова Е.Л., к.т.н., Саркисян В.В., Кирьянов Д.П., Аксенов С.А.

Рабочая программа дисциплины составлена в соответствии с требованиями ФГОС ВО по направлению подготовки (специальности) 09.04.03 «Прикладная информатика».

Составитель(и):

доцент Маслова Елена Владимировна (кафедра прикладных информационных технологий и программирования).

Рабочая программа дисциплины рассмотрена и утверждена на заседании кафедры.

Системный аналитик группы
системной архитектуры департамента
информационных технологий
ООО "СИ СОФТ"  Е.А. Шапошникова

## Приложение

## Тестовые задания для входного контроля знаний обучающихся

**Выберите один правильный ответ.**

1 Алфавит языка программирования – это…
a. набор символов русского языка, используемых при составлении любых текстов, называемых алфавитом языка
b. набор цифр и знаков препинания, используемых при составлении текстов, называемых алфавитом языка
c. разрешенный к использованию набор символов, с помощью которого могут быть образованы слова и величины данного языка
2 Переменная характеризуется
a. именем, типом, значением
b. именем, значением
c. значением, типом
3 Величина, значение которой меняется в процессе выполнения программы, называется
a. константой
b. переменной
c. меткой
4 Операции DIV и MOD допустимы над величинами
a. целочисленного типа
b. вещественного типа
c. символьного типа
5 Какое значение примет переменная X после выполнения оператора X=17 mod 4?
a. 1
b. 4
c. 4.25
6 Дано трехзначное число x=abc. Операция x div 10 mod 10 определяет
a. единицы
b. десятки
c. сотни
7 Цикл FOR используется в том случае, если
a. необходимо, чтобы тело цикла выполнялось хотя бы один раз
b. точно знаете, сколько раз должно быть выполнено тело цикла
c. хотите, чтобы проверка была произведена прежде, чем выполнено тело цикла
8 Массив – это
a. совокупность конечного числа данных различных типов
b. совокупность конечного числа данных одного типа
c. скалярный тип данных
9 Индекс массива определяет

a. положение элементов массива относительно друг друга

b. номер элемента в массиве

c. количество элементов

10 Разветвляющейся называется алгоритмическая конструкция,…

a. в которой выбирается один из нескольких возможных путей (вариантов) вычислительного процесса;

b. в которой все этапы решения задачи выполняются строго последовательно

c. организующая многократное выполнение указанного действия

Задания и вопросы для текущего контроля знаний обучающихся по разделам (темам) учебной дисциплины

**2 Тестирование по результатам освоения лекционного материала и самостоятельной работы**

Перечень вопросов для тестирования:

**Раздел 1 Введение в анализ текстов, базовые методы предобработки и выделения признаков**

1. Можно ли для токенизации использовать информацию о пунктуации и регистре?

Да

Нет

2. Какой символ чаще всего используется для разделения слов в тексте?

В качестве ответа укажите его название с маленькой буквы и без кавычек.

Например: код

3. Для фраз "красным цветом удобно рисовать красные цветки" и "рисуем красные цветки красиво" написать векторы "мешка слов", полученные до лемматизации. Порядок слов алфавитный, при прочих равных более короткие слова идут раньше.

В качестве ответа укажите два вектора, заключенные в квадратные скобки. Числа внутри скобок разделите запятыми без пробелов.

Пример формата ввода: [1,2,3,4,5,6], [1,2,3,4,5,6]

4. Пусть в корпусе 480 документов. В 120 из них встречается слово "кот". В третьем документе слово "кот" встречается 15 раз. Всего в этом документе 60 слов. Чему равен значение TF-IDF слова "кот" в третьем документе? В формуле использовать двоичный логарифм.

В качестве ответа укажите число с точностью до 1 знака после запятой.

Пример формата ввода: 2.7

5. С помощью морфологического анализатора pymorphy2 (актуальная версия библиотеки — 0.9.1) лемматизируйте фразу "у страха глаза ох как велики", для каждого слова выбирая нормальную форму наиболее вероятного кандидата.

Выберите правильный вариант лемматизации из предлагаемых.

a. у страх глаз ох как великий
b. у страшный ох как глаз великий
c. у страх глаз охать как большой
d. у страшный ох как глаз большой
e. у страх глаз охать как великий
f. у страшный охать как глаз великий
g. у страх глаз ох как большой
h. у страшный охать как глаз большой

**Раздел 2 Неглубокие векторные представления слов**

1 Какие из указанных сущностей можно представить в виде числовых векторов для дальнейшей обработки? Выберите все правильные варианты.

a слова
b текстовые документы
c изображения

2 Какие из перечисленных видов векторных представлений по построению имеют разреженную структуру векторов?

a FastText
b One-hot векторы
c GloVe
d word2vec

3 Какую структуру имеет нейронная сеть для обучения word2vec CBOW?

a 2 свёрточных слоя + 1 полносвязный слой + softmax
b 1 свёрточный слой + 1 полносвязный слой + softmax
c 2 полносвязных слоя + softmax
d 1 свёрточный слой + 1 полносвязный слой + softmax

4 Какая размерность сжатых векторных представлений слов чаще используется на практике?

a 10-50
b 200-400
c 1000-2000
d 5000-10000

5 Векторные представления каких сущностей FastText хранит в явном виде?

a только слов
b только символьных N-грамм
c и слов, и символьных N-грамм

**Раздел 3 Классификация текстов**

1 Что из нижеперечисленного можно использовать в качестве признаков при решении задачи классификации предложений? Выберите все правильные варианты

a Векторное представление на основе синтаксического дерева предложения

b Количество букв "a" в предложении

c Мешок слов предложения

d Усреднение векторов GloVe слов предложения с весами, равными нормированной частоте этих слов в Википедии

e Значение заданной функции от скалярной метки одного из классов этого предложения

f Наличие или отсутствие в предложении смайликов

g Метка типа погоды, во время которой автор написал это предложение

2 Как обучается классификатор FastText на одном документе при использовании в качестве признаков только слов?

a На входе все слова, на выходе предсказываются классы

b На входе часть слов, на выходе предсказываются прочие слова

c На входе часть слов, на выходе предсказываются прочие слова и классы

3 Свёрточные нейронные сети позволяют учитывать A без явного использования словарных N-грамм, что находится на месте "A"?

a локальный контекст

b глобальный контекст

c синтаксическую роль

d морфологический состав

4 Какие нейросетевые слои лежат в основе архитектуры DAN?

a Рекурентные

b Полносвязные

c Свёрточные

d Слои пулинга

5 В задаче детекции спама производится бинарная классификация на класс 0 (“спам”) и класс 1 (“обычное письмо”). Тестовая выборка состоит из 6 писем. Для них приведены верные ответы и ответы, данные моделью.

класс 0, ответ модели - 0

класс 0, ответ модели — 1

класс 0, ответ модели — 0

класс 0, ответ модели — 0

класс 1, ответ модели — 1

класс 1, ответ модели — 0

Напишите, чему равна полнота детекции спама этой модели.

В качестве ответа укажите число с точностью до двух знаков после запятой. Пример формата ввода: 1.77

**Раздел 4 Разметка последовательности**

1 Какая из перечисленных задач не является задачей разметки последовательности?

a Определение тональности целого предложения

b Извлечение всех отметок времени из предложения

c Определение грамматических характеристик каждого слова

d Определение переключения кодов (перехода с одного языка на другой. Пример: believe me or not, но я съела целый торт)

2 Пусть дано предложение из трех слов: <s>, w1, w2, w3 </s>. Чему равна вероятность этого предложения в модели биграмм?

a p(<s>) p(w1|<s>) p(w2|w1) p(w3|w2) p(</s>|w3)

b p(<s>) + p(w1|<s>) + p(w2|w1) + p(w3|w2) + p(</s>|w3)

c p(<s>) (p(<s>) + p(w1)) (p(w1) + p(w2)) (p(w2) + p(w3)) (p(w3) + p(</s>))

d p(<s>) p(w1) p(w2) p(w3) p(</s>)

3 Пусть тестовое множество состоит из трех слов – w1, w2, w3. Некоторая языковая модель присваивает следующие вероятности этим словам: p(w1) = 0.25, p(w2) = 0.5, p(w3) = 0.25. Чему равна перплексия языковой модели?

a 3.175

b 0.315

c 0.03125

d 2.828

4 Рассмотрим нейросетевую модель 5-грамм, использующую один полносвязный скрытый слой и один слой для предсказания.

На вход модели подаем 4 вектора слов, на выходе модели получаем предсказание – пятое слов. Размерность одного вектора слова – 300. Размерность скрытого слоя – 256. В словаре 10000 слов. Вектора слов входят в параметры модели.

Матрица векторов слов: 300 x 10000

Скрытый слой: 1200 x 256

Выходной слой: 256 x 10000

Чему равно количество параметров модели?

В качестве ответа укажите целое число. Например: 10

5 Какова вероятность последовательности "<s> кошка пила молоко"?

В качестве ответа укажите число с точностью до 5 знаков после точки. Например: 0.62606

**Раздел 5 Машинный перевод**

1 Рассмотрим модель класса кодировщик-декодировщик для машинного перевода. Отметьте неверное утверждение.

a В качестве декодировщика может быть использована двунаправленная РНС.

b В качестве кодировщика может быть использована двунаправленная РНС.

c Скрытое состояние передается от кодировщика декодировщику.

d Кодировщик получает на вход текст на исходном языке, а декодировщик – на целевом.

2 Какая часть модели класса кодировщик-декодировщик решает задачу языкового моделирования?

a Кодировщик

b Декодировщик

3 Различаются ли обучение и использование (тестирование) моделей машинного перевода?

a Нет, отличий нет

b Да, во время тестирования декодировщик не видит правильных ответов на целевом языке.

c Да, декодировщик обучается как авторегресионная языковая модель, а используется – как неавторегресионная.

d Да, во время обучения декодировщик использует механизм внимания, а во время использования – нет

4 Пусть текущее скрытое состояние декодера – это вектор размерности 3: [10, 5, 10].

Кодировщик насчитывает четыре скрытых состояний, их размерность тоже равна 3. Скрытые состояния кодировщика:

[0, 1, 1]
[5, 0, 1]
[1, 1, 0]
[0, 5, 1]

Используя скалярное произведение в качестве функции близости и softmax, найдите веса внимания и составьте вектор контекста для данного скрытого состояния.

Используйте округление до целого числа. Введите вектор контекста как последовательность чисел через пробел и без скобок. Пример: 1 2 3

5 Выполните второй шаг механизма внимания и получите выход механизма внимания.

Введите выходную матрицу как последовательность чисел через пробел. Начните с элемента матрицы с индексом [11], затем – элемент с индексом [12], и так далее.

Пример: 1.11 2.22 3.33 4.44

**Раздел 6**

**Предобученные языковые модели. Улица Сезам 1**

1 Векторное представление предложения (эмбеддинг предложения) можно получить из векторов слов, входящих в предложение. Для этого необходимо…

a Конкатенировать вектора слов

b Усреднить вектора слов

c Взять вектор первого слова

d Взять вектор символа начала предложения <s>

2 Что общего у моделей skip-thought и универсального кодировщика (universal sentence embedding)?

a Обе модели используют архитектуру кодировщик-декодировщик

b Обе модели обучаются в режиме обучения с учителем

c Обе модели используют рекуррентные нейронные сети

d Для обучения обеих моделей данные должны быть представлены в виде троек последовательных предложений

3 Какие из перечисленных ниже моделей являются авто-регрессионными языковыми моделями?

Выберите все верные варианты ответа.

a doc2vec

b BERT

c ELMo

d GPT2

4 Допустим, что мы спроецирировали на плоскость вектора многозначного слова, полученные с помощью модели ELMo. График, изображающий такую проекцию, поможет выявить кластерную структуру в векторах.

Продолжите рассуждение.

a В идеале, каждый кластер отражает один из смыслов слова.

b Наличие кластерной структуры помогает выявить ошибки в обучающих данных: например, те случаи, когда слово было написано с опечаткой.

c В идеале кластеров должно быть столько, сколько может быть форм у слова.

d Если слово является частеречным омонимом (например, стекло – глагол и существительное, замок – глагол и существительное), кластеров должно быть строго два: по одному на каждую часть речи.

5 Какие формальные задачи используются для обучения модели BERT?

a Выберите все правильные варианты ответа.

b Предсказание слова на месте маски

c Предсказание следующего предложение

d Предсказание следующего слова

e Предсказание контекста по данному слову

**Раздел 7 Предобученные языковые модели. Улица Сезам 2**

1 Ниже перечислено несколько моделей векторных слов и предложений. Выберите все контекстно-зависимые модели:

a Word2vec
b GloVe
c BERT
d XLNet
e Doc2vec

Ни одна из вышеперечисленных

2 Ниже перечислено несколько моделей векторных слов и предложений. Выберите все модели, учитывающие и правые, и левые контексты, то есть, являющиеся двунаправленными:

a Word2vec
b GloVe
c BERT
d XLNet
e ELMo
f Doc2vec

3 От какой модели вы ожидаете лучших показателей качества на бенчмарках GLUE и SuperGLUE?

a BERT
b GPT-2
c ELMo
d XLNet

4 В какой из перечисленных ниже моделей используется языковая модель с перестановками?

a BERT
b GPT-2
c ELMo
d XLNet

5 Каково соотношение между количеством весов в квантизованной модели Q8BERT и модели BERT?

a В квантизованной модели больше весов
b В квантизованной модели меньше весов
c Количество весов в квантизованной модели и в модели BERT совпадает

**Раздел 8 Синтаксис в рамках грамматики зависимостей**

Вам предложены два варианта разбора предложения "Сельский хор из Калмыкии стал российским призером":

A.

B.

1 Какой из представленных выше синтаксических разборов выполнен в грамматике составляющих?

a A
b B

2 В каком из перечисленных ниже предложений нет эллипсиса?

a Мама гладит белье, а дочь – собаку.
b Я буду пить кефир, а ты?
c Дети собирались в школу, пока мама готовила завтрак.
d В Москве сегодня хорошая погода, а в Туле – дождь.

3 Какое из перечисленных ниже предложений является проективным?

a Все коты любят есть сосиски
b Сосиски все коты любят есть
c Есть любят сосиски все коты

4 Зачем нужна операция SWAP в transition-based парсере?

a Для ускорения парсинга
b Для сокращения используемой при парсинге памяти
c Для борьбы с непроективностью

5 Какая архитектура парсеров лучше улавливает локальные связи между словами?

a transition-based парсеры
b graph-based парсеры
c Ни одна не улавливает
d Обе улавливают одинаково хорошо

**Раздел 9 Тематическое моделирование**

1 Какая из указанных метрик качества тематической модели отвечает за сходимость обучающего её процесса?

a Когерентность
b Перплексия
c Разреженность

2 В общем случае при обучении тематической модели количество тем

a Задаётся вручную перед началом обучения
b Определяется алгоритмом обучения автоматически из данных

3 Выпишите вероятностную модель p(w|d) для PLSA (с учётом применения гипотезы условной независимости).

4 Можно ли сказать, что задача оптимизации, решаемая при построении тематической модели в общем случае является невыпуклой?

a Да
b Нет

5 Выпишите строки входной матрицы для тематического моделирования, если коллекция состоит из следующих 3-х документов:

"я любить гулять"
"улица тепло"
"я любить тепло"

Слова в модели упорядочиваются по алфавиту, вещественные числа в матрице округлите до второго знака после запятой. При округлении число, меньшее 0.5 округляется вниз, большее или равное 0.5 – вверх. Содержимое вектора записывается в квадратных скобках, числа и векторы перечисляются в порядке следования через запятую. Дробная часть чисел отделяется от целой точкой, для целых чисел дробная часть НЕ указывается.

Пример ответа: [0.11, 0.22, 0.33], [0.11, 0.22, 0.33]

**Раздел 10 Суммаризация и симплификация текстов**

1 Допустим, что в качестве реферата научной статьи мы хотим использовать все термины, которые употреблены в этой научной статье. Такой подход к составлению реферата относится к:

a Абстрактивной суммаризации
b Экстрактивной суммаризации

2 Допустим, что в качестве реферата новостной статьи мы хотим использовать первые 1-3 предложения в статье. Такой подход к составлению реферата относится к:

a Абстрактивной суммаризации
b Экстрактивной суммаризации

3 Какие векторные представления предложений можно использовать в алгоритме TextRank:

a Счетные модели векторизации предложений
b Нейросетевые модели векторизации предложений
c Модели векторизации предложений на основе язы-ковых моделей
d Любые

4 Рефераты, получаемые с помощью алгоритма TextRank…

a Скорее всего, корректны с точки зрения грамматики
b Скорее всего, содержат грамматические ошибки
c Могут исказить факты, представленные в исходном тексте
d Не учитывают важность предложений в исходном тексте

5 Модель T5 предназначена:

a Исключительно для абстрактивной суммаризации
b Для нескольких задач трансформации последовательности
c Для нескольких задач классификации и трансформации последовательности
d Для машинного перевода

**Раздел 11 QA-системы**

1 "Сценарий чат-бота представляет собой A граф".

Какое слово находится на месте A?

a Планарный

b Непланарный
c Направленный
d Ненаправленный
e Полный

2 Что можно отнести к свойствам Retrieval-Based диалоговой системы?

a Система может генерировать новые уникальные ответы
b Ответы корректно написаны и осмысленны
c Допускает создание на своей основе системы диалоговой общего назначения в широкой области

3 В чём преимущества параллельного кодирования реплик в модели для свободного диалога по сравнению с кросс-кодированием?

a Работает быстрее в режиме применения
b Работает быстрее в режиме обучения
c Работает качественнее
d Требует меньше памяти на хранение векторов реплик

4 Какая задача машинного обучения решается при определении интента?

5 На каком из этапов работы IR-based QA-системы используются легковесные с вычислительной точки зрения методы?

a Отбор текстовых фрагментов
b Поиск подходящей подстроки

**3 Выполнение практических работ**

Задания для практических работ, теоретический и методический материал для самостоятельного изучения представлены на платформе «Открытое образование» в он-лайн-курсе "Анализ текстовых данных". – (https://openedu.ru/course/hse/TEXT/).

**Раздел 1**

Предобработка текстовых данных. Выделение признаков из текстов

Учебное задание: провести стандартную предобработку данных с сайта Lenta.ru.

**Раздел 2 Неглубокие векторные представления слов**

Модели word2vec и fastText

Задание: дообучить модели word2vec и fastText на англоязычных отзывах о фильмах с сайта IMDB.

**Раздел 3 Классификация текстов**

word2vec и fastText для классификации текстов. CNN для классификации текстов

Задание: используя данные отзывов IMDB, построить мо-дель, которая на основе текста отзыва сможет определять, относится ли он к позитивному или к негативному классу.

**Раздел 4 Разметка последовательности**

Языковые модели для генерации текстов

Задание: построить простую вероятностную языковую мо-дель для генерации названий динозавров. ¶

Задание: написать модель для изучения именованных сущностей.

**Раздел 5 Машинный перевод**

Модель кодировщик-декодировщик, модель Трансформер.

Задание: изучить работу seq2seq моделей и механизма attention на задаче нормализации дат.

Задание: изучить работу модели трансформер.

**Раздел 6 Предобученные языковые модели. Улица Сезам 1.**

Классификация текстов.

Задание: построим модель классификации отзывов на фильмы на позитивный и негативный классы с помощью векторных представлений ELMo.

Задание: построить модель, которая сможет предсказать оценку пользователя на основе текста его отзыва.

**Раздел 9 Тематическое моделирование**

Тематическое моделирование

Задание: построить тематические модели для корпуса новостей Lenta.ru

**Раздел 10 Суммаризация и симплификация текстов**

Алгоритм TextRank

Задание: из исходного текста получить более короткий текст, который будет содержать в себе всю (или почти всю) информацию, которая была в исходном тексте.

**Раздел 12 Графы знаний**

Библиотека CoreNLP

Задание: познакомиться с библиотекой Stanford CoreNLP.

## 4 Оценивание обучающегося на экзамене

Промежуточная аттестация производится на основе текущего контроля успеваемости при условии выполнения всех заданий в семестре, предусмотренных фондом оценочных средств по дисциплине, и выполнения итогового контрольного теста с прокторингом на платформе «Открытое образование» в онлайн-курсе "Анализ текстовых данных". – (https://openedu.ru/course/hse/TEXT/)