

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

**УЧЕБНО-МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ**  
**ПО ДИСЦИПЛИНЕ**

*АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ*

Направление подготовки (специальность):

02.04.02 Фундаментальная информатика и информационные технологии

Образовательная программа

Искусственный интеллект и компьютерные науки

**очная форма обучения**

Составитель:

**ЛАГУТИНА Н. С., 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](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](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.uni-yar.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)

## 2. 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	Введение. Обработка естественного языка (Natural Language Processing, NLP). Области применения NLP. Методы NLP. Лингвистические ресурсы.
2	Общая схема предварительной обработки текстов. Преобразование формата. Удаление шума. Выделение единиц текста. Вычисление признаков каждого токена. Маркировка.
3	Морфологический анализ. Четкая морфология на основе словаря. Нечеткая морфология на основе системы правил. Обзор модулей морфологического анализа.
4	Программные инструменты для автоматического анализа естественного языка. Stanza, spaCy. Обзор API. Примеры использования.
5	Эксперименты и оценка качества решения задач. Сбор и разметка корпусов текстов. Обработка текста и расчет числовых параметров. Статистические характеристики для оценки результатов: точность, полнота, F-мера.
6	Классификация текстов. Виды задач классификации. Модели текста. Методы и инструменты классификации.

#### Задания для выполнения рефератов

№	Тема	Описание темы	Источники
1	Национальный корпус русского языка	Структура, состав, функциональные возможности корпуса. Примеры использования.	<a href="https://ruscorpora.ru/new/index.html">https://ruscorpora.ru/new/index.html</a>
2	Тезаурусы WordNet и RussNet. Алгоритмы построения тезаурусов типа WordNet	Описание WordNet. и RussNet. Основные элементы: существительные, прилагательные, глаголы (с примерами). Описание RussNet, Основные элементы и связи. Алгоритмы автоматизации построения тезауруса: предварительная обработка корпуса текстов, выделение терминов, определение отношений.	<a href="http://nsu.ru/xmlui/bitstream/handle/nsu/9086/louk_book.pdf">http://nsu.ru/xmlui/bitstream/handle/nsu/9086/louk_book.pdf</a> <a href="http://www.nsu.ru/xmlui/bitstream/handle/nsu/412/Text_BimenovaZB.pdf">http://www.nsu.ru/xmlui/bitstream/handle/nsu/412/Text_BimenovaZB.pdf</a> <a href="http://project.phil.spbu.ru/RussNet/index_ru.shtml">http://project.phil.spbu.ru/RussNet/index_ru.shtml</a> <a href="https://wordnet.princeton.edu/">https://wordnet.princeton.edu/</a>

3	Алгоритмы вида «обучение с учителем» для выделения ключевых слов из текстов KEA и Maui	Шаги алгоритма. Расчет числовых характеристик слов-кандидатов. Методы выбора ключевых слов. Загрузка и использование KEA.	Официальный сайт KEA <a href="http://community.nzdl.org/kea/index.html">http://community.nzdl.org/kea/index.html</a> Русскоязычная статья о KEA <a href="https://openbooks.itmo.ru/ru/file/6522/6522.pdf">https://openbooks.itmo.ru/ru/file/6522/6522.pdf</a> Статья об алгоритме Maui <a href="https://www.aclweb.org/anthology/D09-1137.pdf">https://www.aclweb.org/anthology/D09-1137.pdf</a>
4	Алгоритмы вида «обучение без учителем» для выделения ключевых слов из текстов TextRank и Topical PageRank	Шаги алгоритма. Расчет числовых характеристик слов-кандидатов. Методы выбора ключевых слов. Загрузка и использование TextRank.	TextRank <a href="https://github.com/summanlp/textrank">https://github.com/summanlp/textrank</a> Статья об алгоритме <a href="https://www.aclweb.org/anthology/W04-3252.pdf">https://www.aclweb.org/anthology/W04-3252.pdf</a> Статья об алгоритме Topical PageRank <a href="https://www.aclweb.org/anthology/D10-1036.pdf">https://www.aclweb.org/anthology/D10-1036.pdf</a>
5	Векторное представление слов	Алгоритмы word2vec. Описание и использование алгоритмов	Статья автора алгоритма word2vec <a href="https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf">https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf</a> Пример системы, реализующей алгоритм <a href="https://www.kaggle.com/pierremegret/gensim-word2vec-tutorial">https://www.kaggle.com/pierremegret/gensim-word2vec-tutorial</a>
6	Алгоритм синтаксического парсера для английского языка	Подробное описание алгоритма. Примеры применения алгоритма	<a href="https://explosion.ai/blog/parsing-english-in-python">https://explosion.ai/blog/parsing-english-in-python</a>
7	Анализ тональности текстов	Методы анализа тональности. Словарные ресурсы для анализа тональности. Оценка результатов. Открытые программные системы для определения тональности. Примеры их использования	Статьи <a href="http://www.dialog-21.ru/media/1451/50.pdf">http://www.dialog-21.ru/media/1451/50.pdf</a> <a href="http://vestnik.psu.ru/docs/2012/1/1/2012111.doc">vestnik.psu.ru/docs/2012/1/1/2012111.doc</a> <a href="http://www.isa.ru/aidt/images/documents/2014-01/25_33.pdf">http://www.isa.ru/aidt/images/documents/2014-01/25_33.pdf</a> Учебник <a href="https://www.hse.ru/data/2017/07/22/1173852769/NLP_and_DA.pdf">https://www.hse.ru/data/2017/07/22/1173852769/NLP_and_DA.pdf</a>
8	Извлечение структурированных данных из текста с помощью Томи-парсер	Методы извлечения информации. Извлечение именованных сущностей. Извлечение атрибутов понятий. Извлечение фактов. Применение Томи-парсера	<u>Томи-парсер</u> <a href="https://yandex.ru/dev/tomita/?turbo=true">https://yandex.ru/dev/tomita/?turbo=true</a> Учебник <a href="https://www.hse.ru/data/2017/07/22/1173852769/NLP_and_DA.pdf">https://www.hse.ru/data/2017/07/22/1173852769/NLP_and_DA.pdf</a>
9	Классификация полнотекстовых документов. Алгоритмы классификации с учителем	Представление данных в задачах классификации текстов. Меры сходства и различий между образами документов. Отбор терминов для классификации. Признаки документов. Оценка результатов. Пример одного из алгоритмов классификации с учителем. Описание алгоритма, пример применения.	Учебник <a href="http://window.edu.ru/resource/465/78465/files/miem_lingvistika.pdf">http://window.edu.ru/resource/465/78465/files/miem_lingvistika.pdf</a> Статья на русском языке <a href="https://cyberleninka.ru/article/n/metody-avtomaticheskoy-klassifikatsii-tekstov">https://cyberleninka.ru/article/n/metody-avtomaticheskoy-klassifikatsii-tekstov</a>
10	Классификация полнотекстовых документов. Алгоритмы классификации без учителя	Понятие кластеризации. Виды алгоритмов кластеризации: k-средних, иерархические, эвристические. Оценка результатов. Пример одного из алгоритмов классификации без учителя. Описание алгоритма,	Учебник <a href="http://window.edu.ru/resource/465/78465/files/miem_lingvistika.pdf">http://window.edu.ru/resource/465/78465/files/miem_lingvistika.pdf</a> Статья на русском языке <a href="http://www.dialog-21.ru/digest/2001/articles/kirichenko/">http://www.dialog-21.ru/digest/2001/articles/kirichenko/</a>

		пример применения.	
11	Методы автоматической рубрикации.	Методы автоматической рубрикации. Оценка качества рубрикации. Пример системы рубрикации текста. Алгоритм работы системы, оценка результатов работы. Корпус текстов, который может использоваться для оценки работы системы.	<a href="https://cyberleninka.ru/article/n/avtomaticheskaya-rubrikatsiya-tekstov-metody-i-problemy">https://cyberleninka.ru/article/n/avtomaticheskaya-rubrikatsiya-tekstov-metody-i-problemy</a>
12	Машинный перевод	Лингвистические стратегии машинного перевода и поколения систем машинного перевода. Автоматический перевод, основанный на правилах. Оценки качества машинного перевода. Статистический машинный перевод. Программные инструменты, предоставляющие API для машинного перевода	<a href="http://bwbooks.net/index.php?id1=4&amp;category=lingvistika&amp;author=leonteva-nn&amp;book=2006">http://bwbooks.net/index.php?id1=4&amp;category=lingvistika&amp;author=leonteva-nn&amp;book=2006</a>
13	Автоматическая проверка орфографии с помощью Яндекс Спеллер	Методы проверки орфографии. API Яндекс Спеллер. Демонстрация работы	Яндекс // Спеллер // URL: <a href="https://tech.yandex.ru/speller/">https://tech.yandex.ru/speller/</a>  Пикалёв Я. С., Вовнянко А. С., Денищенко И. Я. АНАЛИЗ АВТОМАТИЧЕСКИХ СИСТЕМ ПРОВЕРКИ ПРАВОПИСАНИЯ РУССКОГО ЯЗЫКА // Проблемы искусственного интеллекта. – 2018. – №. 2 (9). <a href="https://cyberleninka.ru/article/n/analiz-avtomaticheskikh-sistem-proverki-pravopisaniya-russkogo-yazyka">https://cyberleninka.ru/article/n/analiz-avtomaticheskikh-sistem-proverki-pravopisaniya-russkogo-yazyka</a>
14	Автоматическая проверка орфографии с помощью LanguageTool	Методы проверки орфографии. API LanguageTool. Демонстрация работы	LanguageTool // LanguageTool ПО для проверки грамматики и орфографии // URL: <a href="https://languagetool.org/ru/">https://languagetool.org/ru/</a>  Пикалёв Я. С., Вовнянко А. С., Денищенко И. Я. АНАЛИЗ АВТОМАТИЧЕСКИХ СИСТЕМ ПРОВЕРКИ ПРАВОПИСАНИЯ РУССКОГО ЯЗЫКА // Проблемы искусственного интеллекта. – 2018. – №. 2 (9). <a href="https://cyberleninka.ru/article/n/analiz-avtomaticheskikh-sistem-proverki-pravopisaniya-russkogo-yazyka">https://cyberleninka.ru/article/n/analiz-avtomaticheskikh-sistem-proverki-pravopisaniya-russkogo-yazyka</a>

### Требования к оформлению и защите рефератов.

Реферат оформляется в электронном виде как файл формата pdf, защита осуществляется в виде доклада, сопровождаемого презентацией. Правила оформления реферата <https://kursach37.com/oformlenie-referata-po-gost/>.

### Критерии оценки

«Отлично» – оцениваются рефераты, содержание которых основано на глубоком и всестороннем знании темы, изученной литературы, изложено логично, аргументировано и

в полном объеме. Основные понятия, выводы и обобщения сформулированы убедительно и доказательно.

«Хорошо» – оцениваются рефераты, основанные на твердом знании исследуемой темы. Возможны недостатки в систематизации или в обобщении материала, неточности в выводах. Студент твердо знает основные категории, умело применяет их для изложения материала.

«Удовлетворительно» – оцениваются рефераты, которые базируются 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.

«Неудовлетворительно» – оцениваются рефераты, в которых обнаружено неверное изложение основных вопросов темы, обобщений и выводов нет. Текст реферата целиком или в значительной части дословно переписан из первоисточника без ссылок на него.

### Пример задания для выполнения лабораторных работ

Установите одну из библиотек по обработке текста поддерживающую работу с русским языком. Создайте файл, содержащий текст рассказа или статьи. Например, выберите рассказ А. П. Чехова из электронного ресурса (<http://chehov-lit.ru/chehov/text/rassказы.htm>). Решите следующие задачи:

- а) определите количество слов текста;
- б) определите количество предложений в тексте;
- в) определите количество знаков пунктуации в тексте;
- г) определите долю существительных и долю глаголов относительно всех слов текста;
- д) введите слово и выведите список предложений, в которых оно встречается;
- е) определите количество различных слов в тексте;
- ж) выведите предложения, которые начинаются и заканчиваются на одно и тоже слово;
- з) выведите пары предложений, которые начинаются одним и тем же словом или словосочетанием.

### Критерии оценки

«Отлично» – решены все задачи.

«Хорошо» – решено 6-7 задач.

«Удовлетворительно» – решено 4-5 задач

«Неудовлетворительно» – решено менее 4 задач.

### Вопросы к зачету

1. Перечислите области применения автоматического 11 птй обработки естественного языка.
2. Сформулируйте задачи информационного поиска.
3. Сформулируйте постановку задачи реферирования и аннотирования текста.
4. Опишите задачи по обработке текста в аналитических системах.
5. Сформулируйте постановку задачи автоматического редактирования текста.

6. Опишите задачи по обработке текста в вопросно-ответных и диалоговых системах.
7. Опишите задачи по обработке текста в обучении естественному языку.
8. Сформулируйте постановку задачи автоматической генерации текста
9. Опишите задачи по обработке текста в распознавании и синтезе звучащей речи.
10. Назовите виды лингвистических ресурсов.
11. Опишите общую схему предварительной обработки текстов.
12. Опишите особенности преобразование формата сырого текста.
13. Опишите особенности удаления шума из текста.
14. Перечислите единицы текста и особенности их выделения.
15. Перечислите группы признаков и отдельные признаки токенов.
16. Дайте определение морфологического анализа.
17. Опишите метод четкой морфологии на основе словаря.
18. Опишите метод нечеткой морфологии на основе системы правил.
19. Назовите модули морфологического анализа.
20. Дайте определение стемминга и лемматизации.
21. Опишите алгоритм работы стеммера.
22. Опишите алгоритм работы лемматизатора.
23. Назовите синтаксические маркеры слов и токенов предложений.
24. Назовите программные инструменты для автоматического анализа естественного языка.
25. Опишите пайплайн обработки текста и приведите пример его организации в одной из программных библиотек.
26. Приведите пример токенизации текста.
27. Приведите пример определения морфологических характеристик текста.
28. Приведите пример определения леммы слов предложения.
29. Приведите пример определения роли токенов в предложении.
30. Назовите алгоритмы семантического анализа, используемые в программном инструменте.
31. Опишите и приведите пример определения именованных сущностей.
32. Опишите и приведите пример определения тональности слова.
33. Опишите процесс проведения экспериментов в области автоматического анализа текстов.
34. Сформулируйте особенности сбора и разметки корпуса текстов.
35. Опишите деление результатов эксперимента на группы для расчета метрик качества.
36. Дайте определение метрикам качества: точность, полнота, F-мера.

## 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 программного инструмента для обработки текста.

«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 программного инструмента для обработки текста.

«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 программного инструмента для обработки текста для решения простых задач.

«Неудовлетворительно» – оценивается ответ на вопросы, в котором обнаружено неверное изложение темы, систематизации знаний, обобщений и выводов нет. Навыки практического применения программного инструмента для обработки текста слабые и отрывочные или отсутствуют.