

МИНОБРНАУКИ РОССИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
Ярославский государственный университет им. П.Г.Демидова

УЧЕБНО-МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ
ПО ДИСЦИПЛИНЕ

МЕТОДЫ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ В ИСКУССТВЕННОМ
ИНТЕЛЛЕКТЕ - 2

Направление подготовки (специальность):

02.04.02 Фундаментальная информатика и информационные технологии

Образовательная программа

Искусственный интеллект и компьютерные науки

очная форма обучения

Составитель:

ПАРАМОНОВ ИЛЬЯ ВЯЧЕСЛАВОВИЧ,
К.Ф.-М.Н., ДОЦЕНТ Ф-ТА ИВТ
ЯРГУ ИМ. П.Г. ДЕМИДОВА

г. Ярославль

Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины (модуля)

Основная литература:

- 1 Боярский, К. К. Введение в компьютерную лингвистику : Учебное пособие / К. К. Боярский. – Санкт-Петербург : Университет ИТМО, 2013. – 73 с. (ЭБС IPR BOOKS)

Дополнительная литература:

- 1 Проблемы компьютерной лингвистики и типологии : Сборник научных трудов / «ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ФАКУЛЬТЕТ РОМАНО-GERMANСКОЙ ФИЛОЛОГИИ НАУЧНО-МЕТОДИЧЕСКИЙ ЦЕНТР КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ. – Воронеж : Издательский дом ВГУ, 2017. – 246 с. URL: https://www.elibrary.ru/download/elibrary_35022773_12374093.pdf
- 2 Мамаев, И. Д. Русско-английский словарь основных терминов компьютерной лингвистики / И. Д. Мамаев // Лексикографическая копилка : Сборник научных статей / Под научной редакцией В.В. Гончаровой. – Санкт-Петербург : Санкт-Петербургский государственный экономический университет, 2020. – С. 83-92. URL: https://www.elibrary.ru/download/elibrary_43856875_79359109.pdf

Учебно-методическое обеспечение самостоятельной работы обучающихся по дисциплине (модулю)

- 1 Электронный университет Moodle ЯрГУ URL: <https://moodle.uniyar.ac.ru/>
- 2 Единое окно доступа к образовательным ресурсам URL: <http://window.edu.ru/>

Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины (модуля), включая перечень информационных справочных систем (при необходимости)

- 1 Пакет анализа естественного языка Stanza: официальный сайт. URL: <https://stanfordnlp.github.io/stanza/>
- 2 Пакет анализа естественного языка SpaCy: официальный сайт. URL: <https://spacy.io/>

Перечень информационных технологий, используемых при изучении дисциплины, включая программное обеспечение

1. Интерпретатор Python 3 (свободно распространяемое ПО)
Среда PyCharm Community Edition (свободно распространяемое ПО)

**Учебно-методические указания и рекомендации
к изучению тем лекционных и практических занятий, самостоятельной
работе студентов**

Содержание дисциплины

Наименование раздела дисциплины	Название темы с кратким содержанием
1	Методы классификации документов и предложений. Математическая постановка задачи классификации. Формальные методы определения классификации на различных уровнях лингвистического анализа (морфологическом, синтаксическом, семантическом): кластерный анализ, деревья принятия решений, Байесовский классификатор.
2	Использование искусственных нейронных сетей для решения задач компьютерной лингвистики. Архитектуры encoder-decoder, GRU, LSTM нейросети с вниманием. Языковая модель BERT.
3	Задача классификации текстов по тональности. Понятие тональности. Способы выделения классов тональности. Объективные и субъективные тональные предложения. Аспектный анализ тональности.
4	Задача выделения именованных сущностей из текста. Виды именованных сущностей. Тематическое моделирование текста. Латентно-семантический анализ. Метод сингулярного разложения матрицы. Латентное размещение Дирихле.
5	Диалоговые системы и чат-боты. Особенности диалога на естественном языке. Архитектура диалоговых систем. Обучение диалоговых систем на реальных диалогах. Принципы и инструментарий для разработки чат-ботов.
6	Задача машинного перевода. Лингвистические стратегии машинного перевода и поколения систем машинного перевода. Задачи распознавания речи. Проблема вариативности речи. Лингвистический и статистический подходы к задаче распознавания речи. Скрытые марковские модели. Методы синтеза речи. Устройство TTS-синтезатора речи и модуля лингвистической обработки текста.

Примеры заданий для лабораторных работ

1. Реализовать алгоритм выделения ключевых слов:

- TextRank — это алгоритм без обучения, основанный на графовых методах. Его основная идея - построить граф на основе входного текста, где вершинами будут кандидаты в ключевые слова, а ребрами — совместные появления слов в тексте, и ранжировать вершины с помощью специального графового алгоритма. Чем выше ранг у соседей данной вершины, тем выше будет ранг самой вершины.
- Topical PageRank — это алгоритм без обучения, похожий на TextRank. Также как и TextRank, данный метод строит граф кандидатов в ключевые слова, но использует улучшенный метод ранжирования. Основное различие между этими двумя алгоритмами заключается в том, что Topical PageRank перед стадией ранжирования вершин графа находит темы входного текста методом латентного размещения Дирихле. Данный шаг обеспечивает то, что выделенные ключевые слова будут соответствовать нужным темам. Далее алгоритм выполняет метод TextRank для каждой найденной темы.
- Kea - это алгоритм с обучением для выделения ключевых слов из текстов. Сначала алгоритм находит набор ключевых слов-кандидатов с помощью лексических методов и вычисляет несколько статистических характеристик для каждого кандидата. Затем Kea строит модель предсказаний для тренировочных текстов, у которых уже есть выделенные вручную ключевые слова. На последнем шаге алгоритм применяет метод наивного Байеса для определения ключевых слов тестовых документов. Также Kea может использовать тезаурус для выбора ключевых слов из заданного словаря.
- Maui основывается на Kea и выполняет те же шаги для выделения ключевых слов. Данные методы различаются тем, что Maui вычисляет больше статистических характеристик слов-кандидатов и применяет деревья решений вместо алгоритма наивного Байеса.

2. Написать код для определения характеристик качества выделения ключевых слов: точность, полнота и F-мера.

3. Автоматически построить тезаурус для заданной предметной области со следующими видами связей:

1. эквивалентные: синонимы, лексические варианты, квазисинонимы
2. иерархические: гипонимы, гиперонимы, часть-целое
3. ассоциации

Ниже приведены ссылки на конкретные методы, которые должны использоваться при построении тезауруса.

Методы, выделяющие конкретные связи: синонимы, гиперонимы

1. Синтаксические, использующие существующий тезаурус.

Detection of synonymy links between terms: experiment and results (2001)

Automatic Acquisition of Synonym Resources and Assessment of their Impact on the Enhanced Search in EHRs (2009)

Projecting Corpus-Based Semantic Links on a Thesaurus (1999)

Learning syntactic patterns for automatic hypernym discovery (2004)

2. Синтаксические, не использующие существующий тезаурус.

Extracting Hyponymic Relations from Chinese Free Corpus (2006)

Building a hyponymy lexicon with hierarchical structure (2002)

3. Методы, выделяющие несколько различных типов связей на основе страниц сайтов

Building a Web Thesaurus from Web Link Structure (2003)

Wikipedia Mining for an Association Web Thesaurus Construction (2007)

Методы, выделяющие ассоциативные связи на основе меры похожести

1. Статистические

Метод автоматического построения тезаурусов на основе статистической обработки

текстов на естественном языке (2012)

Automatic construction of networks of concepts characterizing document databases (1992)

Construction of a dynamic Thesaurus and its use for associated information retrieval (1989)

2. Методы кластеризации

An approach to the automatic construction of global thesauri (1990)

3. Синтаксические

Explorations in Automatic Thesaurus Discovery (1994)

4. Комбинированные

Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy (1997) - статистический + синтаксический

Критерии оценки

«Отлично» — Знает и применяет подходящие для решения алгоритмы, выбирает наиболее эффективный алгоритм. Создает полноценное приложение в среде разработки. Поясняет код и изменяет его при необходимости. Анализирует изученный материал, выделяет наиболее значимые для решения задачи факты, научные положения, соблюдает логическую последовательность в выполнении работы

«Хорошо» — Знает и применяет подходящие для решения алгоритмы, выбирает наиболее эффективный алгоритм. Создает полноценное приложение в среде разработки. Поясняет код и изменяет его при необходимости с небольшими неточностями. Выделяет подходящие для решения задачи факты, научные положения, соблюдает логическую последовательность в выполнении работы с небольшими неточностями.

«Удовлетворительно» — Знает и применяет подходящий для решения алгоритм, но с некоторыми ошибками. Создает полноценное приложение в среде разработки. С трудом поясняет код, не может изменить код при усложнении или существенном дополнении задачи. Выделяет подходящие для решения задачи факты, но нарушает логическую последовательность в выполнении работы.

«Неудовлетворительно» — Не может подобрать и реализовать подходящий для решения алгоритм. Не может создать полноценное приложение в среде разработки. Не может пояснить и изменить код. Не знает материал темы задания, нарушает логическую последовательность в выполнении работы.

Примеры вопросов к зачету

1. Математическая постановка задачи классификации.
2. Применение кластерного анализа для решения задачи классификации.
3. Деревья принятия решений
4. Применение Байесовского классификатора для решения задачи классификации.
5. Архитектура encoder-decoder.
6. Нейросети архитектур GRU, LSTM, нейросети с вниманием.
7. Языковая модель BERT и ее применение в компьютерной лингвистике.
8. Задача классификации текстов по тональности. Понятие тональности.
9. Способы выделения классов тональности. Объективные и субъективные тональные предложения.
10. Аспектный анализ тональности.
11. Задача выделения именованных сущностей из текста. Виды именованных сущностей.
12. Латентно-семантический анализ.
13. Метод сингулярного разложения матрицы.
14. Латентное размещение Дирихле.
15. Особенности диалога на естественном языке. Архитектура диалоговых систем.

16. Обучение диалоговых систем на реальных диалогах. Принципы и инструментарий для разработки чат-ботов.
17. Лингвистические стратегии машинного перевода и поколения систем машинного перевода.
18. Задачи распознавания речи. Проблема вариативности речи.
19. Лингвистический и статистический подходы к задаче распознавания речи. Скрытые марковские модели.
20. Методы синтеза речи. Устройство TTS-синтезатора речи и модуля лингвистической обработки текста.

Критерии оценки

«Отлично» – ответ на вопросы показывает всестороннее знание темы, изученной литературы, изложен логично, аргументировано и в полном объеме. Основные понятия, выводы и обобщения сформулированы убедительно и доказательно. Продемонстрированы полные и глубокие навыки практического применения методов компьютерной лингвистики.

«Хорошо» – ответ на вопросы основан на твердом знании темы. Возможны недостатки в систематизации или в обобщении материала, неточности в выводах. Продемонстрированы хорошие навыки практического применения методов компьютерной лингвистики.

«Удовлетворительно» – ответ на вопросы базируется на знании основ предмета, но имеются значительные пробелы в изложении материала, затруднения в его изложении и систематизации, выводы слабо аргументированы, в содержании допущены теоретические ошибки. Продемонстрированы элементарные навыки практического применения методов компьютерной лингвистики.

«Неудовлетворительно» – оценивается ответ на вопросы, в котором обнаружено неверное изложение темы, систематизации знаний, обобщений и выводов нет. Навыки практического применения методов компьютерной лингвистики слабые и отрывочные или отсутствуют.